A UNIFIED FRAMEWORK FOR MULTI-DISTRIBUTION DENSITY RATIO ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Binary density ratio estimation (DRE), the problem of estimating the ratio p_1/p_2 given their empirical samples, provides the foundation for many state-of-the-art machine learning algorithms such as contrastive representation learning and covariate shift adaptation. In this work, we consider a generalized setting where given samples from multiple distributions $p_1, \ldots, p_k$ (for $k > 2$), we aim to efficiently estimate the density ratios between all pairs of distributions. Such a generalization leads to important new applications such as estimating statistical discrepancy among multiple random variables like multi-distribution f-divergence and bias correction via multiple importance sampling. We then develop a general framework from the perspective of Bregman divergence minimization, where each strictly convex multivariate function induces a proper loss for multi-distribution DRE. Moreover, we formally relate multi-distribution density ratio estimation and class probability estimation, theoretically justifying the use of any strictly proper scoring rule composite with a link function for multi-distribution DRE. We show that our framework leads to methods that strictly generalize their counterparts in binary DRE, as well as new methods that show comparable or superior performance on various downstream tasks.

1 INTRODUCTION

Estimating the density ratio between two distributions based on their empirical samples is a central problem in machine learning, which continuously drives progress in this field and finds its applications in many machine learning tasks such as anomaly detection (Hido et al., 2008; Smola et al., 2009; Hido et al., 2011), importance weighting in covariate shift adaptation (Huang et al., 2006; Sugiyama et al., 2007), generative modeling (Uehara et al., 2016; Nowozin et al., 2016; Grover et al., 2019), two-sample test (Sugiyama et al., 2011; Gretton et al., 2012), mutual information estimation and representation learning (Oord et al., 2018; Hjelm et al., 2018). It is such a powerful paradigm because computing density ratio focuses on extracting and preserving contrastive information between two distributions, which is crucial in many tasks.

Despite the tremendous success of binary DRE, in practice, many application scenarios involve more than two probability distributions and developing density ratio estimation methods among multiple distributions has the potential of advancing various applications such as estimating multi-distribution statistical discrepancy measures (Garcia-Garcia & Williamson, 2012), multi-domain transfer learning, bias correction and variance reduction with multiple importance sampling (Elvira et al., 2019), multi-marginal generative modeling (Cao et al., 2019) and multilingual machine translation (Dong et al., 2015; Aharoni et al., 2019).

Although recent years have witnessed significant progress and a continuously increasing trend in developing more sophisticated and advanced methods for binary DRE (Sugiyama et al., 2012; Liu et al., 2017; Rhodes et al., 2020; Kato & Teshima, 2021; Choi et al., 2021), methods for estimating density ratios among multiple distributions remain largely unexplored, besides an empirical exploration of multi-class logistic regression for multi-task learning (Bickel et al., 2008), where the density ratios serve as the resampling weights between the distribution of a pool of examples of multiple tasks and the target distribution for a given task at hand and lead to significant accuracy improvement on HIV therapy screening experiments.

In this work, we propose a unified framework based on expected Bregman divergence minimization, where any strictly convex multivariate function induces a proper loss for multi-distribution DRE, thus generalizing the framework in (Sugiyama et al., 2012) to multi-distribution case. Moreover, based on a new multivariate Bregman identity, we formally relate losses for multi-distribution density ratio estimation and class probability estimation, theoretically justifying the use of any strictly proper scoring rule (e.g., the logarithm score (Good, 1952), the Brier score (Brier et al., 1950) and the pseudo-spherical score (Good, 1971)) composite with a link function for multi-distribution DRE. By choosing a variety of specific convex functions or proper scoring rules, we show that our unified framework leads to methods that strictly generalize their counterparts for binary DRE, as well as new objectives specific to multi-distribution DRE. We demonstrate the effectiveness of our framework, and study and compare the empirical performance of its different instantiations on various downstream tasks that rely on accurate multi-distribution density ratio estimation.

2 PRELIMINARIES

2.1 MULTI-CLASS EXPERIMENTS

In multi-class experiments, we have a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with joint distribution $D(X, Y)$, where $\mathcal{X}$ is the sample space and $\mathcal{Y} = [k] := \{1, \dots, k\}$ is the finite label space. Define the probability simplex as $\Delta_k := \{\boldsymbol{p} \in \mathbb{R}^k_{\geq 0} | \mathbf{1}^\top \boldsymbol{p} = 1\}$. According to chain rule of probability, any joint distribution $D(X, Y)$ can be decomposed into class priors $\pi_i := \mathbb{P}(Y = i)$ and class conditionals $P_i(x) := \mathbb{P}(X = x|Y = i)$ for $i \in [k]$, or into sample marginal $M(x) := \mathbb{P}(X = x)$ and class probability function $\boldsymbol{\eta} : \mathcal{X} \to \Delta_k$ (i.e., $\eta_i(x) = \mathbb{P}(Y = i|X = x)$). We write $\boldsymbol{\eta}(x)$ as a vector $\boldsymbol{\eta}$ and omit x when it is clear from context. Thus we can also represent the joint distribution as $D = (\boldsymbol{\pi}, P_1, \dots, P_k)$ (where $\boldsymbol{\pi} \in \Delta_k$) or $(M, \boldsymbol{\eta})$. For any $i \in [k]$, we assume P_i has density p_i with respect to the Lebesgue measure.

Remark on notations. To avoid confusion, we would like to emphasize that the class probability is denoted as $\eta_i(x) = \mathbb{P}(Y = i|X = x)$ and the class conditional is denoted as $P_i(x) = \mathbb{P}(X = x|Y = i)$ with density $p_i(x)$. The former further satisfies the normalization constraint: $\forall x \in \mathcal{X}, \sum_{i=1}^k \eta_i(x) = 1$, while i in the latter one only serves as the index for k different distributions.

In multi-class classification, given independent and identically distributed (i.i.d.) samples from the joint distribution $D(X, Y)$, we want to learn a probabilistic classifier $\hat{\boldsymbol{\eta}} : \mathcal{X} \to \Delta_k$ to approximate the true class probability function $\boldsymbol{\eta}$ by minimizing the following ℓ-risk:

$$\mathcal{L}_{\text{CPE}}(\hat{\boldsymbol{\eta}}; D) = \mathbb{E}_{D(x,y)}[\ell(y, \hat{\boldsymbol{\eta}}(x))] = \mathbb{E}_{x\sim M}[\mathbb{E}_{y\sim\boldsymbol{\eta}(x)}[\ell(y, \hat{\boldsymbol{\eta}}(x))]] = \mathbb{E}_{x\sim M}[L(\boldsymbol{\eta}(x), \hat{\boldsymbol{\eta}}(x))] \quad (1)$$

where $\ell : [k] \times \Delta_k \to \mathbb{R}$ is the *loss* function for using the class predictor $\hat{\boldsymbol{\eta}}(x)$ when the true class is y, and $L : \Delta_k \times \Delta_k \to \mathbb{R}$ is the *expected loss* of $\hat{\boldsymbol{\eta}}(x)$ under the true class probability $\boldsymbol{\eta}(x)$.

Definition 1 (Proper loss)**.** *A loss function ℓ is proper if the corresponding expected loss satisfies: $\forall P, Q \in \Delta_k, L(P, Q) \geq L(P, P)$. It is strictly proper if the equality holds only when $P = Q$.*

In statistical decision theory (Gneiting & Raftery, 2007), the negative proper loss is also called *proper scoring rule* (i.e., $S(y, \hat{\boldsymbol{\eta}}(x)) = -\ell(y, \hat{\boldsymbol{\eta}}(x))$), which assesses the utility of the prediction. Properness of a loss is desirable in multi-class classification because it encourages the class probability estimator $\hat{\boldsymbol{\eta}}$ to match the true class probability function $\boldsymbol{\eta}$. An important property of proper loss is summarized in the following theorem:

Definition 2 (Bregman divergence)**.** *Given a differentiable convex function $\phi : \mathcal{S} \to \mathbb{R}$ defined on a convex set $\mathcal{S} \subset \mathbb{R}^d$ and two points $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$, the Bregman divergence from $\boldsymbol{x}$ to $\boldsymbol{y}$ is defined as:*

$$\mathbf{B}_\phi(\boldsymbol{x}, \boldsymbol{y}) := \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \boldsymbol{x} - \boldsymbol{y}, \nabla\phi(\boldsymbol{y}) \rangle \quad (2)$$

Theorem 1 ((Gneiting & Raftery, 2007); Proposition 7 in (Vernet et al., 2011))**.** *Given a proper loss ℓ and the corresponding expected loss L, for any $P, Q \in \Delta_k$, the generalized entropy function $\underline{L}(P) := \inf_{Q\in\Delta_k} L(P, Q) = L(P, P)$ is concave; when $\underline{L}$ is differentiable, the regret or excess risk of a predictor Q over the Bayes-optimal P is the Bregman divergence induced by the convex function $f = -\underline{L}$:*

$$\text{reg}(P, Q; \ell) := L(P, Q) - L(P, P) = \mathbf{B}_f(P, Q) \quad (3)$$

Given the Bregman divergence representation of the point-wise regret in Theorem 1 and the ℓ-risk in Equation (1), the excess risk of a class probability estimator $\hat{\boldsymbol{\eta}}$ over the Bayes optimal $\boldsymbol{\eta}$ is:

$$\begin{aligned}\operatorname{reg}(\hat{\boldsymbol{\eta}}; M, \boldsymbol{\eta}, \ell) :=& \mathcal{L}_{\text{CPE}}(\hat{\boldsymbol{\eta}}; D) - \mathcal{L}_{\text{CPE}}(\boldsymbol{\eta}; D) = \mathbb{E}_{M(x)}[L(\boldsymbol{\eta}(x), \hat{\boldsymbol{\eta}}(x)) - L(\boldsymbol{\eta}(x), \boldsymbol{\eta}(x))] \\ =& \mathbb{E}_{M(x)}[\mathbf{B}_f(\boldsymbol{\eta}(x), \hat{\boldsymbol{\eta}}(x))]\end{aligned} \tag{4}$$

2.2 Multi-distribution f-Divergence

Csiszaŕ's f-divergence is a popular way to measure the discrepancy between two probability distributions. Specifically, given two distributions P, Q and a convex function $f : \mathbb{R}_+ \to \mathbb{R} \cup \{\pm\infty\}$ satisfying $f(1) = 0$, the f-divergence between P and Q is defined as $\mathbf{D}_f(P||Q) = \mathbb{E}_Q[f(\mathrm{d}P/\mathrm{d}Q)]$. In the following, we will introduce the multi-distribution extension of f-divergence (Garcia-Garcia & Williamson, 2012).

Definition 3 (Multi-distribution f-divergence)**.** *For k probability distributions $P_1, \ldots, P_k$ on a common probability space $(\mathcal{X}, \sigma(\mathcal{X}))$ with densities $p_1, \ldots, p_k$, given multi-variate closed convex function $f : \mathbb{R}_+^{k-1} \to \mathbb{R} \cup \{\pm\infty\}$ satisfying $f(\mathbf{1}) = 0$, the multi-distribution f-divergence between $P_1, \ldots, P_{k-1}$ and P_k is defined as:*

$$\mathbf{D}_f\left(P_1, \ldots, P_{k-1} || P_k\right) = \mathbb{E}_{p_k(x)}\left[f\left(\frac{p_1(x)}{p_k(x)}, \ldots, \frac{p_{k-1}(x)}{p_k(x)}\right)\right] \tag{5}$$

2.3 Connecting Density Ratios and Class Probabilities via Link Function

Inspired by the definition in Eq. (5), we consider the following canonical density ratio vector (more discussion about this choice can be found in Section 3.2): $\boldsymbol{r}(x) = (r_1(x), \ldots, r_k(x))$ where $r_i(x) := p_i(x)/p_k(x)$ and $r_k(x) = 1$. Then we can connect a density ratio vector $\boldsymbol{r}(x) \in \mathbb{R}_+^{k-1} \times \{1\}$ and a class probability vector $\boldsymbol{\eta}(x) \in \Delta_k$ via an invertible link function.

According to Bayes' theorem, we have:

$$\frac{\mathbb{P}(X = x, Y = i)}{\mathbb{P}(X = x, Y = k)} = \frac{\pi_i p_i(x)}{\pi_k p_k(x)} = \frac{M(x)\eta_i(x)}{M(x)\eta_k(x)} \Leftrightarrow r_i(x) = \frac{p_i(x)}{p_k(x)} = \frac{\pi_k}{\pi_i} \cdot \frac{\eta_i(x)}{\eta_k(x)}. \tag{6}$$

Thus we define the following multi-distribution link function $\Psi_{\text{dr}} : \Delta^k \to \mathbb{R}_+^{k-1} \times \{1\}$ as a natural generalization of the binary DRE link function (Menon & Ong, 2016; Vernet et al., 2011):

$$[\Psi_{\text{dr}}(\boldsymbol{\eta})]_i := \frac{\pi_k}{\pi_i} \cdot \frac{\eta_i}{\eta_k} = r_i, \text{ for all } i \in [k]. \tag{7}$$

Given Eq. (7) and the normalization constraint $\sum_{i \in [k]} \eta_i = 1$, we obtain the inverse link function:

$$[\Psi_{\text{dr}}^{-1}(\boldsymbol{r})]_i := \frac{\pi_i r_i}{\sum_{j \in [k]} \pi_j r_j} = \eta_i, \text{ for all } i \in [k]. \tag{8}$$

Thus given knowledge of the prior distribution $\boldsymbol{\pi}$ (which can also be easily estimated from empirical samples), one can transform a class probability estimator into a density ratio estimator via $\hat{\boldsymbol{r}} = \Psi_{\text{dr}}(\hat{\boldsymbol{\eta}})$ and vice versa via $\hat{\boldsymbol{\eta}} = \Psi_{\text{dr}}^{-1}(\hat{\boldsymbol{r}})$.

3 A Unified Framework for Multi-distribution DRE

3.1 Multi-distribution Density Ratio Estimation Problem Setup

Following the basic formulation of multi-class experiments in Section 2.1, we now introduce the problem setup of multi-distribution density ratio estimation (DRE). Recall that $\mathcal{X}$ is the common data domain and $P_1, \ldots, P_k$ are k different distributions defined on $\mathcal{X}$ with densities $p_1, \ldots, p_k$. Suppose we are given n_i i.i.d. samples $\{x_j^{(i)}\}_{j=1}^{n_i}$ from each distribution P_i. The goal of multi-distribution DRE is to estimate the density ratios between all pairs of distributions $\{r_{ij} := p_i/p_j\}_{i,j \in [k]}$ from the i.i.d. datasets $\{\{x_j^{(i)}\}_{j=1}^{n_i}\}_{i=1}^{k}$. In this paper, we assume that the density ratios are always well-defined on domain $\mathcal{X}$ (e.g., when the distributions have strictly positive densities), which is also a common assumption in binary DRE problem (Kanamori et al., 2009; Kato & Teshima, 2021).

A naive approach towards this problem is to separately estimate each density p_i from $\{x_j^{(i)}\}_{j=1}^{n_i}$ and then plug in p_i and p_j to get r_{ij}. However, as previous theoretical works (Kpotufe, 2017; Nguyen et al., 2007; Kanamori et al., 2012; Que & Belkin, 2013) suggest, directly estimating density ratios has many advantages in practical settings. Specifically, we know that (1) optimal convergence rates depend only on the smoothness of the density ratio and not on the densities; (2) optimal rates depend only on the intrinsic dimension of data, thus escaping the curse of dimension in density estimation. Inspired by these observations in binary DRE, this paper aims to develop a general framework for directly estimating multi-distribution density ratios. Moreover, we also theoretically prove that various interesting facts (Menon & Ong, 2016; Sugiyama et al., 2012), which hold in the binary case, extend to our multi-distribution case in Section 4.

While most previous works focus on DRE in binary cases, multi-distribution DRE has many important downstream applications. For example, given any integrable function $\phi : \mathcal{X} \to \mathbb{R}$, suppose we want to use importance sampling to estimate the expectation of ϕ with respect to a target distribution Q with density q w.r.t. the base measure:

$$\mathbb{E}_{q(x)}[\phi(x)] = \int_{\mathcal{X}} q(x)\phi(x)\mathrm{d}x = \int_{\mathcal{X}} p(x)\frac{q(x)}{p(x)}\phi(x)\mathrm{d}x = \mathbb{E}_{p(x)}\left[r(x)\cdot\phi(x)\right] \tag{9}$$

where we use the density ratio $r = p/q$ to correct the bias caused by using samples from the proposal distribution p rather than the target distribution q. However, in practice, finding a good proposal is critical yet challenging (Owen & Zhou, 2000). An alternative and more robust strategy is to use a population of different proposals (sampling schemes) and use a set of density ratios to correct the bias, which is also known as multiple importance sampling (MIS) (Cappé et al., 2004; Elvira et al., 2015). Given k different proposals $p_1, \ldots, p_k$, the MIS estimation of the expectation is given by:

$$\mathbb{E}_{q(x)}[\phi(x)] = \sum_{i=1}^{k} \omega_i \mathbb{E}_{p_i(x)}\left[\frac{q(x)}{p_i(x)}\phi(x)\right] \tag{10}$$

where ω_i is the weight for each proposal p_i and satisfies $\sum_i \omega_i = 1$. Thus a more efficient and accurate multi-distribution DRE method will lead to better MIS. In the context of multi-source off-policy policy evaluation (Kallus et al., 2021), the proposals correspond to a set of demonstration policies and the target distribution is the query policy whose performance we want to evaluate from the offline multi-souce demonstrations; in the context of multi-domain transfer learning setting (covariate shift adaptation) (Bickel et al., 2008; Dinh et al., 2013), the proposals correspond to a set of data generating distributions (e.g. multiple source domains or various data augmentation strategies) and the target is the test distribution we care about. Estimating multi-distribution density ratios also allows us to compute important information quantities among multiple random variables such as the multi-distribution f-divergence in Equation (5), which can be used to analyze various kinds of discrepancy and correlations between multiple random variables and further has the potential of inspiring new generative models for multiple marginal matching problem (Cao et al., 2019).

3.2 Multi-distribution DRE via Bregman Divergence Minimization

Inspired by the success of Bregman divergence minimization for unifying various DRE methods in the binary case (Sugiyama et al., 2012), in this section, we propose a general framework for solving the multi-distribution density ratio estimation problem. First, we discuss our modeling choices. Although our goal is to estimate $\binom{k}{2}$ density ratios (between all possible pairs), the solution set $\{r_{ij} := p_i/p_j\}_{i,j\in[k]}$ actually has $k-1$ degrees of freedom (e.g., $r_{ik} = r_{ij}\cdot r_{jk}$). Thus without loss of generality, we parametrize the following $k-1$ density ratio models $\hat{\boldsymbol{r}}_{\boldsymbol{\theta}} = (\hat{r}_{\theta_1}, \ldots, \hat{r}_{\theta_{k-1}})$ to approximate the true canonical density ratios $\boldsymbol{r} = (r_1, \ldots, r_{k-1})$, where $r_i := p_i/p_k$ for $i \in [k-1]$. For the simplicity of notation, we will omit the dependence on the parameters $\boldsymbol{\theta}$ and write our density ratio models as $\hat{\boldsymbol{r}} = (\hat{r}_1, \ldots, \hat{r}_{k-1})$. An advantage of such modeling choice is that any density ratio can be recovered within one step of computation $\frac{p_i}{p_j} = \frac{p_i/p_k}{p_j/p_k} = \frac{r_i}{r_j}$, thus avoiding large compounding error while naturally ensuring consistency within the solution set (i.e., if we parametrize $\hat{r}_{ij}$, $\hat{r}_{jk}$ and $\hat{r}_{ik}$ respectively, we have to make sure they satisfy $\hat{r}_{ik} = \hat{r}_{ij}\cdot\hat{r}_{jk}$).

Since our goal is to optimize $\hat{\boldsymbol{r}}$ to approximate the true density ratios $\boldsymbol{r}$, we consider to use Bregman divergence (Def. 2) to measure the discrepancy between $\boldsymbol{r}$ and $\hat{\boldsymbol{r}}$. Specifically, for any strictly convex

function $f : \mathbb{R}_+^{k-1} \to \mathbb{R}$ and $\forall x \in \mathcal{X}$, we have the following point-wise optimization problem:

$$\min_{\hat{\boldsymbol{r}}(x) \in \mathbb{R}_+^{k-1}} \mathbf{B}_f(\boldsymbol{r}(x), \hat{\boldsymbol{r}}(x)) = f(\boldsymbol{r}(x)) - f(\hat{\boldsymbol{r}}(x)) - \langle \nabla f(\hat{\boldsymbol{r}}(x)), \boldsymbol{r}(x) - \hat{\boldsymbol{r}}(x) \rangle \tag{11}$$

which corresponds to the difference between the value of f at $\boldsymbol{r}$, and the value of the first-order Taylor expansion of f around point $\hat{\boldsymbol{r}}$ evaluated at point $\boldsymbol{r}$. Although the current formulation can be understood as a regression problem from $\hat{\boldsymbol{r}}(x)$ to the true density ratios $\boldsymbol{r}(x)$, we actually only have i.i.d. samples $x \sim p_1, \ldots, p_k$ instead of the true targets $\boldsymbol{r}(x)$. In this case, we consider to use the following expected Bregman divergence to measure the overall discrepancy from the true density ratios $\boldsymbol{r}$ to the density ratio models $\hat{\boldsymbol{r}}$:

$$\mathcal{L}_{\text{DRE}}(\hat{\boldsymbol{r}}; D) = \int_{\mathcal{X}} p_k(x) \Big(f(\boldsymbol{r}(x)) - f(\hat{\boldsymbol{r}}(x)) - \langle \nabla f(\hat{\boldsymbol{r}}(x)), \boldsymbol{r}(x) - \hat{\boldsymbol{r}}(x) \rangle \Big) \mathrm{d}x \tag{12}$$

$$= \mathbb{E}_{p_k(x)} \left[\langle \nabla f(\hat{\boldsymbol{r}}(x)), \hat{\boldsymbol{r}}(x) \rangle - f(\hat{\boldsymbol{r}}(x)) \right] - \sum_{i \in [k-1]} \mathbb{E}_{p_i(x)} [\partial_i f(\hat{\boldsymbol{r}}(x))] + C \tag{13}$$

where $C := \int_{\mathcal{X}} p_k(x) f(\boldsymbol{r}(x)) \mathrm{d}x = \mathbf{D}_f(P_1, \ldots, P_{k-1} \| P_k)$ is a constant with respect to $\hat{\boldsymbol{r}}$ and the equality comes from the fact that $p_k \cdot (r_1, \ldots, r_{k-1}) = (p_1, \ldots, p_{k-1})$ according to the definition of $\boldsymbol{r}$. The rationale behind the above choice is that it allows us to get an unbiased estimation of the discrepancy between $\boldsymbol{r}$ and $\hat{\boldsymbol{r}}$ only using i.i.d. samples from $p_1, \ldots, p_k$. Specifically, since C is a constant, we have the following optimization problem over $\hat{\boldsymbol{r}}$ to approximate the true density ratios (where each expectation $\mathbb{E}_{p_i}$ can be empirically estimated using samples from p_i):

$$\min_{\hat{\boldsymbol{r}}: \mathcal{X} \to \mathbb{R}_+^{k-1}} \mathbb{E}_{p_k(x)} \left[\langle \nabla f(\hat{\boldsymbol{r}}(x)), \hat{\boldsymbol{r}}(x) \rangle - f(\hat{\boldsymbol{r}}(x)) \right] - \sum_{i \in [k-1]} \mathbb{E}_{p_i(x)} \left[\partial_i f(\hat{\boldsymbol{r}}(x)) \right] \tag{14}$$

Interestingly, the above multi-distribution DRE formulation, which is based on Bregman divergence minimization, can be alternatively derived from the perspective of variational estimation of multi-distribution f-divergence. In the following, We briefly discuss such an interpretation of Eq. (14).

Based on Fenchel duality, we can represent any strictly convex function $f : \mathbb{R}_+^{k-1} \to \mathbb{R} \cup \{+\infty\}$ through its conjugate function $f^*(\boldsymbol{s}) := \max_{\boldsymbol{r} \in \mathbb{R}_+^{k-1}} \langle \boldsymbol{s}, \boldsymbol{r} \rangle - f(\boldsymbol{r})$ as:

$$f(\boldsymbol{r}(x)) = \max_{\boldsymbol{s}: \mathcal{X} \to \mathbb{R}^{k-1}} \langle \boldsymbol{r}(x), \boldsymbol{s}(x) \rangle - f^*(\boldsymbol{s}(x)), \quad \text{for any } x \in \mathcal{X}. \tag{15}$$

In order to estimate the multi-distribution f-divergence defined in Eq. (5) only using samples from $P_1, \ldots, P_k$ (instead of their density information), we consider the following variational representation of multi-distribution f-divergence by substituting Eq. (15) into Eq. (5):

$$\mathbf{D}_f(P_1, \ldots, P_{k-1} || P_k) = - \min_{\boldsymbol{s}: \mathcal{X} \to \mathbb{R}^{k-1}} \left[- \sum_{i \in [k-1]} \mathbb{E}_{p_i(x)} [\boldsymbol{s}(x)]_i + \mathbb{E}_{p_k(x)} f^*(\boldsymbol{s}(x)) \right] \tag{16}$$

We then have the following lemma revealing the equivalence between the optimization problem in Eq. (14) and Eq. (16).

Proposition 1 (DRE via variational estimation of multi-distribution f-divergence)**.** *Given a strictly convex function $f : \mathbb{R}_+^{k-1} \to \mathbb{R} \cup \{+\infty\}$, the optimization problem in Eq. (14) (induced by minimizing expected Bregman divergence $\mathbf{B}_f(\boldsymbol{r}, \hat{\boldsymbol{r}})$) is equivalent to the one in Eq. (16) (for variational estimation of multi-distribution f-divergence) under change of variables satisfying: $\nabla f(\hat{\boldsymbol{r}}(x)) = \boldsymbol{s}(x)$, $\forall x \in \mathcal{X}$.*

4 Connecting Losses for Multi-class Classification and DRE

Inspired by the link and inverse link function connecting density ratio estimators and class probability estimators introduced in Section 2.3, there has been existing theoretical works that established the connections between the *losses* of binary classification and binary DRE (Menon & Ong, 2016). However, despite the empirical exploration in (Bickel et al., 2008), the relationship between losses of multi-class classification and multi-distribution DRE has not been theoretically understood.

In Section 2.1, we have shown that the exact minimization of the excess risk for any strictly proper loss ℓ results in the true class probability function $\boldsymbol{\eta}$, and consequently gives us the true density ratio $\boldsymbol{r}$ through the link function $\Psi_{\mathrm{dr}}(\boldsymbol{\eta})$. In the following, we take a further step to show that essentially the procedure of minimizing any strictly proper loss is equivalent to minimizing an expected Bregman divergence between the true density ratios $\boldsymbol{r}$ and the approximate density ratios $\hat{\boldsymbol{r}}$, thus generalizing the theoretical results in binary case (Menon & Ong, 2016) to the multi-distribution case and justifying the validity of using any strictly proper scoring rule (e.g. Brier score (Brier et al., 1950) and pseudo-spherical score (Good, 1971)) for multi-distribution DRE. All proofs for this section can be found in Appendix A.3. We start by introducing a new multivariate Bregman identity that may be of independent interest.

Lemma 1 (Multivariate Bregman Identity)**.** *Given a convex function $f : \mathbb{R}^{k-1} \to \mathbb{R}$, we can define an associated function $f^{\circledast}(u_1, \ldots, u_{k-1}) = (1 + \sum_{i\in[k-1]} u_i) f\left(\frac{1}{1+\sum_{i\in[k-1]} u_i} \cdot \boldsymbol{u}\right)$. We can show that (i) $f^{\circledast}$ is convex and (ii) for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{k-1}$, their associated Bregman divergences satisfy:*

$$\mathbf{B}_f\left(\frac{1}{1+\sum_{i\in[k-1]} u_i} \cdot \boldsymbol{u}, \frac{1}{1+\sum_{i\in[k-1]} v_i} \cdot \boldsymbol{v}\right) = \frac{1}{1+\sum_{i\in[k-1]} u_i} \mathbf{B}_{f^{\circledast}}(\boldsymbol{u}, \boldsymbol{v}). \tag{17}$$

One can then apply Lemma 1 with $u_i = \frac{\pi_i}{\pi_k} r_i$ and $v_i = \frac{\pi_i}{\pi_k} \hat{r}_i$ for each $i \in [k-1]$ and use the fact that $\mathbf{B}_{f_\pi^{\circledast}}(\boldsymbol{r}, \hat{\boldsymbol{r}}) = \mathbf{B}_{f^{\circledast}}(\boldsymbol{u}, \boldsymbol{v})$ for $f_\pi^{\circledast}(\boldsymbol{r}) = f^{\circledast}(\frac{1}{\pi_k}\boldsymbol{\pi} \circ \boldsymbol{r})$ to establish the following connection between the optimality gap of density ratio estimators and class probability estimators, where we use $\boldsymbol{a} \circ \boldsymbol{b}$ to denote the element-wise product between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, and $\boldsymbol{\pi}_{[1:k-1]} \in \mathbb{R}^{k-1}$ as the vector when restricting $\boldsymbol{\pi}$ onto its first $k-1$ coordinates.

Proposition 2. *For any convex function $f : \mathbb{R}_+^{k-1} \to \mathbb{R}$, and two density ratio vectors $\boldsymbol{r}(x)$ and $\hat{\boldsymbol{r}}(x)$, one can construct corresponding class probability vectors $\boldsymbol{\eta}(x) = \Psi_{\mathrm{dr}}^{-1}(\boldsymbol{r}(x))$ and $\hat{\boldsymbol{\eta}}(x) = \Psi_{\mathrm{dr}}^{-1}(\hat{\boldsymbol{r}}(x))$ through the inverse link function in Eq. (8), and obtain:*

$$\mathbf{B}_f\left(\boldsymbol{\eta}(x), \hat{\boldsymbol{\eta}}(x)\right) = \frac{\pi_k}{\pi_k + \sum_{i\in[k-1]} \pi_i r_i(x)} \mathbf{B}_{f_\pi^{\circledast}}\left(\boldsymbol{r}(x), \hat{\boldsymbol{r}}(x)\right) \text{ for all } x \in \mathcal{X}, \tag{18}$$

where we define the convex function $f_\pi^{\circledast}$ induced by some prior distribution $\pi \in \Delta_k$ as

$$f_\pi^{\circledast}(r_1, \ldots, r_{k-1}) := \left(1 + \sum_{i\in[k-1]} \pi_i r_i / \pi_k\right) \cdot f\left(\frac{\boldsymbol{\pi}_{[1:k-1]} \circ \boldsymbol{r}}{\pi_k + \sum_{i\in[k-1]} \pi_i r_i}\right). \tag{19}$$

Combining Proposition 2 with the Bregman divergence representation of the point-wise regret for a proper risk ℓ for multi-class classification in Eq. (4), we provide the following main theorem that interprets the minimization of multi-class classification regret as multi-distribution DRE under expected Bregman divergence minimization.

Theorem 2. *Given any strictly proper loss ℓ, for any joint data distribution $D(X, Y)$ with class prior $\pi \in \Delta_k$, the multi-class classification regret defined in Eq. (4) satisfies that:*

$$\mathrm{reg}(\hat{\boldsymbol{\eta}}; M, \boldsymbol{\eta}, \ell) = \pi_k \mathbb{E}_{p_k(x)} \mathbf{B}_{f_\pi^{\circledast}}(\boldsymbol{r}(x), \hat{\boldsymbol{r}}(x)), \tag{20}$$

for $f_\pi^{\circledast}$ as defined in Eq. (19), and $\boldsymbol{r} = \Psi_{\mathrm{dr}}(\boldsymbol{\eta})$ and $\hat{\boldsymbol{r}} = \Psi_{\mathrm{dr}}(\hat{\boldsymbol{\eta}})$ as defined in Eq. (7).

Theorem 2 generalizes a known equivalence between density ratio estimation and class probability estimation in the binary case (see Section 5 in (Menon & Ong, 2016)), and serves as a theoretical justification for a new equivalence in the more complicated multi-class experiments. Besides, in comparison to the binary case result, we also provide a simpler proof, loosen the assumptions on the twice-differentiability of convex function f induced by the proper loss ℓ (i.e., $f = -\underline{L}$, see Theorem 1 for more details), and generalize the argument to an arbitrary prior distribution $\pi \in \Delta^k$ instead of the uniform prior case $\pi_1 = \pi_2 = 1/2$ considered in (Menon & Ong, 2016).

Moreover, we notice that multi-distribution f-divergence among class conditionals $P_1, \ldots, P_k$ also corresponds to the statistical information measure in multi-class experiments (DeGroot, 1962) (defined as the gap between the prior and posterior generalized entropy). Since we have established the equivalence between multi-class DRE (Eq. (14)) and variational estimation of multi-distribution f-divergence (Eq. (16)), we can show by choosing particular convex functions (associated with the loss ℓ for multi-class classification), multi-distribution DRE can be viewed as estimating the statistical information measure in multi-class experiments. See detailed discussions in Appendix A.3.1.

5 EXAMPLES OF MULTI-DISTRIBUTION DRE

In the binary density ratio matching under Bregman divergence framework (Sugiyama et al., 2012), we can choose various convex functions to recover popular binary DRE methods such as KLIEP (Sugiyama et al., 2008), LSIF (Kanamori et al., 2009) and Logistic Regression (Franklin, 2005). In this section, we provide some instantiations of our multi-distribution DRE framework. Specifically, Section 3.2 suggests that any strictly convex multivariate function $f : \mathbb{R}_+^{k-1} \to \mathbb{R}$ induces a proper loss for multi-distribution DRE, and Section 4 justifies that any strictly proper scoring rule composite with Ψ_{dr} can also be used for multi-distribution DRE. We briefly discuss some choices of the convex function or proper scoring rule, and we provide detailed derivations in Appendix A.4.

5.1 METHODS INDUCED BY CONVEX FUNCTIONS

Multi-class Logistic Regression. From Section 2.3, we know that there is a one-to-one correspondence between a class probability estimator and a density ratio estimator: $\hat{\boldsymbol{r}} = \Psi_{\mathrm{dr}} \circ \hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\eta}} = \Psi_{\mathrm{dr}}^{-1} \circ \hat{\boldsymbol{r}}$. For the clarity of presentation, here we assume the class prior distribution $\boldsymbol{\pi}$ is uniform such that $\hat{r}_i(x) = \hat{\eta}_i(x)/\hat{\eta}_k(x)$ and $\hat{\eta}_i(x) = \hat{r}_i(x)/\sum_{j=1}^k \hat{r}_j(x)$. To recover the loss of multi-class logistic regression, we choose the following convex function to be $f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \frac{1}{k}\sum_{i=1}^k \hat{r}_i \log\left(\hat{r}_i/\sum_{j=1}^k \hat{r}_j\right)$. In this case, the loss in Eq. (14) reduces to:

$$\frac{1}{k}\mathbb{E}_{p_k(x)}\left[\log\left(\sum_{j=1}^k \hat{r}_j(x)\right)\right] - \frac{1}{k}\sum_{i=1}^{k-1}\mathbb{E}_{p_i(x)}\left[\log\left(\frac{\hat{r}_i(x)}{\sum_{j=1}^k \hat{r}_j(x)}\right)\right] = -\left(\frac{1}{k}\sum_{i=1}^k \mathbb{E}_{p_i(x)}[\log \hat{\eta}_i(x)]\right) \tag{21}$$

We provide discussions for the general case (non-uniform prior $\boldsymbol{\pi}$) in Appendix A.4.1. Interestingly, we noticed that the above convex function also gives rise to the multi-distribution Jensen-Shannon divergence (Garcia-Garcia & Williamson, 2012) (also known as the information radius (Sibson, 1969), $\mathbf{D}_f(P_1, \ldots, P_k) = \frac{1}{k}\sum_{i=1}^k D_{\mathrm{KL}}(P_i \| \frac{1}{k}\sum_{j=1}^k P_j)$) up to a constant of $\log k$.

Multi-LSIF. When the convex function associated with the Bregman divergence is chosen to be $f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \frac{1}{2}\sum_{i=1}^{k-1}(\hat{r}_i - 1)^2 = \frac{1}{2}\|\hat{\boldsymbol{r}} - \mathbf{1}\|^2$, the loss in Eq. (14) reduces to:

$$\frac{1}{2}\sum_{i=1}^{k-1}\mathbb{E}_{p_k(x)}\left[\hat{r}_i^2(x) - 1\right] - \sum_{i=1}^{k-1}\mathbb{E}_{p_i(x)}\left[\hat{r}_i(x) - 1\right] = \frac{1}{2}\sum_{i=1}^{k-1}\mathbb{E}_{p_k(x)}\left[(\hat{r}_i(x) - r_i(x))^2\right] - C \tag{22}$$

where $C = \mathbb{E}_{p_k(x)}\left[\|\boldsymbol{r}(x) - 1\|^2\right]$ is a constant w.r.t. $\hat{\boldsymbol{r}}$ and the minimizer to the above loss function matches the true density ratios, which strictly generalizes the Least-Squares Importance Fitting (LSIF) (Kanamori et al., 2009) method to the multi-distribution case.

Besides, we also consider the following simple convex functions that either strictly generalize their binary DRE counterparts as above, or lead to completely new methods for multi-distribution DRE:

- **Multi-KLIEP.** $f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \sum_{i=1}^{k-1}(\hat{r}_i \log \hat{r}_i - \hat{r}_i) = \langle \hat{\boldsymbol{r}}, \log(\hat{\boldsymbol{r}})\rangle - \|\hat{\boldsymbol{r}}\|_1$. This strictly generalizes the Kullback–Leibler Importance Estimation Procedure (KLIEP) (Sugiyama et al., 2008) to the multi-distribution case. See Appendix A.4.3 for more details.
- **Power.** $f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \sum_{i=1}^{k-1}\hat{r}_i^\alpha = \|\hat{\boldsymbol{r}}\|_\alpha^\alpha$, for $\alpha > 1$. This strictly generalizes the robust DRE method in (Sugiyama et al., 2012), which recovers Multi-KLIEP when $\alpha \to 1$ and Multi-LSIF when $\alpha = 2$. See Appendix A.4.4 for more details.
- **Quadratic.** $f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \hat{\boldsymbol{r}}^\top \boldsymbol{H}\hat{\boldsymbol{r}} + \boldsymbol{q}^\top \hat{\boldsymbol{r}}$, for any positive definite matrix $\boldsymbol{H} \succ 0$. When $\boldsymbol{H}$ is the identity matrix and $\boldsymbol{q} = (-2, \ldots, -2)$, this is equivalent to Multi-LSIF.
- **LogSumExp.** $f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \alpha \log\left(\sum_{i=1}^{k-1}\exp(\hat{r}_i/\alpha)\right)$ for $\alpha > 0$.

In principle, we can use any desired strictly convex function $f : \mathbb{R}_+^{k-1} \to \mathbb{R}$ within the optimization problem in Eq. (14), implying the great potential of our unified framework for discovering novel objectives for multi-distribution DRE. In terms of modeling flexibility, the curvature of different convex functions encode different inductive biases that may favor various downstream applications and we leave the design of more suitable convex functions for DRE as exciting future avenues.

5.2 METHODS INDUCED BY PROPER SCORING RULES COMPOSITE WITH Ψ_{dr}

From Section 4, we know that any strictly proper loss $\ell : [k] \times \Delta_k \to \mathbb{R}$ (or strictly proper scoring rule $S(i, \hat{\boldsymbol{\eta}}) = -\ell(i, \hat{\boldsymbol{\eta}})$) in conjunction with the link function Ψ_{dr} also induces valid losses for multi-distribution DRE:

$$\min_{\hat{\boldsymbol{r}}: \mathcal{X} \to \mathbb{R}_+^{k-1}} \mathbb{E}_{D(x,y)}[\ell(y, \hat{\boldsymbol{\eta}}(x))] = \mathbb{E}_{x \sim M, y \sim \boldsymbol{\eta}(x)}[\ell(y, \Psi_{\mathrm{dr}}^{-1}(\hat{\boldsymbol{r}}(x)))] \tag{23}$$

In this work, we consider using the following classic proper scoring rules (Gneiting & Raftery, 2007), where $\hat{\boldsymbol{\eta}}$ is parametrized as $\Psi_{\mathrm{dr}}^{-1}(\hat{\boldsymbol{r}})$ (i.e. $\hat{\eta}_i = \pi_i \hat{r}_i / \sum_{j=1}^k \pi_j \hat{r}_j$):

- **Logarithm score.** (Good, 1952) The loss is specified as $\ell(i, \hat{\boldsymbol{\eta}}) = -\log(\hat{\eta}_i)$, which also recovers the loss of multi-class logistic regression in Section 5.1.
- **Brier score.** (Brier et al., 1950) The loss is specified as $\ell(i, \hat{\boldsymbol{\eta}}) = -2\hat{\eta}_i + \sum_{j=1}^k \hat{\eta}_j^2 + 1$.
- **Logarithm pseudo-spherical score.** (Good, 1971; Fujisawa & Eguchi, 2008) The loss is specified as $\ell(i, \hat{\boldsymbol{\eta}}) = -\log\left(\frac{\hat{\eta}_i^{\alpha-1}}{(\sum_{j=1}^k \hat{\eta}_j^{\alpha})^{(\alpha-1)/\alpha}}\right)$ for $\alpha > 1$.

6 EXPERIMENTS

In this section, we verify the validity of our framework, as well as study and compare the various instantiations introduced in Section 5, on a variety of tasks that all rely on accurate multi-distribution density ratio estimation. In particular, the tasks we consider include density ratio estimation among multiple multivariate Gaussian distributions, anomaly detection on CIFAR-10 (Krizhevsky et al., 2009), multi-target MNIST Generation (LeCun et al., 1998) and multi-distribution off-policy policy evaluation. We discuss the basic problem setups, evaluation metrics and experimental results in the following and we provide more experimental details for each task in Appendix A.5.

Synthetic Data Experiments. We first apply our methods to estimate density ratios among $k = 5$ multivariate Gaussian distributions with different mean vectors and identity covariance matrix. We conducted experiments for various data dimensions ranging from 2 to 50. Since Gaussian distributions have tractable densities, we know the ground-truth density ratio functions and we calculate the mean absolute error (MAE) between all $\binom{k}{2}$ true density ratios and the learned ones:

$$\mathrm{MAE}(\boldsymbol{r}, \hat{\boldsymbol{r}}; M(x)) = \frac{2}{k(k-1)} \mathbb{E}_{M(x)} \left[\sum_{1 \le i < j \le k} |r_{ij}(x) - \hat{r}_{ij}(x)| \right]$$

where density ratio between p_i and p_j is recovered by $\hat{r}_i / \hat{r}_j$ as discussed in Section 3.2. We summarize the results in Table 1, from which we can see that multi-class logistic regression and Brier score composite with Ψ_{dr} show superior performance in this task.

OOD Detection on CIFAR-10. Suppose we have k different distributions $p_1, \ldots, p_k$, where $p_k = \sum_{i \in [k-1]} \alpha_i p_i$, ($\sum_{i \in [k-1]} \alpha_i = 1$ and $\forall i, \alpha_i > 0$). For each distribution p_i ($i \le k-1$), samples from the mixture distribution p_k contain both inlier samples and outlier samples. The goal of this task is to identify the inlier samples from the pool of mixture samples. In particular, we use the learned density ratio $\hat{r}_i$ as the score function to retrieve the inlier samples of p_i, since the true density ratio function $r_i = p_i / \sum_{j \in [k-1]} \alpha_j p_j$ tend to be larger for samples from p_i and smaller for samples from other distributions. In this case, we calculate the average AUROC for each scoring function.

Multi-target MNIST Generation. DRE can be used in the sampling-importance-resampling (SIR) paradigm (Liu & Chen, 1998; Doucet et al., 2000). Suppose we want to obtain samples from $p_1, \ldots, p_{k-1}$ while we have a large set of samples from p_k. For each $i \in [k-1]$, we can use the density ratio function $\hat{r}_i$ in conjunction with SIR to approximately sample from the target distribution p_i (Algorithm 1 in (Grover et al., 2019)). For this task, we evaluate if the SIR samples for target distribution p_i contains the correct proportion of classes/properties (10 digit numbers in MNIST) and we use $\frac{1}{k-1} \sum_{i=1}^{k-1} \sum_{j=1}^{10} |h_{ij} - \hat{h}_{ij}|$ as the evaluation metric, where h_{ij} and $\hat{h}_{ij}$ denote the desired proportion and sampled proportion for property j in each target-generation task i.

Table 1: Mean absolute error for multi-distribution density ratio estimation among five multivariate Gaussian distributions. Results are averaged across three random seeds.

Method	Dim = 2	Dim = 5	Dim = 10	Dim = 20	Dim = 30	Dim = 40	Dim = 50
Random Init	1.724 ± 0.03	1.723 ± 0.008	1.728 ± 0.02	1.765 ± 0.017	1.749 ± 0.009	1.753 ± 0.002	1.768 ± 0.008
Multi-LR	$\mathbf{0.044} \pm 0.003$	$\mathbf{0.048} \pm 0.005$	$\mathbf{0.061} \pm 0.002$	$\mathbf{0.07} \pm 0.001$	$\mathbf{0.081} \pm 0.002$	$\mathbf{0.089} \pm 0.001$	$\mathbf{0.098} \pm 0.002$
Multi-KLIEP	0.051 ± 0.002	0.066 ± 0.002	0.074 ± 0.0	0.089 ± 0.002	0.105 ± 0.005	0.112 ± 0.004	0.123 ± 0.003
Multi-LSIF	0.073 ± 0.006	0.097 ± 0.001	0.109 ± 0.005	0.124 ± 0.003	0.144 ± 0.004	0.141 ± 0.005	0.158 ± 0.004
Power	0.054 ± 0.003	0.073 ± 0.001	0.081 ± 0.004	0.104 ± 0.003	0.117 ± 0.003	0.123 ± 0.005	0.135 ± 0.004
Brier	$\mathbf{0.042} \pm 0.002$	$\mathbf{0.056} \pm 0.003$	$\mathbf{0.066} \pm 0.003$	$\mathbf{0.081} \pm 0.002$	$\mathbf{0.086} \pm 0.002$	$\mathbf{0.094} \pm 0.002$	$\mathbf{0.105} \pm 0.001$
Spherical	0.103 ± 0.007	0.106 ± 0.006	0.115 ± 0.004	0.121 ± 0.005	0.125 ± 0.006	0.132 ± 0.003	0.138 ± 0.011
LogSumExp	0.231 ± 0.067	0.198 ± 0.034	0.184 ± 0.013	0.179 ± 0.014	0.184 ± 0.009	0.192 ± 0.01	0.193 ± 0.003
Quadratic	0.148 ± 0.033	0.186 ± 0.028	0.218 ± 0.011	0.219 ± 0.018	0.226 ± 0.018	0.236 ± 0.023	0.254 ± 0.014

Table 2: Results for CIFAR-10 OOD detection, MNIST multi-target generation and multi-distribution off-policy policy evaluation error based on learned density ratios. $\uparrow$ means higher is better and $\downarrow$ means lower is better. Results of top 3 methods for each task are bold. Results are averaged across three random seeds.

Method	CIFAR-10 OOD ($\uparrow$)	MNIST Generation ($\downarrow$)	Off-policy Evaluation ($\downarrow$)
Random Init	0.499 ± 0.017	1.598 ± 0.063	1377.68 ± 379.76
Multi-LR	$\mathbf{0.854} \pm 0.009$	$\mathbf{0.156} \pm 0.014$	62.43 ± 12.87
Multi-KLIEP	0.828 ± 0.005	0.281 ± 0.050	110.89 ± 35.33
Multi-LSIF	0.801 ± 0.008	0.274 ± 0.027	71.09 ± 1.12
Power ($\alpha = 1.5$)	0.816 ± 0.007	0.224 ± 0.036	$\mathbf{53.43} \pm 20.73$
Brier	$\mathbf{0.849} \pm 0.010$	$\mathbf{0.107} \pm 0.022$	71.21 ± 17.34
Spherical ($\alpha = 1.8$)	$\mathbf{0.853} \pm 0.010$	$\mathbf{0.145} \pm 0.041$	/
LogSumExp ($\alpha = 5$)	0.782 ± 0.012	/	$\mathbf{52.02} \pm 9.16$
Quadratic	0.804 ± 0.009	/	$\mathbf{55.10} \pm 11.92$

Multi-distribution Off-policy Policy Evaluation. Suppose we have k different reinforcement learning policies $p_i(a|s)$, each inducing an occupancy measure (Syed et al., 2008) (i.e, state-action distribution) $\rho_i(s, a)$. Density ratios allow us to conduct off-policy policy evaluation, which estimates the expected return (sum of reward) of target policies $p_1, \ldots, p_{k-1}$ given trajectories sampled from a source policy p_k. In this case, we evaluate the following metric to assess the quality of the learned density ratios ($\tau = \{(s_t, a_t)\}_{t=1}^T$ denotes a sequence of state-action pairs):

$$\frac{1}{k-1} \sum_{i=1}^{k-1} \left| \mathbb{E}_{p_k(\tau)} \left[\sum_{t=1}^{T} \hat{r}_i(s_t, a_t) r(s_t, a_t) \right] - \mathbb{E}_{p_i(\tau)} \left[\sum_{t=1}^{T} r(s_t, a_t) \right] \right|$$

We summarized the results for CIFAR-10 OOD detection, multi-target MNIST generation and multi-distribution off-policy policy evaluation in Table 2 (omitted results indicates the corresponding method performs worse than listed methods by a large margin on the specific task). We can see that methods induced by proper scoring rules such as multi-class logistic regression, Brier score and pseudo-spherical score tend to have the best performance on the first two tasks. And surprisingly, methods induced by some simple multivariate convex functions such as the LogSumExp and the quadratic function show excellent performance on the third task. These results demonstrate the advantage of our framework in the sense that it offers us extreme flexibility for designing new objectives for multi-distribution DRE that are more suitable for various downstream applications.

7 Conclusion

In this paper, we focus on the generalized problem of efficiently estimating density ratios among multiple distributions. We propose a general framework based on expected Bregmand divergence minimization, where each strictly convex function induces a proper loss for multi-distribution DRE. Furthermore, we theoretically prove an equivalence between the problem of class probability estimation and density ratio estimation in the context of multi-class experiments, generalizing previous results in binary case and justifying the use of any strictly proper scoring rules for multi-distribution DRE. Finally, we demonstrated the value of our framework on various downstream tasks.

REFERENCES

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*, 2019.

Steffen Bickel, Jasmina Bogojeska, Thomas Lengauer, and Tobias Scheffer. Multi-task learning for hiv therapy screening. In *Proceedings of the 25th international conference on Machine learning*, pp. 56–63, 2008.

Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Jiezhang Cao, Langyuan Mo, Yifan Zhang, Kui Jia, Chunhua Shen, and Mingkui Tan. Multi-marginal wasserstein gan. *Advances in Neural Information Processing Systems*, 32:1776–1786, 2019.

Olivier Cappé, Arnaud Guillin, Jean-Michel Marin, and Christian P Robert. Population monte carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, 2004.

Kristy Choi, Madeline Liao, and Stefano Ermon. Featurized density ratio estimation. *arXiv preprint arXiv:2107.02212*, 2021.

Morris H DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

Cuong V Dinh, Robert PW Duin, Ignacio Piqueras-Salazar, and Marco Loog. Fidos: A generalized fisher based feature extraction method for domain shift. *Pattern Recognition*, 46(9):2510–2518, 2013.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1723–1732, 2015.

Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.

John Duchi, Khashayar Khosravi, and Feng Ruan. Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics*, 46(6B):3246–3275, 2018.

Víctor Elvira, Luca Martino, David Luengo, and Mónica F Bugallo. Efficient multiple importance sampling estimators. *IEEE Signal Processing Letters*, 22(10):1757–1761, 2015.

Víctor Elvira, Luca Martino, David Luengo, and Mónica F Bugallo. Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155, 2019.

James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

Dario Garcia-Garcia and Robert C Williamson. Divergences and risks for multiclass experiments. In *Conference on Learning Theory*, pp. 28–1. JMLR Workshop and Conference Proceedings, 2012.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

IJ Good. Rational decisions. *Journal of the Royal Statistical Society*, pp. 107–114, 1952.

IJ Good. Comment on "measuring information and uncertainty" by robert j. buehler. *Foundations of Statistical Inference*, pp. 337–339, 1971.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Aditya Grover, Jiaming Song, Alekh Agarwal, Kenneth Tran, Ashish Kapoor, Eric Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. *arXiv preprint arXiv:1906.09531*, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Inlier-based outlier detection via direct density ratio estimation. In *2008 Eighth IEEE international conference on data mining*, pp. 223–232. IEEE, 2008.

Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and information systems*, 26(2):309–336, 2011.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19:601–608, 2006.

Nathan Kallus, Yuta Saito, and Masatoshi Uehara. Optimal off-policy evaluation from multiple logging policies. In *International Conference on Machine Learning*, pp. 5247–5256. PMLR, 2021.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012.

Masahiro Kato and Takeshi Teshima. Non-negative bregman divergence minimization for deep direct density ratio estimation. In *International Conference on Machine Learning*, pp. 5320–5333. PMLR, 2021.

Samory Kpotufe. Lipschitz density-ratios, structured data, and data-driven tuning. In *Artificial Intelligence and Statistics*, pp. 1320–1328. PMLR, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jun S Liu and Rong Chen. Sequential monte carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443):1032–1044, 1998.

Song Liu, Akiko Takeda, Taiji Suzuki, and Kenji Fukumizu. Trimmed density ratio estimation. *arXiv preprint arXiv:1703.03216*, 2017.

Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pp. 304–313. PMLR, 2016.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *NIPS*, pp. 1089–1096, 2007.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 271–279, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

Qichao Que and Mikhail Belkin. Inverse density as an inverse problem: The fredholm equation approach. *arXiv preprint arXiv:1304.5575*, 2013.

Benjamin Rhodes, Kai Xu, and Michael U Gutmann. Telescoping density-ratio estimation. *arXiv preprint arXiv:2006.12204*, 2020.

Robin Sibson. Information radius. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(2):149–160, 1969.

Alex Smola, Le Song, and Choon Hui Teo. Relative novelty detection. In *Artificial Intelligence and Statistics*, pp. 536–543. PMLR, 2009.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Von Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, volume 7, pp. 1433–1440. Citeseer, 2007.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Masashi Sugiyama, Taiji Suzuki, Yuta Itoh, Takafumi Kanamori, and Manabu Kimura. Least-squares two-sample test. *Neural networks*, 24(7):735–751, 2011.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

Umar Syed, Michael Bowling, and Robert E Schapire. Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning*, pp. 1032–1039, 2008.

Masatoshi Uehara, Issei Sato, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Generative adversarial nets from a density ratio estimation perspective. *arXiv preprint arXiv:1610.02920*, 2016.

Elodie Vernet, Robert Williamson, Mark Reid, et al. Composite multiclass losses. 2011.

A PROOFS

A.1 PROOFS FOR SECTION 2

Theorem 1 ((Gneiting & Raftery, 2007); Proposition 7 in (Vernet et al., 2011))**.** *Given a proper loss ℓ and the corresponding expected loss L, for any $P, Q \in \Delta_k$, the generalized entropy function $\underline{L}(P) := \inf_{Q \in \Delta_k} L(P, Q) = L(P, P)$ is concave; when $\underline{L}$ is differentiable, the regret or excess risk of a predictor Q over the Bayes-optimal P is the Bregman divergence induced by the convex function $f = -\underline{L}$:*

$$\mathrm{reg}(P, Q; \ell) := L(P, Q) - L(P, P) = \mathbf{B}_f(P, Q) \tag{3}$$

Proof. For completeness, we provide the proof here. First, we can check that $\underline{L}(P) : \Delta_k \to \mathbb{R}$ is a concave function. Define $\mathbf{L}(P)$ to be the vector $(\ell(1, P), \ldots, \ell(k, P))$. Then the entropy function can be represented as $\underline{L}(P) = L(P, P) = \mathbb{E}_{y \sim P}[\ell(y, P)] = P^\top \mathbf{L}(P)$ and similarly $L(P, Q) = P^\top \mathbf{L}(Q)$. For $\lambda \in [0, 1]$ and $P, Q \in \Delta_k$, we have:

$$\begin{aligned}
\underline{L}(\lambda P + (1-\lambda)Q) &= (\lambda P + (1-\lambda)Q)^\top \mathbf{L}(\lambda P + (1-\lambda)Q) \\
&= \lambda P^\top \mathbf{L}(\lambda P + (1-\lambda)Q) + (1-\lambda) Q^\top \mathbf{L}(\lambda P + (1-\lambda)Q) \\
&\geq \lambda P^\top \mathbf{L}(P) + (1-\lambda) Q^\top \mathbf{L}(Q) = \lambda \underline{L}(P) + (1-\lambda)\underline{L}(Q)
\end{aligned}$$

where the inequality is because ℓ is proper. Thus $\underline{L}$ is a concave function. Next we show that the excess risk is a Bregman divergence with convex function $-\underline{L}$. First, observe that

$$L(P, Q) = P^\top \mathbf{L}(Q) = Q^\top \mathbf{L}(Q) + (P - Q)^\top \mathbf{L}(Q)$$

Because ℓ is proper, we have:

$$\begin{aligned}
0 \leq L(P, Q) - L(P, P) &= Q^\top \mathbf{L}(Q) + (P - Q)^\top \mathbf{L}(Q) - P^\top \mathbf{L}(P) \\
&= -\underline{L}(P) - (-\underline{L}(Q)) - (P - Q)^\top (-\mathbf{L}(Q))
\end{aligned}$$

Rearrange the term we get $-\underline{L}(P) \geq (-\underline{L}(Q)) + (-\mathbf{L}(Q))^\top (P - Q)$ and therefore $-\mathbf{L}(Q)$ is a subderivative of $-\underline{L}$. When $-\underline{L}$ is differentiable, its subdifferential contains exactly one subderivative and $-\mathbf{L}(Q) = \nabla(-\underline{L}(Q))$. Therefore, we have $\mathrm{reg}(P, Q) = L(P, Q) - L(P, P) = f(P) - f(Q) - \langle \nabla f(Q), P - Q \rangle = \mathbf{B}_f(P, Q)$ with $f = -\underline{L}$. □

A.2 PROOFS FOR SECTION 3.2

Proposition 1 (DRE via variational estimation of multi-distribution f-divergence)**.** *Given a strictly convex function $f : \mathbb{R}_+^{k-1} \to \mathbb{R} \cup \{+\infty\}$, the optimization problem in Eq. (14) (induced by minimizing expected Bregman divergence $\mathbf{B}_f(\boldsymbol{r}, \hat{\boldsymbol{r}})$) is equivalent to the one in Eq. (16) (for variational estimation of multi-distribution f-divergence) under change of variables satisfying: $\nabla f(\hat{\boldsymbol{r}}(x)) = \boldsymbol{s}(x)$, $\forall x \in \mathcal{X}$.*

Proof of Proposition 1. We first recall that the optimization problem for multi-distribution DRE is of the form

$$\min_{\hat{\boldsymbol{r}}: \mathcal{X} \to \mathbb{R}_+^{k-1}} \mathbb{E}_{p_k(x)} \left[\langle \nabla f(\hat{\boldsymbol{r}}(x)), \hat{\boldsymbol{r}}(x) \rangle - f(\hat{\boldsymbol{r}}(x)) \right] - \sum_{i \in [k-1]} \mathbb{E}_{p_i(x)} [\partial_i f(\hat{\boldsymbol{r}}(x))] \tag{24}$$

and one can use the Fenchel-dual convex conjugate to represent the f-divergence as

$$\mathbf{D}_f(P_1, \cdots, P_{k-1} || P_k) = - \min_{\boldsymbol{s}: \mathcal{X} \to \mathbb{R}^{k-1}} \left[- \sum_{i \in [k-1]} \mathbb{E}_{p_i(x)} [\boldsymbol{s}(x)]_i + \mathbb{E}_{p_k(x)} f^*(\boldsymbol{s}(x)) \right] \tag{25}$$

By first-order optimality condition of convex functions, for any $x \in \mathcal{X}$ the optimal solution $\overline{\boldsymbol{s}}(x)$ for Eq. (25) satisfies that

$$\forall i \in [k-1], x \in \mathcal{X}, \;\; p_i(x) - \partial_i f^*(\overline{\boldsymbol{s}}(x)) p_k(x) = 0 \iff \frac{p_i(x)}{p_k(x)} = \partial_i f^*(\overline{\boldsymbol{s}}(x))$$

Therefore $\overline{\boldsymbol{r}}(x) = \nabla f^*(\overline{\boldsymbol{s}}(x))$ recovers the true density ratios.

Now we show that under change of variable $\boldsymbol{s}(x) = \nabla f(\hat{\boldsymbol{r}}(x))$, one can write the problem in Eq. (25) equivalently as the one in Eq. (24). First due to the property of the convex conjugate function ($f^{**} = f$), we have:

$$f^*(\boldsymbol{s}(x)) = \min_{\boldsymbol{h}(x)\in\mathbb{R}^{k-1}} \langle \boldsymbol{s}(x), \boldsymbol{h}(x)\rangle - f(\boldsymbol{h}(x))$$

Substituting $\boldsymbol{s}(x)$ with $\nabla f(\hat{\boldsymbol{r}}(x))$, we have:

$$f^*(\boldsymbol{s}(x)) = \min_{\boldsymbol{h}(x)\in\mathbb{R}^{k-1}} \langle \nabla f(\hat{\boldsymbol{r}}(x)), \boldsymbol{h}(x)\rangle - f(\boldsymbol{h}(x)) \tag{26}$$

Taking derivative w.r.t. $\boldsymbol{h}$ and due to the strict convexity of f ($\nabla f(\boldsymbol{a}) = \nabla f(\boldsymbol{b}) \Leftrightarrow \boldsymbol{a} = \boldsymbol{b}$), we know that the minimum of Eq. (26) achieves at $\overline{\boldsymbol{h}}(x) = \hat{\boldsymbol{r}}(x)$. Thus we have:

$$f^*(\boldsymbol{s}(x)) = \langle \nabla f(\hat{\boldsymbol{r}}(x)), \hat{\boldsymbol{r}}(x)\rangle - f(\hat{\boldsymbol{r}}(x)) \tag{27}$$

Plugging Eq. (27) and $\boldsymbol{s}(x) = \nabla f(\hat{\boldsymbol{r}}(x))$ back to the optimization problem in Eq. (25), we can get the following equivalent problem by flipping a sign of the objective function without changing the optimal solution:

$$\min_{\hat{\boldsymbol{r}}:\mathcal{X}\to\mathbb{R}^{k-1}} \mathbb{E}_{p_k(x)}\left[\langle \nabla f(\hat{\boldsymbol{r}}(x)), \hat{\boldsymbol{r}}(x)\rangle - f(\hat{\boldsymbol{r}}(x))\right] - \sum_{i\in[k-1]} \mathbb{E}_{p_i(x)} \partial_i f(\hat{\boldsymbol{r}}(x)),$$

which is the same as the one in (24). □

A.3 Proofs for Section 4

Lemma 1 (Multivariate Bregman Identity)**.** *Given a convex function $f : \mathbb{R}^{k-1} \to \mathbb{R}$, we can define an associated function $f^{\circledast}(u_1, \ldots, u_{k-1}) = (1 + \sum_{i\in[k-1]} u_i) f\left(\frac{1}{1+\sum_{i\in[k-1]} u_i} \cdot \boldsymbol{u}\right)$. We can show that (i) $f^{\circledast}$ is convex and (ii) for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{k-1}$, their associated Bregman divergences satisfy:*

$$\mathbf{B}_f\left(\frac{1}{1+\sum_{i\in[k-1]} u_i} \cdot \boldsymbol{u}, \frac{1}{1+\sum_{i\in[k-1]} v_i} \cdot \boldsymbol{v}\right) = \frac{1}{1+\sum_{i\in[k-1]} u_i} \mathbf{B}_{f^{\circledast}}(\boldsymbol{u}, \boldsymbol{v}). \tag{17}$$

Proof of Lemma 1. For simplicity of notations we let $u_k = v_k = 1$ for arbitrary $u, v \in \mathbb{R}^{k-1}$. We first prove the convexity of $f^{\circledast}$ by definition. Given any two points $u, v \in \mathbb{R}^{k-1}$ and $\theta \in [0, 1]$, one has

$$\begin{aligned}
&f^{\circledast}(\theta\boldsymbol{u} + (1-\theta)\boldsymbol{v}) \\
=&\left(\sum_{i\in[k]} (\theta u_i + (1-\theta)v_i)\right) \cdot f\left(\frac{1}{\sum_{i\in[k]}(\theta u_i + (1-\theta)v_i)} \cdot (\theta\boldsymbol{u} + (1-\theta)\boldsymbol{v})\right) \\
=&\left(\theta \sum_{i\in[k]} u_i + (1-\theta)\sum_{i\in[k]} v_i\right) \cdot f\left(\frac{1}{\theta\sum_{i\in[k]} u_i + (1-\theta)\sum_{i\in[k]} v_i} \cdot (\theta\boldsymbol{u} + (1-\theta)\boldsymbol{v})\right) \\
\overset{(\star)}{\leq}&\theta\left(\sum_{i\in[k]} u_i\right) f\left(\frac{1}{\sum_{i\in[k]} u_i}\boldsymbol{u}\right) + (1-\theta)\left(\sum_{i\in[k]} v_i\right) f\left(\frac{1}{\sum_{i\in[k]} v_i}\boldsymbol{v}\right) \\
=&\theta f^{\circledast}(\boldsymbol{u}) + (1-\theta) f^{\circledast}(\boldsymbol{v}).
\end{aligned}$$

Here for inequality $(\star)$ we use the fact that for any convex function $g : \mathbb{R}^n \to \mathbb{R}$, the perspective function $h(t, x) : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ defined as $h(t, x) := tg(x/t)$ is a function jointly convex in (t, x).

Now to see the identity holds, note we can write

$$\begin{aligned}
\text{LHS} =& f\left(\frac{1}{\sum_{i\in[k]} u_i} \cdot \boldsymbol{u}\right) - f\left(\frac{1}{\sum_{i\in[k]} v_i} \cdot \boldsymbol{v}\right) \\
&- \left\langle \nabla f\left(\frac{1}{\sum_{i\in[k]} v_i} \cdot \boldsymbol{v}\right), \frac{1}{\sum_{i\in[k]} u_i} \cdot \boldsymbol{u} - \frac{1}{\sum_{i\in[k]} v_i} \cdot \boldsymbol{v}\right\rangle
\end{aligned}$$

and that

$$
\begin{aligned}
\text{RHS} =& f\left(\frac{1}{\sum_{i\in[k]} u_i}\cdot \boldsymbol{u}\right) - \frac{\sum_{i\in[k]} v_i}{\sum_{i\in[k]} u_i}\cdot f\left(\frac{1}{\sum_{i\in[k]} v_i}\cdot \boldsymbol{v}\right) - \frac{1}{\sum_{i\in[k]} u_i}\left\langle \nabla f^{\circledast}(\boldsymbol{v}), \boldsymbol{u}-\boldsymbol{v}\right\rangle \\
\overset{(i)}{=}& f\left(\frac{1}{\sum_{i\in[k]} u_i}\cdot \boldsymbol{u}\right) - \frac{\sum_{i\in[k]} v_i}{\sum_{i\in[k]} u_i}\cdot f\left(\frac{1}{\sum_{i\in[k]} v_i}\cdot \boldsymbol{v}\right) \\
& - \frac{1}{\sum_{i\in[k]} u_i}\left\langle f\left(\frac{1}{\sum_{i\in[k]} v_i}\cdot \boldsymbol{v}\right)\mathbf{1} + \left(\mathbf{I} - \frac{1}{\sum_{i\in[k]} v_i}\mathbf{1}\boldsymbol{v}^\top\right)\nabla f\left(\frac{1}{\sum_{i\in[k]} v_i}\cdot \boldsymbol{v}\right), \boldsymbol{u}-\boldsymbol{v}\right\rangle \\
\overset{(ii)}{=}& f\left(\frac{1}{\sum_{i\in[k]} u_i}\cdot \boldsymbol{u}\right) - \left[\frac{\sum_{i\in[k]} v_i}{\sum_{i\in[k]} v_i} + \frac{\sum_{i\in[k]}(u_i - v_i)}{\sum_{i\in[k]} u_i}\right]\cdot f\left(\frac{1}{\sum_{i\in[k]} v_i}\cdot \boldsymbol{v}\right) \\
& + \frac{1}{\sum_{i\in[k]} u_i}\left\langle \nabla f\left(\frac{1}{\sum_{i\in[k]} v_i}\cdot \boldsymbol{v}\right), \left(\mathbf{I} - \frac{1}{\sum_{i\in[k]} v_i}\boldsymbol{v}\mathbf{1}^\top\right)(\boldsymbol{u}-\boldsymbol{v})\right\rangle \\
=& f\left(\frac{1}{\sum_{i\in[k]} u_i}\cdot \boldsymbol{u}\right) - f\left(\frac{1}{\sum_{i\in[k]} v_i}\cdot \boldsymbol{v}\right) \\
& + \left\langle \nabla f\left(\frac{1}{\sum_{i\in[k]} v_i}\cdot \boldsymbol{v}\right), \frac{1}{\sum_{i\in[k]} u_i}\cdot \boldsymbol{u} - \frac{1}{\sum_{i\in[k]} u_i}\cdot \boldsymbol{v} - \frac{\sum_{i\in[k]}(u_i - v_i)}{\left(\sum_{i\in[k]} u_i\right)\left(\sum_{i\in[k]} v_i\right)}\cdot \boldsymbol{v}\right\rangle \\
=& f\left(\frac{1}{\sum_{i\in[k]} u_i}\cdot \boldsymbol{u}\right) - f\left(\frac{1}{\sum_{i\in[k]} v_i}\cdot \boldsymbol{v}\right) \\
& + \left\langle \nabla f\left(\frac{1}{\sum_{i\in[k]} v_i}\cdot \boldsymbol{v}\right), \frac{1}{\sum_{i\in[k]} u_i}\cdot \boldsymbol{u} - \frac{1}{\sum_{i\in[k]} v_i}\cdot \boldsymbol{v}\right\rangle,
\end{aligned}
$$

where we use (i) the gradient formula that $\nabla f^{\circledast}(\boldsymbol{v}) = \left(\mathbf{I} - \frac{1}{\sum_{i\in[k]} v_i}\mathbf{1}\boldsymbol{v}^\top\right)\nabla f\left(\frac{1}{\sum_{i\in[k]} v_i}\cdot \boldsymbol{v}\right)$ by definition of $f^{\circledast}$, and (ii) rearranging terms and that $\langle \boldsymbol{A}\boldsymbol{v}, \boldsymbol{u}\rangle = \langle \boldsymbol{u}, \boldsymbol{A}^\top \boldsymbol{v}\rangle$.

Thus, we have shown that LHS $=$ RHS and concludes the proof. □

Proposition 2. *For any convex function $f : \mathbb{R}_+^{k-1} \to \mathbb{R}$, and two density ratio vectors $\boldsymbol{r}(x)$ and $\hat{\boldsymbol{r}}(x)$, one can construct corresponding class probability vectors $\boldsymbol{\eta}(x) = \Psi_{\mathrm{dr}}^{-1}(\boldsymbol{r}(x))$ and $\hat{\boldsymbol{\eta}}(x) = \Psi_{\mathrm{dr}}^{-1}(\hat{\boldsymbol{r}}(x))$ through the inverse link function in Eq. (8), and obtain:*

$$
\mathbf{B}_f\left(\boldsymbol{\eta}(x), \hat{\boldsymbol{\eta}}(x)\right) = \frac{\pi_k}{\pi_k + \sum_{i\in[k-1]} \pi_i r_i(x)}\mathbf{B}_{f_\pi^{\circledast}}\left(\boldsymbol{r}(x), \hat{\boldsymbol{r}}(x)\right) \text{ for all } x \in \mathcal{X}, \tag{18}
$$

where we define the convex function $f_\pi^{\circledast}$ induced by some prior distribution $\pi \in \Delta_k$ as

$$
f_\pi^{\circledast}(r_1, \ldots, r_{k-1}) := \left(1 + \sum_{i\in[k-1]} \pi_i r_i/\pi_k\right)\cdot f\left(\frac{\boldsymbol{\pi}_{[1:k-1]} \circ \boldsymbol{r}}{\pi_k + \sum_{i\in[k-1]} \pi_i r_i}\right). \tag{19}
$$

Proof of Proposition 2. Given any $x \in \mathcal{X}$, the equality follows by applying Lemma 1 with $\boldsymbol{u} = \frac{1}{\pi_k}\boldsymbol{\pi}_{[1:k-1]} \circ \boldsymbol{r}(x)$ and $\boldsymbol{v} = \frac{1}{\pi_k}\boldsymbol{\pi}_{[1:k-1]} \circ \hat{\boldsymbol{r}}(x)$. To see why this is true, note that we have by definition of $\eta_i(x)$ and $\hat{\eta}_i(x)$ that (here $\circ$ implies element-wise multiplication)

$$
\boldsymbol{\eta}(x) = \frac{\boldsymbol{\pi} \circ \boldsymbol{r}(x)}{\pi_k + \sum_{j\in[k-1]} \pi_j r_j(x)} = \frac{1}{1 + \sum_{i\in[k-1]} u_i}\cdot \boldsymbol{u}, \text{ and similarly } \hat{\boldsymbol{\eta}}(\boldsymbol{x}) = \frac{1}{1 + \sum_{i\in[k-1]} v_i}\cdot \boldsymbol{v}.
$$

Consequently applying Lemma 1 implies that

$$
\mathbf{B}_f(\boldsymbol{\eta}(x), \hat{\boldsymbol{\eta}}(x)) = \frac{1}{1 + \sum_{i\in[k-1]} u_i}\mathbf{B}_{f^{\circledast}}(\boldsymbol{u}, \boldsymbol{v}) \tag{28}
$$

Note that given any convex function $f^{\circledast}$, we consider its composition with linear map function as

$$f_{\pi}^{\circledast}(\boldsymbol{r}) = f^{\circledast}\left(\frac{1}{\pi_k}\boldsymbol{\pi}_{[1:k-1]} \circ \boldsymbol{r}\right) = f^{\circledast}(\boldsymbol{u}).$$

We note that linear composition preserves convexity and Bregman divergence equality, i.e. we have

$$\begin{aligned}
&\mathbf{B}_{f^{\circledast}}(\boldsymbol{u}, \boldsymbol{v}) = f^{\circledast}(\boldsymbol{u}) - f^{\circledast}(\boldsymbol{v}) - \langle \nabla f^{\circledast}(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v} \rangle \\
=& f^{\circledast}\left(\frac{1}{\pi_k}\boldsymbol{\pi}_{1:k-1} \circ \boldsymbol{r}\right) - f^{\circledast}\left(\frac{1}{\pi_k}\boldsymbol{\pi}_{1:k-1} \circ \hat{\boldsymbol{r}}\right) - \left\langle \nabla f^{\circledast}\left(\frac{1}{\pi_k}\boldsymbol{\pi}_{1:k-1} \circ \hat{\boldsymbol{r}}\right), \frac{1}{\pi_k}\boldsymbol{\pi}_{1:k-1} \circ (\boldsymbol{r} - \hat{\boldsymbol{r}}) \right\rangle \\
\overset{(\star)}{=}& f_{\pi}^{\circledast}(\boldsymbol{r}) - f_{\pi}^{\circledast}(\hat{\boldsymbol{r}}) - \langle \nabla f_{\pi}^{\circledast}(\boldsymbol{r}), \boldsymbol{r} - \hat{\boldsymbol{r}} \rangle = \mathbf{B}_{f_{\pi}^{\circledast}}(\boldsymbol{r}, \hat{\boldsymbol{r}}),
\end{aligned} \tag{29}$$

where for equality $(\star)$ we use chain rule for taking derivatives of the linear composite mapping. Combining Equations (28) and (29) and replacing $\boldsymbol{u} = \frac{1}{\pi_k}\boldsymbol{\pi}_{[1:k-1]} \circ \boldsymbol{r}$ gives the desired result. □

Theorem 2. *Given any strictly proper loss ℓ, for any joint data distribution $D(X, Y)$ with class prior $\pi \in \Delta_k$, the multi-class classification regret defined in Eq. (4) satisfies that:*

$$\operatorname{reg}(\hat{\boldsymbol{\eta}}; M, \boldsymbol{\eta}, \ell) = \pi_k \mathbb{E}_{p_k(x)} \mathbf{B}_{f_{\pi}^{\circledast}}(\boldsymbol{r}(x), \hat{\boldsymbol{r}}(x)), \tag{20}$$

for $f_{\pi}^{\circledast}$ as defined in Eq. (19), and $\boldsymbol{r} = \Psi_{\mathrm{dr}}(\boldsymbol{\eta})$ and $\hat{\boldsymbol{r}} = \Psi_{\mathrm{dr}}(\hat{\boldsymbol{\eta}})$ as defined in Eq. (7).

Proof of Theorem 2. Given the multi-class classification regret under some proper loss ℓ in Eq. (4) and Proposition 2 we have:

$$\begin{aligned}
\operatorname{reg}(\hat{\boldsymbol{\eta}}; M, \boldsymbol{\eta}, \ell) &:= \mathcal{L}_{\mathrm{CPE}}(\hat{\boldsymbol{\eta}}; D) - \mathcal{L}_{\mathrm{CPE}}(\boldsymbol{\eta}; D) = \mathbb{E}_{M(x)}[\mathbf{B}_f(\boldsymbol{\eta}(x), \hat{\boldsymbol{\eta}}(x))] \\
&\overset{(i)}{=} \sum_{i \in [k]} \pi_i \mathbb{E}_{p_i(x)} \mathbf{B}_f(\boldsymbol{\eta}(x), \hat{\boldsymbol{\eta}}(x)) = \mathbb{E}_{p_k(x)} \left(\sum_{i \in [k]} \pi_i \frac{p_i(x)}{p_k(x)} \mathbf{B}_f(\boldsymbol{\eta}(x), \hat{\boldsymbol{\eta}}(x)) \right) \\
&\overset{(ii)}{=} \mathbb{E}_{p_k(x)} \left(\left(\pi_k + \sum_{i \in [k-1]} \pi_i r_i(x) \right) \cdot \mathbf{B}_f(\boldsymbol{\eta}(x), \hat{\boldsymbol{\eta}}(x)) \right) \\
&\overset{(iii)}{=} \pi_k \mathbb{E}_{p_k(x)} \mathbf{B}_{f_{\pi}^{\circledast}}(\boldsymbol{r}(x), \hat{\boldsymbol{r}}(x)),
\end{aligned}$$

where we use (i) the definition of marginal distribution $M(x) = \sum_{i \in [k]} \pi_i p_i(x)$, (ii) the definition of density ratio that $r_i(x) = p_i(x)/p_k(x)\ \forall x \in \mathcal{X}, i \in [k]$, and (iii) Proposition 2 with the consistent definitions of $f_{\pi}^{\circledast}$ and $\boldsymbol{r}$, $\hat{\boldsymbol{r}}$ as stated in the theorem. □

A.3.1 Information Measure in Multi-class Experiments

In this section, we show that multi-distribution density ratio estimation can be viewed as estimating the statistical information measure (DeGroot, 1962) in multi-class experiments, under appropriate choices for the convex function f.

We first introduce the following definitions in multi-class experiments. For $\boldsymbol{p} \in \Delta_k$, any proper loss function $\ell : [k] \times \Delta^k \to \mathbb{R}$ induces a generalized entropy:

$$H_{\ell}(\boldsymbol{p}) := \inf_{\boldsymbol{q} \in \Delta^k} \sum_{i \in [k]} p_i \ell(i, \boldsymbol{q}),$$

which measures the uncertainty of the task. Given a multi-class experiment $D = (\boldsymbol{\pi}, P_1, \ldots, P_k) = (M, \boldsymbol{\eta})$ and the generalized entropy $H_{\ell} : \Delta^k \to \mathbb{R}$ (which is closed concave), the information measure in a multi-class experiment (DeGroot, 1962; Duchi et al., 2018) is defined as the gap between the prior and posterior generalized entropy:

$$\mathcal{I}_{H_{\ell}}(D) = H_{\ell}(\boldsymbol{\pi}) - \mathbb{E}_{M(x)}[H_{\ell}(\boldsymbol{\eta}(x))].$$

We next introduce the following connections between multi-distribution f-divergence, generalized entropy and information measure in multi-class experiments. Specifically, Duchi et al. (2018) proved an equivalence between the gap of prior and posterior Bayes risks and the multi-distribution f-divergence induced by a convex function f depending on ℓ and the prior $\boldsymbol{\pi}$, demonstrating the utility of multi-distribution f-divergence for experimental design of multi-class classification.

Theorem 3 ((Duchi et al., 2018)). *Given a proper loss ℓ, its associated generalized entropy H_ℓ and a multi-class distribution $D = (\boldsymbol{\pi}, P_1, \ldots, P_k) = (M, \boldsymbol{\eta})$, we can define a closed convex function $f_{\ell,\boldsymbol{\pi}} : \mathbb{R}_+^{k-1} \to \mathbb{R} \cup \{\pm\infty\}$ as*

$$f_{\ell,\boldsymbol{\pi}}(\boldsymbol{t}) := \sup_{\boldsymbol{\nu} \in \Delta^k} \left(H_\ell(\boldsymbol{\pi}) - \sum_{i \in [k-1]} \pi_i \ell(i, \boldsymbol{\nu}) t_i - \pi_k \ell(k, \boldsymbol{\nu}) \right) \tag{30}$$

We can then express the information measure of multi-class experiments as the multi-distribution f-divergence induced by Eq. (30):

$$\begin{aligned} \mathcal{I}_{H_\ell}(D) &= H(\boldsymbol{\pi}) - \mathbb{E}_{M(x)}[H_\ell(\boldsymbol{\eta}(x))] = \inf_{\boldsymbol{\nu} \in \Delta^k} \sum_{i \in [k]} \pi_i \ell(i, \boldsymbol{\nu}) - \inf_{\hat{\boldsymbol{\eta}}} \mathcal{L}(\hat{\boldsymbol{\eta}}; D) \\ &= \mathbf{D}_{f_{\ell,\boldsymbol{\pi}}}(P_1, \ldots, P_{k-1} \| P_k). \end{aligned}$$

Given Theorem 3 and Proposition 1, we know that multi-distribution density ratio estimation by minimizing expected Bregman divergence (Eq. (14)), induced by the convex function $f_{\ell,\boldsymbol{\pi}}$ defined in Eq. (30), corresponds to estimating the statistical information measure in multi-class classification experiments.

A.4 Examples of Multi-distribution DRE

A.4.1 Multi-class Logistic Regression

From Section 2.3, we know that there is a one-to-one correspondence between a class probability estimator and a density ratio estimator through the link and the inverse link function: $\hat{\boldsymbol{r}} = \Psi_{\mathrm{dr}} \circ \hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\eta}} = \Psi_{\mathrm{dr}}^{-1} \circ \hat{\boldsymbol{r}}$. When the class prior distribution $\boldsymbol{\pi}$ is uniform, we have:

$$\hat{r}_i(x) = \frac{\hat{\eta}_i(x)}{\hat{\eta}_k(x)} \;\text{ and }\; \hat{\eta}_i(x) = \frac{\hat{r}_i(x)}{\sum_{j=1}^k \hat{r}_j(x)}, \;\text{ for all } i \in [k], x \in \mathcal{X}. \tag{31}$$

To recover the loss of multi-class logistic regression used in (Bickel et al., 2008), we choose the following convex function (where we use $\hat{r}_k = 1$):

$$f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \frac{1}{k} \sum_{i=1}^k \hat{r}_i \log \left(\frac{\hat{r}_i}{\sum_{j=1}^k \hat{r}_j} \right) \tag{32}$$

$$\partial_i f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \frac{1}{k} \log \left(\frac{\hat{r}_i}{\sum_{j=1}^k \hat{r}_j} \right) \quad \text{for } i \in [k-1] \tag{33}$$

Thus the loss in Eq. (14) reduces to:

$$\begin{aligned} &\frac{1}{k} \mathbb{E}_{p_k(x)} \left[\sum_{i=1}^{k-1} \hat{r}_i(x) \log \frac{\hat{r}_i(x)}{\sum_{j=1}^k \hat{r}_j(x)} - \sum_{i=1}^{k-1} \hat{r}_i(x) \log \hat{r}_i(x) + \left(\sum_{i=1}^k \hat{r}_i(x) \right) \log \left(\sum_{i=1}^k \hat{r}_i(x) \right) \right] - \\ &\frac{1}{k} \sum_{i=1}^{k-1} \mathbb{E}_{p_i(x)} \left[\log \left(\frac{\hat{r}_i(x)}{\sum_{j=1}^k \hat{r}_j(x)} \right) \right] \\ =& \frac{1}{k} \mathbb{E}_{p_k(x)} \left[\hat{r}_k(x) \log \left(\sum_{i=1}^k \hat{r}_i(x) \right) \right] - \frac{1}{k} \sum_{i=1}^{k-1} \mathbb{E}_{p_i(x)} \left[\log \left(\frac{\hat{r}_i(x)}{\sum_{j=1}^k \hat{r}_j(x)} \right) \right] \\ \overset{(i)}{=} & - \left(\frac{1}{k} \sum_{i=1}^k \mathbb{E}_{p_i(x)} [\log(\hat{\boldsymbol{\eta}}_i(x))] \right) \end{aligned}$$

where (i) is because $\hat{r}_k(x) = 1, \; \forall x \in \mathcal{X}$ and Eq. (31).

When the class prior $\boldsymbol{\pi}$ is not uniform, from Section 2.3, we know that the link and inverse link connecting density ratio estimators and class probability estimators are:

$$\hat{r}_i = \frac{\pi_k}{\pi_i} \cdot \frac{\hat{\eta}_i}{\hat{\eta}_k} \quad \text{and} \quad \hat{\eta}_i = \frac{\pi_i \hat{r}_i}{\sum_{j \in [k]} \pi_j \hat{r}_j}, \; \text{for all } i \in [k], x \in \mathcal{X}. \tag{34}$$

In this case, we choose the following convex function (where we use $\hat{r}_k = 1$):

$$f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \sum_{i=1}^{k} \pi_i \hat{r}_i \log \pi_i \hat{r}_i - \left(\sum_{i=1}^{k} \pi_i \hat{r}_i \right) \log \left(\sum_{i=1}^{k} \pi_i \hat{r}_i \right) \tag{35}$$

$$\partial_i f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \pi_i \log \left(\frac{\pi_i \hat{r}_i}{\sum_{j=1}^{k} \pi_i \hat{r}_j} \right) \quad \text{for } i \in [k-1] \tag{36}$$

Note that when $\boldsymbol{\pi}$ is uniform distribution, Eq. (35) reduces to Eq. (32).

The loss in Eq. (14) reduces to:

$$\mathbb{E}_{p_k(x)} \left[\pi_k \hat{r}_k(x) \log \left(\sum_{i=1}^{k} \pi_i \hat{r}_i(x) \right) - \pi_k \hat{r}_k(x) \log(\pi_k \hat{r}_k(x)) \right] - \sum_{i=1}^{k-1} \mathbb{E}_{p_i(x)} \left[\pi_i \log \left(\frac{\pi_i \hat{r}_i(x)}{\sum_{j=1}^{k} \pi_j \hat{r}_j(x)} \right) \right]$$

$$= - \left(\sum_{i=1}^{k} \pi_i \mathbb{E}_{p_i(x)} [\log(\hat{\eta}_i(x))] \right)$$

which corresponds to the multi-class logistic regression loss for the class probability estimators $\hat{\boldsymbol{\eta}}$.

Remark. Interestingly, we noticed that the multi-distribution f-divergence associated with the convex function in Eq. (32) is the multi-distribution Jensen-Shannon divergence (Garcia-Garcia & Williamson, 2012) (also known as information radius (Sibson, 1969)) up to a constant of $\log k$:

$$\begin{aligned}
\mathbf{D}_f(P_1, \ldots, P_k) &= \mathbb{E}_{p_k(x)} \left[f \left(\frac{p_1(x)}{p_k(x)}, \ldots, \frac{p_{k-1}(x)}{p_k(x)} \right) \right] \\
&= \frac{1}{k} \mathbb{E}_{p_k(x)} \left[\sum_{i=1}^{k} \frac{p_i(x)}{p_k(x)} \log \left(\frac{p_i(x)}{\sum_{j=1}^{k} p_k(x)} \right) \right] \\
&= \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}_{p_i(x)} \left[\log \left(\frac{p_i(x)}{\frac{1}{k} \sum_{j=1}^{k} p_j(x)} \right) \right] - \log k \\
&= \frac{1}{k} \sum_{i=1}^{k} D_{\mathrm{KL}} \left(P_i \| \frac{1}{k} \sum_{j=1}^{k} P_j \right) - \log k
\end{aligned}$$

A.4.2 Least-squares Importance Fitting

When the convex function associated with the Bregman divergence is chosen to be:

$$f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \frac{1}{2} \sum_{i=1}^{k-1} (\hat{r}_i - 1)^2 = \frac{1}{2} \|\hat{\boldsymbol{r}} - \mathbf{1}\|^2 \tag{37}$$

$$\partial_i f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \hat{r}_i - 1 \quad \text{for } i \in [k-1] \tag{38}$$

The loss in Eq. (14) reduces to:

$$\begin{aligned}
&\mathbb{E}_{p_k(x)} \left[\sum_{i=1}^{k-1} (\hat{r}_i^2(x) - \hat{r}_i(x)) - \frac{1}{2} \sum_{i=1}^{k-1} (\hat{r}_i^2(x) - 2\hat{r}_i(x) + 1) \right] - \sum_{i=1}^{k-1} \mathbb{E}_{p_i(x)} \left[\hat{r}_i(x) - 1 \right] \\
=& \frac{1}{2} \mathbb{E}_{p_k(x)} \left[\sum_{i=1}^{k-1} (\hat{r}_i^2(x) - 1) \right] - \sum_{i=1}^{k-1} \mathbb{E}_{p_i(x)} \left[\hat{r}_i(x) - 1 \right] \\
=& \frac{1}{2} \sum_{i=1}^{k-1} \mathbb{E}_{p_k(x)} \left[\hat{r}_i^2(x) - 1 - 2 \frac{p_i(x)}{p_k(x)} (\hat{r}_i(x) - 1) \right] \\
=& \frac{1}{2} \sum_{i=1}^{k-1} \mathbb{E}_{p_k(x)} \left[(\hat{r}_i(x) - r_i(x))^2 \right] - C
\end{aligned}$$

where $C = \mathbb{E}_{p_k(x)} \left[\|\boldsymbol{r}(x) - 1\|^2 \right]$ is a constant w.r.t. $\hat{\boldsymbol{r}}$ and the minimizer to the above loss function matches the true density ratios, which strictly generalizes the Least-Squares Importance Fitting (LSIF) (Kanamori et al., 2009) method to the multi-distribution case.

A.4.3 KL IMPORTANCE ESTIMATION PROCEDURE

When the convex function associated with the Bregman divergence is chosen to be:

$$f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \sum_{i=1}^{k-1} (\hat{r}_i \log \hat{r}_i - \hat{r}_i) = \langle \hat{\boldsymbol{r}}, \log(\hat{\boldsymbol{r}}) \rangle - \|\hat{\boldsymbol{r}}\|_1 \tag{39}$$

$$\partial_i f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \log \hat{r}_i \quad \text{for } i \in [k-1] \tag{40}$$

The loss in Eq. (14) reduces to:

$$\begin{aligned} &\mathbb{E}_{p_k(x)} \left[\sum_{i=1}^{k-1} \hat{r}_i(x) \log \hat{r}_i(x) - \sum_{i=1}^{k-1} (\hat{r}_i(x) \log \hat{r}_i(x) - \hat{r}_i(x)) \right] - \sum_{i=1}^{k-1} \mathbb{E}_{p_i(x)}[\log \hat{r}_i(x)] \\ =&\mathbb{E}_{p_k(x)} \left[\sum_{i=1}^{k-1} \hat{r}_i(x) \right] - \sum_{i=1}^{k-1} \mathbb{E}_{p_i(x)}[\log \hat{r}_i(x)] \end{aligned} \tag{41}$$

This is equivalent to the following constrained optimization problem:

$$\begin{aligned} &\arg\min_{\hat{\boldsymbol{r}}: \mathcal{X} \to \mathbb{R}^{k-1}} \sum_{i=1}^{k-1} D_{\mathrm{KL}}(p_i(x) \| \hat{r}_i(x) \cdot p_k(x)) = \sum_{i=1}^{k-1} \mathbb{E}_{p_i(x)} \left[\log \left(\frac{p_i(x)}{\hat{r}_i(x) \cdot p_k(x)} \right) \right] \\ =& \arg\min_{\hat{\boldsymbol{r}}: \mathcal{X} \to \mathbb{R}^{k-1}} - \sum_{i=1}^{k-1} \mathbb{E}_{p_i(x)}[\log \hat{r}_i(x)] \\ &\text{s.t.} \quad \mathbb{E}_{p_k(x)}[\hat{r}_i(x)] = 1 \text{ and } \hat{r}_i(x) \geq 0, \text{ for all } i \in [k-1]. \end{aligned}$$

which strictly generalizes the Kullback–Leibler Importance Estimation Procedure (KLIEP) (Sugiyama et al., 2008) to the multi-distribution case.

A.4.4 BASU'S POWER DIVERGENCE FOR ROBUST DRE

For some $\alpha > 1$, we choose the following convex function (the α-norm of a vector):

$$f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \sum_{i=1}^{k-1} \hat{r}_i^{\alpha} = \|\hat{\boldsymbol{r}}\|_{\alpha}^{\alpha} \tag{42}$$

$$\partial_i f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \alpha \hat{r}_i^{\alpha - 1} \tag{43}$$

The loss in Eq. (14) reduces to:

$$\begin{aligned} &\mathbb{E}_{p_k(x)} \left[\sum_{i=1}^{k-1} \alpha \hat{r}_i^{\alpha}(x) - \sum_{i=1}^{k-1} \hat{r}_i^{\alpha}(x) \right] - \sum_{i=1}^{k-1} \mathbb{E}_{p_i(x)} \left[\alpha \hat{r}_i^{\alpha-1}(x) \right] \\ =&\sum_{i=1}^{k-1} \mathbb{E}_{p_k(x)} \left[(\alpha - 1) \hat{r}_i^{\alpha}(x) \right] - \sum_{i=1}^{k-1} \mathbb{E}_{p_i(x)} \left[\alpha \hat{r}_i^{\alpha-1}(x) \right] \end{aligned} \tag{44}$$

To understand the robustness of this formulation, for each $i \in [k-1]$, we take the derivative of Eq. (44) w.r.t. the parameters in the density ratio model $\hat{r}_i$ and equate it to zero, and we get the following parameter estimation equation:

$$\mathbb{E}_{p_k(x)}[\hat{r}_i^{\alpha-1}(x) \nabla \hat{r}_i(x)] - \mathbb{E}_{p_i(x)}[\hat{r}_i^{\alpha-2}(x) \nabla \hat{r}_i(x)] = \mathbf{0} \tag{45}$$

Now we apply the same analysis to the multi-distribution KLIEP method in Eq. (41) and we get the following equation (for each $i \in [k-1]$):

$$\mathbb{E}_{p_k(x)}[\nabla \hat{r}_i(x)] - \mathbb{E}_{p_i(x)}[\hat{r}_i^{-1}(x) \nabla \hat{r}_i(x)] = \mathbf{0} \tag{46}$$

Comparing Eq. (45) with Eq. (46), we can see that the power divergence DRE method in Eq. (44) is a weighted version of the multi-distribution KLIEP method, where the weight $\hat{r}_i^{\alpha-1}(x)$ controls

the importance of the samples. In some scenario where the outlier samples tend to have density ratios less than one, they will have less influence on the parameter estimation, which generalizes the binary Basu's power divergence robust DRE method (Sugiyama et al., 2012) to the multi-distribution case. Another interesting observation is that when $\alpha \to 1$, Eq. (45) recovers the KLIEP gradient in Eq. (46); when $\alpha = 2$, the power divergence DRE in Eq. (44) recovers the multi-distribution LSIF method in Section A.4.2.

A.4.5 More Examples

When the convex function is chosen to be the Log-Sum-Exp type function (for $\alpha > 0$):

$$f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \alpha \log \left(\sum_{i=1}^{k-1} \exp(\hat{r}_i/\alpha) \right) \tag{47}$$

$$\partial_i f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \frac{\exp(\hat{r}_i/\alpha)}{\sum_{i=1}^{k-1} \exp(\hat{r}_i/\alpha)} \tag{48}$$

The loss in Eq. (14) can be written as:

$$\mathbb{E}_{p_k(x)} \left[\sum_{i=1}^{k-1} \frac{\hat{r}_i(x) \exp(\hat{r}_i(x)/\alpha)}{\sum_{j=1}^{k-1} \exp(\hat{r}_j(x)/\alpha)} - \alpha \log \left(\sum_{i=1}^{k-1} \exp(\hat{r}_i(x)/\alpha) \right) \right] - \sum_{i=1}^{k-1} \mathbb{E}_{p_i(x)} \left[\frac{\exp(\hat{r}_i(x)/\alpha)}{\sum_{j=1}^{k-1} \exp(\hat{r}_j(x)/\alpha)} \right]$$

We can similarly derive loss functions induced by other convex functions such as the quadratic function $f(\hat{r}_1, \ldots, \hat{r}_{k-1}) = \hat{\boldsymbol{r}}^\top \boldsymbol{H} \hat{\boldsymbol{r}} + \boldsymbol{q}^\top \hat{\boldsymbol{r}}$, for some positive definite matrix $\boldsymbol{H} \succ 0$.

A.5 More Experimental Details

We provide more details about the problem setup of each task used in our empirical study.

For the synthetic data experiments, we use $k = 5$ multivariate Gaussian distributions with identity covariance matrix and different mean vectors:

$$\begin{aligned}
\boldsymbol{\mu}_1 &= (1, 0, 0, \ldots)^d \\
\boldsymbol{\mu}_2 &= (-1, 0, 0, \ldots)^d \\
\boldsymbol{\mu}_3 &= (0, 1, 0, \ldots)^d \\
\boldsymbol{\mu}_4 &= (0, -1, 0, \ldots)^d \\
\boldsymbol{\mu}_5 &= (1, 0, 0, \ldots)^d
\end{aligned}$$

We use such design so that the density ratios are almost surely well-defined and the numerical optimization with respect to the canonical density ratio vector $\hat{\boldsymbol{r}} = (\hat{r}_1, \ldots, \hat{r}_{k-1})$ is more stable. We use a two-layer Multi-Layer Perceptron (MLP) (Linear$(d, 32) \to$ Linear$(32, 32) \to$ Linear$(32, k - 1)$) with ReLU activation function to realize the density ratio model.

For CIFAR-10 OOD detection experiments, we set $k = 4$ and we construct each distribution as: p_1 - samples labeled {airplane, automobile, bird}; p_2 - samples labeled {cat, deer, dog, frog}; p_3 - samples labeled {horse, ship, truck} and p_4 - a uniform mixture of these distributions. We use a standard convolution neural network in the PyTorch tutorial[1] with $k-1$ outputs to realize the density ratio model.

For MNIST multi-target generation experiments, we use $k = 6$ and we construct each distribution as: p_1 - samples labeled {0,1}; p_2 - samples labeled {2,3}; p_3 - samples labeled {4,5}; p_4 - samples labeled {6,7}; p_5 - samples labeled {8,9}; p_6 - a mixture of these distributions. We use a two-layer convolutional neural network (Conv(1,32, 3, 1) $\to$ Conv(32, 64, 3, 1) $\to$ Linear(9216, 128) $\to$ Linear(128, $k - 1$)) with ReLU activation function to realize the density ratio model.

For multi-distribution off-policy policy evaluation experiments, we conducted experiments on the Half-Cheetah environment in OpenAI Gym (Brockman et al., 2016). We use soft actor-critic algorithm (Haarnoja et al., 2018) to obtain five different policies with average expected return of

[1] https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html

$\{3811, 5277, 6444, 7397, 5728\}$ respectively and we learn density ratios between their induced occupancy measures (state-action distributions). We use a three-layer MLP (Linear$(17, 256) \rightarrow$ Linear$(256, 256) \rightarrow$ Linear$(256, 256) \rightarrow$ Linear$(256, k-1)$) with ReLU activation function to realize the density ratio model.