Non-Negative Bregman Divergence Minimization for Deep Direct Density Ratio Estimation

Masahiro Kato [1] **Takeshi Teshima** [2]

Abstract

Density ratio estimation (DRE) is at the core of various machine learning tasks such as anomaly detection and domain adaptation. In existing studies on DRE, methods based on *Bregman divergence* (BD) minimization have been extensively studied. However, BD minimization when applied with highly flexible models, such as deep neural networks, tends to suffer from what we call *train-loss hacking*, which is a source of *overfitting* caused by a typical characteristic of empirical BD estimators. In this paper, to mitigate train-loss hacking, we propose a *non-negative correction* for empirical BD estimators. Theoretically, we confirm the soundness of the proposed method through a generalization error bound. Through our experiments, the proposed methods show a favorable performance in inlier-based outlier detection.

1. Introduction

Density ratio estimation (DRE) has attracted a great deal of attention as an essential task in various machine learning problems, such as regression under a covariate shift (Shimodaira, 2000; Reddi et al., 2015), learning with noisy labels (Liu & Tao, 2014; Fang et al., 2020), anomaly detection (Smola et al., 2009; Hido et al., 2011; Abe & Sugiyama, 2019), two-sample testing (Keziou & Leoni-Aubin, 2005; Kanamori et al., 2010; Sugiyama et al., 2011a), causal inference (Uehara et al., 2020), change point detection (Kawahara & Sugiyama, 2009), and binary classification only from positive and unlabeled data (PU learning; Kato et al., 2019). For instance, anomaly detection is not easy to apply based on standard machine learning methods since anomalous data are often scarce; however, we can solve this by estimating the density ratio when anomaly-free unlabeled test data are available (Hido et al., 2008).

Among the various approaches to DRE, we focus on the Bregman divergence (BD) minimization framework (Bregman, 1967; Sugiyama et al., 2011b), which is a general framework that unifies various DRE methods, such as moment matching (Huang et al., 2007; Gretton et al., 2009), probabilistic classification (Qin, 1998; Cheng & Chu, 2004), density matching (Nguyen et al., 2010), and density-ratio fitting (Kanamori et al., 2009). Kato et al. (2019) proposed using the risk of PU learning for DRE, which also falls within this framework (Appendix A).

Existing methods have mainly adopted linear-in-parameter models for DRE (Kanamori et al., 2012). On the other hand, recent studies in machine learning have suggested that deep neural networks achieve significantly high performances for various tasks, such as computer vision (CV) (Krizhevsky et al., 2012) and natural language processing (NLP) (Bengio et al., 2001). These findings motivate us to use deep neural networks for DRE.

However, when using deep neural networks in combination with empirical BD minimization, we often observe a serious *overfitting* problem as experimentally demonstrated in Figure 2 of Section 5.1 and Figure 4 of Appendix F.1.1. We observe that this is mainly because a model of the density ratio $r(X) = \frac{q(X)}{p(X)}$ between two probability densities p and q becomes large for high-dimensional data when using a flexible model. As the intuition behind this phenomenon, a flexible model can overfit the samples generated from $p(X)$ as if there were no common support between $p(X)$ and $q(X)$ (Figure 1). This hypothesis is inspired by Kiryo et al. (2017), which reports a similar problem in PU learning. In the case of DRE through BD minimization, we conjecture that this phenomenon is caused by the objective function that monotonically decreases with respect to $r(X)$, misleading the model $r(X)$ to take on as large a value as possible on the specific data points X. Note that even if $r(X)$ is bounded, this phenomenon still manifests as the model takes on the largest possible value within its output range. Rhodes et al. (2020) and Ansari et al. (2020) independently found related problems in DRE. Whereas Kiryo et al. (2017) and Rhodes et al. (2020) call their phenomena *over-*

[1]CyberAgent Inc., Tokyo, Japan [2]The University of Tokyo, Tokyo, Japan. Correspondence to: Masahiro Kato <masahiro_kato@cyberagent.co.jp>, Takeshi Teshima <teshima@ms.k.u-tokyo.ac.jp>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

fitting and *density-chasm problem*, respectively, we refer to our problem as *train-loss hacking* because this problem is specific to methods based on BD minimization, and we can still observe this issue even when the true $r(x) = \frac{p(x)}{q(x)}$ is not significantly large. This problem is discussed in more detail in Section 2.3 (Figure 1).

Owing to this property, training a density ratio model with a flexible model tends to result in either a diverging empirical BD estimator or a model sticking to the upper bound of its output range. For instance, when the empirical BD divergence is not lower bounded, it often numerically diverges to negative infinity. Even when the loss function has a lower bound (see *BKL* and *Bounded uLSIF* introduced in Sections 2.2 and 5.1), the trained models tend to stick to the largest possible value of their output ranges (Bounded uLSIF in Figure 2 and BKL-NN in Figure 4). Although train-loss hacking has rarely been discussed in existing studies on DRE, this problem is often encountered while using deep neural networks, as experimentally shown in Section 5.1. One reason for this is that the existing studies use linear-in-parameter models (Kanamori et al., 2012) or simple shallow neural networks (Nam & Sugiyama, 2015; Abe & Sugiyama, 2019) for the density ratio models, which tend to be inflexible in that they do not cause such a phenomenon.

To mitigate the train-loss hacking, we propose a general procedure to modify the empirical BD estimator. First, from the empirical BD divergence, we separate the term causing the train-loss hacking. We then apply a *non-negative correction* to the term to make the model consistent with a constraint that should be satisfied within the population. Our idea of this correction is inspired by Kiryo et al. (2017). However, their idea of a non-negative correction is only applicable to the binary classification setting; thus, we require a non-trivial rewriting of the BD to generalize the approach to our problem. We call our proposed objective function the *non-negative BD* (nnBD). The proposed method can be regarded as a generalization of the method proposed by Kiryo et al. (2017).

Our main contributions are (1) proposal of a general procedure to modify an empirical BD estimator to enable DRE with flexible models, (2) theoretical justification of the proposed estimator, and (3) experimental validation of the proposed method using benchmark data.

2. Problem setting

Let $\mathcal{X}^{\mathrm{nu}} \subseteq \mathbb{R}^d$ and $\mathcal{X}^{\mathrm{de}} \subseteq \mathbb{R}^d$ be the spaces of the d-dimensional *covariates*. Here, "nu" and "de" indicate the numerator and denominator. Let p_{nu} and p_{de} be the probability densities over $\mathcal{X}^{\mathrm{nu}}$ and $\mathcal{X}^{\mathrm{de}}$, respectively. We have independent and identically distributed (i.i.d.) samples from these distributions: $\mathbf{X}^{\mathrm{nu}} = \{X_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{nu}}$ and $\mathbf{X}^{\mathrm{de}} = \{X_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{de}}$.

Basic assumption and goal. Throughout this paper, we assume that $p_{\mathrm{nu}}(X)$ and $p_{\mathrm{de}}(X)$ are strictly positive over $\mathcal{X}^{\mathrm{nu}}$ and $\mathcal{X}^{\mathrm{de}}$, respectively. We also assume $\mathcal{X}^{\mathrm{nu}} \subseteq \mathcal{X}^{\mathrm{de}}$, which is a typical assumption in the literature on DRE, e.g., Section 2.1 of Kanamori et al. (2009). The goal of DRE is to estimate $r^*(x) = \frac{p_{\mathrm{nu}}(x)}{p_{\mathrm{de}}(x)}$ from the samples $\mathbf{X}^{\mathrm{nu}}$ and $\mathbf{X}^{\mathrm{de}}$.

Additional notation. Let $\mathbb{E}_{\mathrm{nu}}$ and $\mathbb{E}_{\mathrm{de}}$ denote the expectations with respect to $p_{\mathrm{nu}}(X)$ and $p_{\mathrm{de}}(X)$, respectively; in addition, $\hat{\mathbb{E}}_{\mathrm{nu}}$ and $\hat{\mathbb{E}}_{\mathrm{de}}$ denote the sample averages over $\{X_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ and $\{X_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$, respectively.

2.1. Density ratio matching by BD minimization

Among existing DRE methods, we focus on *density ratio matching through BD minimization* (DRM-BD; Sugiyama et al., 2011b), which is a framework that unifies various DRE methods (Gretton et al., 2009; Sugiyama et al., 2008; Kanamori et al., 2009; Nguyen et al., 2010).

DRM-BD estimates the density ratio by minimizing the objective function derived as follows: Let $(b_r, B_r) \subset [0, \infty)$, and let $f : (b_r, B_r) \to \mathbb{R}$ be a twice continuously differentiable convex function with a bounded derivative ∂f (Table 1). We quantify the discrepancy from the true density ratio function r^* to a density ratio model r by

$$\mathrm{BD}_f(r^* \| r) := \mathbb{E}_{\mathrm{de}}\left[\partial f(r(X))r(X) - f(r(X))\right] - \mathbb{E}_{\mathrm{nu}}\left[\partial f(r(X))\right] \quad (1)$$

which is equal to the BD (Bregman, 1967) defined as $\mathbb{E}_{\mathrm{de}}[\ddot{\mathrm{BD}}_f(r^*(X)\|r(X))]$, where

$$\ddot{\mathrm{BD}}_f(t^* \| t) := f(t^*) - f(t) - \partial f(t)(t^* - t),$$

ignoring the constant $\overline{\mathrm{BD}} = \mathbb{E}_{\mathrm{de}}\big[f(r^*(X))\big]$. Then, given a hypothesis class $\mathcal{H}$, DRM-BD estimates r^* using a minimizer of the sample analog of (1):

$$\widehat{\mathrm{BD}}_f(r) := \hat{\mathbb{E}}_{\mathrm{de}}\Big[\partial f\big(r(X_i)\big)r(X_i) - f\big(r(X_i)\big)\Big] - \hat{\mathbb{E}}_{\mathrm{nu}}\Big[\partial f\big(r(X_j)\big)\Big]. \quad (2)$$

2.2. Examples of DRE

Sugiyama et al. (2011b) showed that BD minimization can unify various DRE methods. Furthermore, Menon & Ong (2016) showed an equivalence between conditional probability estimation and DRE from the BD minimization perspective. In addition, by generalizing the results of du Plessis et al. (2015) and Kato et al. (2019), we derive a novel method for DRE from PU learning in Appendix B. We summarize

Table 1. Summary of DRE methods (Sugiyama et al., 2011b). For PULogLoss, we use $C < \frac{1}{\overline{R}}$.

Method	$f(t)$	Lower bound of $\widehat{\mathrm{BD}}_f$	Reference
LSIF	$(t-1)^2/2$	Not bounded	Kanamori et al. (2009)
Kernel Mean Matching	$(t-1)^2/2$	Not bounded	Gretton et al. (2009)
UKL	$t\log(t) - t$	Not bounded	Nguyen et al. (2010)
KLIEP	$t\log(t) - t$	Not bounded	Sugiyama et al. (2008)
BKL (LR)	$t\log(t) - (1+t)\log(1+t)$	Bounded	Hastie et al. (2001)
PULogLoss	$C\log(1-t) + Ct(\log(t) - \log(1-t))$ for $0 < t < 1$	Not bounded	Kato et al. (2019)

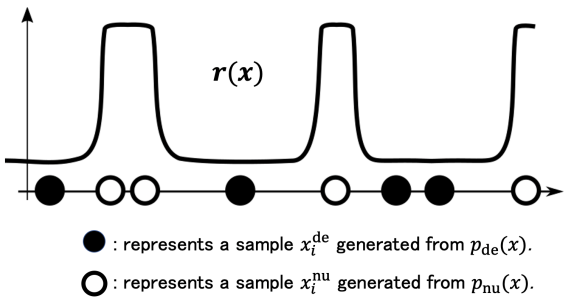

Figure 1. Illustration of the train-loss hacking phenomenon. Given finite data points, a sufficiently flexible model r can easily make the second term of the objective function $\widehat{\mathrm{BD}}_f$ largely negative, resulting in an unreasonable minimizer.

the DRE methods in Table 1. Here, the empirical risks of *least-square importance fitting* (LSIF), *unnormalized Kullback–Leibler* (UKL) divergence, *binary Kullback–Leibler* (BKL) divergence, and *PU learning with log Loss* (PULogLoss) are given as

$$\begin{aligned}
\widehat{\mathrm{BD}}_{\mathrm{LSIF}}(r) &:= \frac{1}{2}\hat{\mathbb{E}}_{\mathrm{de}}[r^2(X_i)] - \hat{\mathbb{E}}_{\mathrm{nu}}[r(X_j)], \\
\widehat{\mathrm{BD}}_{\mathrm{UKL}}(r) &:= \hat{\mathbb{E}}_{\mathrm{de}}\left[r(X_i)\right] - \hat{\mathbb{E}}_{\mathrm{nu}}\left[\log\left(r(X_j)\right)\right], \\
\widehat{\mathrm{BD}}_{\mathrm{BKL}}(r) &:= -\hat{\mathbb{E}}_{\mathrm{de}}\left[\mathrm{BKL1}(X_i)\right] - \hat{\mathbb{E}}_{\mathrm{nu}}\left[\mathrm{BKL2}(X_j)\right], \\
\widehat{\mathrm{BD}}_{\mathrm{PU}}(r) &:= -\hat{\mathbb{E}}_{\mathrm{de}}\left[\log\left(1 - r(X_i)\right)\right] \\
&\quad + C\hat{\mathbb{E}}_{\mathrm{nu}}\left[-\log\left(r(X_j)\right) + \log\left(1 - r(X_j)\right)\right],
\end{aligned}$$

where $0 < C < \frac{1}{\overline{R}}$, $\overline{R}$ is an upper bound on r^*, $\mathrm{BKL1}(x) = \log\left(\frac{1}{1+r(x)}\right)$, and $\mathrm{BKL2}(x) = \log\left(\frac{r(x)}{1+r(x)}\right)$. Here, $\widehat{\mathrm{BD}}_{\mathrm{LSIF}}(r)$ and $\widehat{\mathrm{BD}}_{\mathrm{PU}}(r)$ correspond to LSIF and PULogLoss, respectively. We can derive the *Kullback–Leibler importance estimation procedure* (KLIEP) and logistic regression-based DRE (LR) from $\widehat{\mathrm{BD}}_{\mathrm{UKL}}(r)$ and $\widehat{\mathrm{BD}}_{\mathrm{BKL}}(r)$. In $\widehat{\mathrm{BD}}_{\mathrm{PU}}(r)$, we restrict the model's output to be within $(0, 1)$ and the estimated model r becomes an estimator of Cr^*. Details of these methods are provided in Appendix A.

2.3. Train-loss hacking problem

DRM-BD with neural networks often suffers from an overfitting. For instance, in Section 5.1, we show that the LSIF with neural networks suffers from a serious overfitting issue. We conjecture that a conceivable cause of the overfitting in DRE-BD is the *train-loss hacking*. This hypothesis is inspired by Kiryo et al. (2017), which tackled a similar problem in PU learning; see Appendix C for a brief review.

Train-loss hacking is a phenomenon in which $r(X_j)$ increases to a large value for $\{X_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{nu}}(x)$ when a model is trained to minimize an objective function consisting of multiple separate samples. Recall that, in the case of DRE-BD, the objective function (2) consists of two empirical averages. Among these two, we can make the second term $-\hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right]$ small by making the model to take as large values as possible at the specific input points $\mathbf{X}^{\mathrm{nu}}$ (Figure 1). In fact, since f is a convex function, ∂f is an increasing function; hence, this term monotonically decreases as r increases in $\mathbf{X}^{\mathrm{nu}}$. As a result, when there is no lower bound on $-\hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right]$, it often numerically diverges to negative infinity[1]. Even when there is a lower bound on $-\hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right]$, $r(X_j)$ tends to take on the largest possible value of its output range at the points $\mathbf{X}^{\mathrm{nu}}$. Naively "capping" the model's output, e.g., composing $r(X)$ with $\max\{\cdot, B\}$ where $B > 0$ is a constant, fails to remedy the issue as the model still tends to stick to its largest possible value. We experimentally demonstrate this by implementing the Bounded uLSIF (Figure 2).

This is a critical issue, since merely making the output large on $\mathbf{X}^{\mathrm{nu}}$ is unlikely to be a reasonable training criterion for DRE, and it may only lead to an unreasonable density ratio estimator (Figure 1). The issue becomes salient when the model has a high flexibility. If the hypothesis class has an extremely limited flexibility, this may not be an issue since the remaining term $\hat{\mathbb{E}}_{\mathrm{de}}\left[\partial f\left(r(X_i)\right)r(X_i) - f\left(r(X_i)\right)\right]$ is likely to introduce a trade-off. However, when highly flexible models such as deep neural networks are employed, the model can easily fit to $\mathbf{X}^{\mathrm{nu}}$ and $\mathbf{X}^{\mathrm{de}}$ separately (Figure 1).

[1] Even if we use a bounded model for r, it still tends to diverge during the numerical computation.

3. Deep direct DRE based on non-negative risk estimator

Although DRE with flexible models suffers from serious train-loss hacking, we still have a strong motivation to use them for applications, such as CV and NLP. In this section, we describe our approach to modify the DRM-BD objective function to mitigate the train-loss hacking problem.

3.1. Non-negative BD

To alleviate the train-loss hacking problem, we propose a *non-negative BD estimator* that modifies an empirical BD estimator (2) to be robust against the problem. The proposed method is inspired by Kiryo et al. (2017), which suggested a non-negative correction to the empirical risk of PU learning based on the knowledge that a part of the population risk is non-negative. However, in DRE, it is not straightforward to employ this approach because we do not know which part of the population risk (1) is non-negative. In this paper, by assuming an upper bound $\overline{R}$ on the density ratio r^*, we detect which part of the risk of DRE (1) is non-negative in the population. Then, we apply a non-negative correction to an empirical BD estimator (2) based on the non-negativity of the corresponding part of the population risk. This non-negative correction also corresponds to a generalization of non-negative PU learning (Kiryo et al., 2017).

To enable our approach to mitigate the train-loss hacking, we apply the following assumption:

Assumption 1. The density ratio r^* is bounded from above, i.e., $\overline{R} = \sup_{X \in \mathcal{X}^{\mathrm{de}}} r^*(X) < \infty$.

Then, we arbitrarily specify a constant C such that $0 < C < \frac{1}{\overline{R}}$. Using C, we make the following assumption.

Assumption 2. A function $\tilde{f}$ defined by

$$\partial f(t) = C\big(\partial f(t)t - f(t)\big) + \tilde{f}(t) \tag{3}$$

is bounded from above.

Then, we rewrite the DRM-BD objective (1) as

$$\mathrm{BD}_f(r^* \| r) = \underbrace{\mathbb{E}_{\mathrm{de}}[\ell_1(r(X))] - C\mathbb{E}_{\mathrm{nu}}[\ell_1(r(X))]}_{(*)} + \mathbb{E}_{\mathrm{nu}}\ell_2(r(X)) - (1-C)A, \tag{4}$$

where ℓ_1 and ℓ_2 are

$$\ell_1(t) := \partial f(t)t - f(t) + A, \quad \ell_2(t) := -\tilde{f}(t),$$

and A is a constant such that $\ell_1(t) \geq 0$ for all $t \in (b_r, B_r)$. Now, we make the following observation.

Observation. The $(*)$ part in (4) is non-negative since both ℓ_1 and $p_{\mathrm{de}} - Cp_{\mathrm{nu}}$ are non-negative under Assumption 1 and $0 < C < \frac{1}{\overline{R}}$.

Based on this observation, we propose using the following modified empirical risk:

$$\widehat{\mathrm{nnBD}}_f(r) := \Big(\hat{\mathbb{E}}_{\mathrm{de}}\big[\ell_1(r(X_i))\big] - C\hat{\mathbb{E}}_{\mathrm{nu}}\big[\ell_1(r(X_j))\big]\Big)_+ + \hat{\mathbb{E}}_{\mathrm{nu}}\big[\ell_2(r(X_j))\big] \tag{5}$$

where $(\cdot)_+ := \max\{0, \cdot\}$. Note that the nonnegativity of $(*)$ is always satisfied in the population quantity; however, it can be violated in finite samples, allowing for train-loss hacking. Our *deep direct DRE* (D3RE) is based on minimizing $\widehat{\mathrm{nnBD}}_f(r)$ over a hypothesis class of the density ratio $\mathcal{H} \subset \{r : \mathbb{R}^d \to (b_r, B_r)\}$, where $0 \leq b_r < \overline{R} < B_r$.

Instantiations of the D3RE objective functions. The above strategy can be instantiated with various functions f proposed for DRM-BD. Here, we introduce nnBD corresponding to LSIF, UKL, BKL, and PULogLoss as follows:

$$\widehat{\mathrm{nnBD}}_{\mathrm{LSIF}}(r) := -\hat{\mathbb{E}}_{\mathrm{nu}}\left[r(X_j) - \frac{C}{2}r^2(X_j)\right] + \left(\frac{1}{2}\hat{\mathbb{E}}_{\mathrm{de}}\left[r^2(X_i)\right] - \frac{C}{2}\hat{\mathbb{E}}_{\mathrm{nu}}\left[r^2(X_j)\right]\right)_+,$$

$$\widehat{\mathrm{nnBD}}_{\mathrm{UKL}}(r) := -\hat{\mathbb{E}}_{\mathrm{nu}}\left[\log\big(r(X_j)\big) - Cr(X_j)\right] + \left(\hat{\mathbb{E}}_{\mathrm{de}}\left[r(X_i)\right] - C\hat{\mathbb{E}}_{\mathrm{nu}}\left[r(X_j)\right]\right)_+,$$

$$\widehat{\mathrm{nnBD}}_{\mathrm{BKL}}(r) := -\hat{\mathbb{E}}_{\mathrm{nu}}\left[\mathrm{BKL2}(X_j) + C\mathrm{BKL1}(X_j)\right] + \left(-\hat{\mathbb{E}}_{\mathrm{de}}\left[\mathrm{BKL1}(X_i)\right] + C\hat{\mathbb{E}}_{\mathrm{nu}}\left[\mathrm{BKL1}(X_j)\right]\right)_+,$$

$$\widehat{\mathrm{nnBD}}_{\mathrm{PU}}(r) := -C\hat{\mathbb{E}}_{\mathrm{nu}}\left[\log\big(r(X_j)\big)\right] + \left(C\hat{\mathbb{E}}_{\mathrm{nu}}\left[\log\big(1 - r(X_j)\big)\right] - \hat{\mathbb{E}}_{\mathrm{de}}\left[\log\big(1 - r(X_i)\big)\right]\right)_+.$$

More detailed derivation of $\tilde{f}$ is in Appendix B.

The algorithm for D3RE is described in Algorithm 1. For training with a large amount of data, we adopt a stochastic optimization by splitting the dataset into mini-batches. In stochastic optimization, we separate the samples into N mini-batches as $(\{X_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu},j}}, \{X_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de},j}})$ $(j = 1, \ldots, N$, where $n_{\mathrm{nu},j}$ and $n_{\mathrm{de},j}$ are the sample sizes for each mini-batch. Then, we consider the sample average in each mini-batch. Let $\hat{\mathbb{E}}_{\mathrm{nu}}^j$ and $\hat{\mathbb{E}}_{\mathrm{de}}^j$ be sample averages over $\{X_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu},j}}$ and $\{X_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de},j}}$. In addition, we use regularization, such as L1 and L2 penalties, as denoted by $\mathcal{R}(r)$.

To improve the performance, we can heuristically employ *gradient ascent* from Kiryo et al. (2017) when $\hat{\mathbb{E}}_{\mathrm{de}}\big[\ell_1(r(X))\big] - C\hat{\mathbb{E}}_{\mathrm{nu}}\big[\ell_1(r(X))\big]$ becomes less than 0, i.e., the model is updated in the direction that increases the term. Note that gradient ascent is not essential in D3RE, and we can obtain similar results even without it (see experiments in Appendix F.1.2). We recommend practitioners to

use a gradient ascent and those concerned with a theoretical guarantee to use a plain gradient descent.

Choice of C. Although we use an upper bound of the density ratio in the formulation, we do not require a tight one. The main role of the upper bound is to prevent the density ratio model from diverging, and as long as we successfully prevent divergence, the proposed algorithms work well. We find that D3RE is robust against a loose specification of the upper bound to a certain extent in our experiment (the right graph in Figure 2 of Section 5.1). Thus, in practice, selecting C does not require accurate knowledge of $\overline{R}$. In fact, in inlier-based outlier detection experiments, the proposed methods under a loose specification of the upper bound achieve a preferable performance. However, selecting a hyper-parameter C that is much smaller than $1/\overline{R}$ may damage the empirical performance as shown in Section 5.1. This, of course, does not mean that $1/\overline{R}$ should not be small; if $1/\overline{R}$ is small, C can also be small.

Non-negative PU learning. Kiryo et al. (2017) proposed a non-negative correction for PU learning (nnPU). In this paper, we propose a non-negative correction for DRE, inspired by Kiryo et al. (2017); however, our extension is nontrivial because the relationship between DRE and PU learning has not been well understood. Another contribution of this paper is that it clarifies the relationship between DRE and PU learning, as described in Appendix A and Section 5. We find that the class-prior in PU learning corresponds to the upper bound of r^* in DRE, and that as Sugiyama et al. (2012) generalized DRE in terms of BD divergence minimization, the risk of PU learning can also be generalized through BD divergence minimization. This finding had been implied by Kato et al. (2019), although it had not been formally shown. This finding clarifies the relationship between DRE and PU learning, and thus makes it possible to apply the non-negative correction to DRE, such as nnPU.

3.2. Motivation and intuitive justification of D3RE

Here, we describe how the above non-negative risk correction alleviates the train-loss hacking problem.

D3RE and an unbounded empirical risk. First, we consider the case where the empirical BD is unbounded. First, we assume the following on $\tilde{f}(t)$ of (3). Assumption 2 is satisfied by most of the loss functions which appear in the previously proposed DRE methods (see Appendix B for examples). Under Assumption 2, because $\tilde{f}(t)$ is bounded above, the train-loss hacking $-\hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right] \to -\infty$ to minimize the empirical risk (2) is caused by

$$-\hat{\mathbb{E}}_{\mathrm{nu}}\left[C\left\{\partial f(r(X_j))r(X_j) - f(r(X_j))\right\}\right] \to -\infty$$

Algorithm 1 D3RE

Input: Training data $\left\{X_j^{\mathrm{nu}}\right\}_{j=1}^{n_{\mathrm{nu}}}$ and $\left\{X_i^{\mathrm{de}}\right\}_{i=1}^{n_{\mathrm{de}}}$, the algorithm for stochastic optimization such as Adam (Kingma & Ba, 2015), the learning rate γ, the regularization coefficient and function λ and $\mathcal{R}(r)$, and a constant $C > 0$.
Output: A density ratio estimator $\hat{r}$.
while No stopping criterion has been met: **do**
  Create N mini-batches

$$\left\{\left(\left\{X_j^{\mathrm{nu}}\right\}_{j=1}^{n_{\mathrm{nu},k}}, \left\{X_i^{\mathrm{de}}\right\}_{i=1}^{n_{\mathrm{de},k}}\right)\right\}_{k=1}^{N}.$$

  for $k = 1$ to N **do**
    if $\hat{\mathbb{E}}_{\mathrm{de}}^k\left[\ell_1(r(X))\right] - C\hat{\mathbb{E}}_{\mathrm{nu}}^k\left[\ell_1(r(X))\right] \geq 0$: **then**
      Gradient decent: set gradient

$$\nabla_r\left\{\hat{\mathbb{E}}_{\mathrm{nu}}^k\left[\ell_2(r(X))\right] + \hat{\mathbb{E}}_{\mathrm{de}}^k\left[\ell_1(r(X))\right] - C\hat{\mathbb{E}}_{\mathrm{nu}}^k\left[\ell_1(r(X))\right] + \lambda\mathcal{R}(r)\right\}.$$

    else
      Gradient ascent: set gradient

$$\nabla_r\left\{-\hat{\mathbb{E}}_{\mathrm{de}}^k\left[\ell_1(r(X))\right] + C\hat{\mathbb{E}}_{\mathrm{nu}}^k\left[\ell_1(r(X))\right] + \lambda\mathcal{R}(r)\right\}.$$

    end if
    Update r with the gradient and the learning rate γ.
  end for
end while

because

$$\underbrace{-\hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right]}_{\to -\infty} = \underbrace{-\hat{\mathbb{E}}_{\mathrm{nu}}\left[\tilde{f}(r(X_j))\right]}_{\text{Bounded}} \underbrace{-\hat{\mathbb{E}}_{\mathrm{nu}}\left[C\left\{\partial f(r(X_j))r(X_j) - f(r(X_j))\right\}\right]}_{\to -\infty}.$$

This observation implies that our non-negative correction (5) prevents train-loss hacking by effectively introducing the correction to the problematic term ($*$ in (4)).

D3RE and a bounded empirical risk. Next, we consider the case where $\partial f\left(r(X_j)\right.$ or model $r(x)$ is bounded. Even in these cases, train-loss hacking can occur. For instance, if $\partial f(t) = \log(t) - \log(1+t)$ (BKL), $\partial f(t)$ is upper-bounded by 0, and $-\hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right]$ does not diverge to $-\infty$. However, we can infinitely decrease $-\hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right]$ to 0 by making $r(X_j) \to \infty$, which causes train-loss hacking. On the other hand, when $r(x)$ is upper-bounded, we can minimize $-\hat{\mathbb{E}}_{\mathrm{nu}}\left[\partial f\left(r(X_j)\right)\right]$ by training $r(X_j)$ to stick to the upper bound at $\left\{X_i^{\mathrm{nu}}\right\}_{i=1}^{n_{\mathrm{nu}}}$. Therefore, the upper-bounding $\partial f\left(r(X_j)\right.$ or model $r(x)$ does not solve the train-loss hacking. However, for these cases, the proposed non-negative

risk correction approach is empirically shown to be effective, as shown in Figure 2 of Section 5.1 and Figure 4 of Appendix F.1.1. In these results, Bounded LSIF and BKL correspond to the upper bounding of the model $r(x)$ and $\partial f(r(X_j)$, respectively. Experimentally, DRE methods without the non-negative correction fail to learn the density ratio, while the non-negative correction succeeded in stabilizing the performance.

4. Theoretical justification of D3RE

In this section, we confirm the validity of D3RE by providing a generalization error bound. We derive two types of guarantees, one in terms of the BD risk and the other the L^2-distance. Given $n \in \mathbb{N}$ and a distribution p, we define the *Rademacher complexity* $\mathcal{R}_n^p$ of a function class $\mathcal{H}$ as $\mathcal{R}_n^p(\mathcal{H}) := \mathbb{E}_p \mathbb{E}_\sigma \left[\sup_{r\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i r(X_i)\right|\right]$, where $\{\sigma_i\}_{i=1}^n$ are independent uniform sign variables and $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p$. We omit r^* from the notation BD_f when there is no ambiguity.

4.1. Generalization error bound on BD

Let $I_r := (b_r, B_r)$. Theorem 4 in Appendix H provides a generalization error bound in terms of the Rademacher complexities of a hypothesis class and the following assumption.

Assumption 3. Following (i)–(iv) hold:

- **(i)** there exists an empirical risk minimizer $\hat{r} \in \arg\min_{r\in\mathcal{H}} \widehat{\mathrm{nnBD}}_f(r)$ and a population risk minimizer $\bar{r} \in \arg\min_{r\in\mathcal{H}} \mathrm{BD}_f(r)$;
- **(ii)** $B_\ell := \sup_{t\in I_r}\{\max\{|\ell_1(t)|, |\ell_2(t)|\}\} < \infty$;
- **(iii)** ℓ_1 (resp. ℓ_2) is L_{ℓ_1}-Lipschitz (resp. L_{ℓ_2}-Lipschitz) on I_r;
- **(iv)** $\inf_{r\in\mathcal{H}}(\mathbb{E}_{\mathrm{de}} - C\mathbb{E}_{\mathrm{nu}})\ell_1(r(X)) > 0$.

For the boundedness and Lipschitz continuity in Assumption 3 to hold for the loss functions involving a logarithm (UKL, BKL, PU), a technical assumption $b_r > 0$ is sufficient.

Then, we introduce Assumption 4 (Golowich et al., 2019, Theorem 1) to bound the complexity of the hypothesis class.

Assumption 4 (Neural networks with bounded complexity)**.** The probability densities p_{nu} and p_{de} have bounded supports: $\sup_{x\in\mathcal{X}^{\mathrm{de}}} \|x\| < \infty$, and a hypothesis class $\mathcal{H}$ consists of real-valued neural networks of depth L over the domain $\mathcal{X}$, where each parameter matrix W_j has the Frobenius norm at most $B_{W_j} \geq 0$ and 1-Lipschitz activation functions φ_j that are positive-homogeneous (i.e., φ_j is applied element-wise and $\varphi_j(\alpha t) = \alpha\varphi_j(t)$ for all $\alpha \geq 0$).

Under Assumption 4, Lemma 3 in Appendix I reveals $\mathcal{R}_{n_{\mathrm{nu}}}^{p_{\mathrm{nu}}}(\mathcal{H}) = \mathcal{O}(1/\sqrt{n_{\mathrm{nu}}})$ and $\mathcal{R}_{n_{\mathrm{de}}}^{p_{\mathrm{de}}}(\mathcal{H}) = \mathcal{O}(1/\sqrt{n_{\mathrm{de}}})$. By combining these results with Theorem 4 in Appendix H, we obtain the following theorem.

Theorem 1 (Generalization error bound for D3RE)**.** *Under Assumptions 3 and 4, for any $\delta \in (0,1)$, we have with probability at least $1-\delta$,*

$$\mathrm{BD}_f(\hat{r}) - \mathrm{BD}_f(\bar{r}) \leq \frac{\kappa_1}{\sqrt{n_{\mathrm{de}}}} + \frac{\kappa_2}{\sqrt{n_{\mathrm{nu}}}} + 2(1+C)B_\ell \exp\left(-\frac{2\alpha^2}{({B_\ell}^2/n_{\mathrm{de}}) + (C^2{B_\ell}^2/n_{\mathrm{nu}})}\right) + B_\ell\sqrt{8\left(\frac{1}{n_{\mathrm{de}}} + \frac{(1+C)^2}{n_{\mathrm{nu}}}\right)\log\frac{1}{\delta}},$$

where κ_1, κ_2 are constants that depend on $C, f, B_{p_{\mathrm{de}}}, B_{p_{\mathrm{nu}}}, L$, and B_{W_j}.

See Remark 5 in Appendix H for more explicit forms of κ_1 and κ_2. This generalization error bound provides theoretical guarantees for various applications. For instance, by defining $f(t) = \log(1-t) + Ct(\log(t) - \log(1-t))$ for $0 < t < 1$, $\mathrm{BD}_f(r)$ becomes the risk functional of PU learning (see Appendix A for the derivation). Then, the generalization error bound provides a classification error bound for PU learning, which is a special case of the binary classification problem Kiryo et al. (2017). Note that the dependency of the bound on $f, B_{p_{\mathrm{de}}}, B_{p_{\mathrm{nu}}}, L$, and B_{W_j} is typical for classification with Lipschitz functions (Bartlett & Mendelson, 2003, Corollary 15). The third term of the RHS corresponds to the bias caused by the use of non-negative correction.

4.2. Estimation error bound on L^2 norm

Next, we derive an estimation error bound for $\hat{r}$ on the L^2 norm. We aim to derive the standard convergence rate of non-parametric regression; that is, under the appropriate conditions, the order of $\|\hat{r} - r^*\|_{L^2(p_{\mathrm{de}})}$ is nearly $\mathcal{O}_{\mathbb{P}}(1/(n_{\mathrm{de}} \wedge n_{\mathrm{nu}}))$ (Kanamori et al., 2012). Note that unlike the generalization error bound of the BD, we require a stronger assumption on the loss function, namely, a strong convexity. In Theorem 1, for a multilayer perception with ReLU activation function (Definition 3), we derive the convergence rate of the L^2 distance, which is the same rate as that of the nonparametric regression using the Gaussian kernel and the LSIF loss (Kanamori et al., 2012). This result also corresponds to a faster convergence rate than Theorem 1. The proof is shown in Appendix J. To complement this result, we empirically investigate the estimator error using an artificially generated dataset with the known true density ratio in Section 5.2.

Theorem 2 (L^2 Convergence rate)**.** *Assume f is μ-strongly convex. Let $\mathcal{H}$ be defined as in Definition 3 and assume*

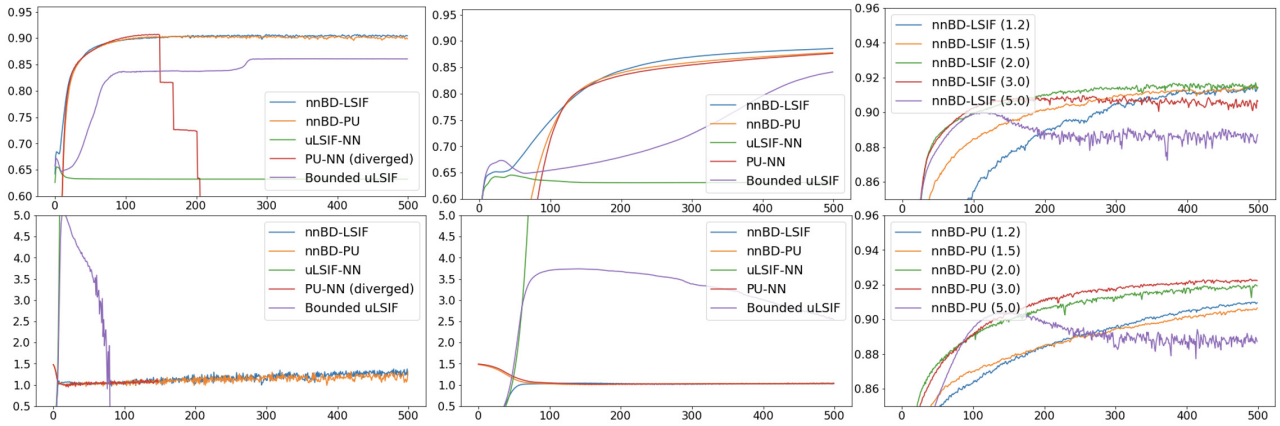

Figure 2. Results of Section 5.1. The horizontal axes represent epochs. Left and center figures: The results under different learning rates, 1×10^{-4} and 1×10^{-5}, respectively, where the vertical axes of the upper graphs show the AUROCs, and those of the lower graphs show $\hat{\mathbb{E}}_{\mathrm{de}}[\hat{r}(X_i)]$. Right figure: Results of sensitivity analyses, where the vertical axes show the AUROCs.

$r^* = \frac{p_{\mathrm{nu}}}{p_{\mathrm{de}}} \in \mathcal{H}$. *In addition, assume the same conditions as Theorem 3. Then, for any* $0 < \gamma < 2$, *as* $n_{\mathrm{de}}, n_{\mathrm{nu}} \to \infty$,

$$\|\hat{r} - r^*\|_{L^2(p_{\mathrm{de}})} \leq \mathcal{O}_{\mathbb{P}}\left(\left(\min\{n_{\mathrm{de}}, n_{\mathrm{nu}}\}\right)^{-1/(2+\gamma)}\right).$$

This L^2 distance bound is useful for statistical inference. For example, double/debiased machine learning with cross-fitting proposed by Chernozhukov et al. (2016) allows for semiparametric inference under estimators of nuisance parameters with appropriate convergence rates. By using cross-fitting, Uehara et al. (2020) proposed causal inference under covariate shifts when the convergence rate of the density ratio satisfies an appropriate convergence rate. Further, it is expected to be applied to two-sample homogeneity using neural networks, as in Kanamori et al. (2010).

5. Experiments

We experimentally show how existing DRE methods fail and D3RE succeeds when using neural networks[2].

5.1. Experiments with image data

To investigate how D3RE prevents train-loss hacking from occurring, we consider a setting of inlier-based outlier detection. For binary labels $y \in \{-1, +1\}$, we consider training a classifier only from $p(X \mid y = +1)$ and $p(X)$ to find a positive data point in the test data sampled from $p(X)$. The goal is to maximize the area under the receiver operating characteristic (AUROC) curve, which is a criterion often used for anomaly detection, by estimating the density ratio $r^*(X) = p(X \mid y = +1)/p(X)$. We construct positive and negative datasets from the `CIFAR-10` (Krizhevsky, 2009) dataset with 10 classes. The positive dataset comprises 'airplane,' 'automobile,' 'ship,' and 'truck'; the negative dataset comprises 'bird,' 'cat,' 'deer,' 'dog,' 'frog,' and 'horse.' We use $1,000$ positive samples generated from $p(X \mid y = +1)$ and $1,000$ unlabeled samples generated from $p(X)$ to train the models. Then, we calculate the AUROCs using $10,000$ test samples generated from $p(X)$. In this case, it is desirable to set $C < \frac{1}{2}$ because $\frac{p(X|y=+1)}{0.5p(X|y=+1)+0.5p(X|y=-1)} = \frac{1}{0.5+0.5\frac{p(X|y=-1)}{p(X|y=+1)}}$. For demonstrative purposes, we use a basic CNN architecture from the PyTorch tutorial (Paszke et al., 2019). The details are shown in Appendix D.2. The model is trained by the Adam optimizer (Kingma & Ba, 2015) without a weight decay and with the parameters $(\beta_1, \beta_2, \epsilon)$ fixed at the default values of the implementation in PyTorch (Paszke et al., 2019), namely $(0.9, 0.999, 10^{-8})$.

First, we compare two of the proposed estimators, $\widehat{\mathrm{nnBD}}_{\mathrm{PU}}$ (nnBD-PU) and $\widehat{\mathrm{nnBD}}_{\mathrm{LSIF}}$ (nnBD-LSIF), with two existing estimators, $\widehat{\mathrm{BD}}_{\mathrm{PU}}$ (PU-NN) and $\widehat{\mathrm{BD}}_{\mathrm{LSIF}}$ (uLSIF-NN) with neural networks. We use the logistic loss for PULogLoss. In addition, we conduct an experiment with uLSIF-NN using a naively capped model $\tilde{r}(X) = \min\{r(X), 1/C\}$ (Bounded uLSIF). We fix the hyperparameter C at $1/3$. We report the results for two learning rates, 1×10^{-4} and 1×10^{-5}. We conduct 10 trials, and report the average AUROCs. We also compute $\hat{\mathbb{E}}_{\mathrm{de}}[\hat{r}(X)]$, which should be close to 1 when the density ratio is successfully estimated since $\int (p(x|y = +1)/p(X))p(X)dx = 1$. These results are shown in the left and center figures in Figure 2. In all cases, the proposed estimators outperform the other methods. We consider that the instabilities of PU-NN and LISF-NN are caused by the unboundedness of the objective function (also see Kiryo et al. (2017), where similar experimental results are reported). The results also demonstrate that naive

[2] A code of the conducted experiments is available at `https://github.com/MasaKat0/D3RE`.

Table 2. Results of Section 5.2: MSEs and SDs of DRE using synthetic datasets. The lowest MSE methods are highlighted in bold.

		uLSIF	LSIF-NN	D3RE (nnBD-LSIF)								
				$C=0.8$	$C=1$	$C=2$	$C=3$	$C=4$	$C=5$	$C=10$	$C=15$	$C=20$
$\dim=10$	MSE	2.378	1.272	1.750	1.695	1.191	0.964	0.873	**0.833**	0.948	1.079	1.170
	SD	1.143	0.413	0.570	0.563	0.523	0.487	0.459	0.424	0.370	0.331	0.387
$\dim=20$	MSE	1.684	2.694	1.704	1.646	1.307	**1.272**	1.337	1.444	2.066	2.697	3.098
	SD	0.372	0.409	0.380	0.368	0.328	0.297	0.283	0.288	0.285	0.346	0.374
$\dim=30$	MSE	1.786	3.724	1.811	1.747	**1.488**	1.577	1.798	2.019	3.238	4.306	5.432
	SD	0.456	0.460	0.459	0.449	0.411	0.400	0.401	0.379	0.370	0.464	0.543
$\dim=50$	MSE	1.791	8.717	1.817	1.753	**1.609**	1.818	2.194	2.614	4.848	6.955	8.798
	SD	0.562	1.518	0.571	0.555	0.513	0.503	0.484	0.465	0.488	0.597	0.672
$\dim=100$	MSE	1.723	4.849	1.748	1.693	**1.626**	1.860	2.226	2.709	5.528	8.605	11.557
	SD	0.574	4.182	0.575	0.571	0.540	0.532	0.495	0.563	0.672	0.790	1.140

capping (Bounded uLSIF) fails to prevent train-loss hacking from occurring and leads to suboptimal behavior. As discussed in Sections 2.3 and 3.2, naive capping is insufficient for this problem because an unreasonable model such that $r(X_i^{\mathrm{de}}) = 0$ and $r(X_j^{\mathrm{nu}}) = 1/C$ can still be a minimizer by decreasing one part of the empirical BD, e.g., $\frac{1}{2}\widehat{\mathbb{E}}_{\mathrm{de}}\left[r^2(X)\right] - \frac{C}{2}\widehat{\mathbb{E}}_{\mathrm{nu}}\left[r^2(X)\right] = 0 - \frac{1}{2C}$.

Next, we investigate the sensitivity of D3RE to the hyperparameter C. We choose C from $\{1/1.2, 1/1.5, 1/2.0.1/3.0, 1/5.0\}$. The other settings remain unchanged from the previous experiment, where the exact upper bound $\overline{R}$ is 2.0. The results are shown on the right-hand side of Figure 2. While estimators with $1/C \simeq 2.0$ show a superior performance, the method is robust to the choice of C to a certain extent. Additional experimental results are reported in Appendix F.1.

Note that this experimental setting is similar to that of PU learning (Elkan & Noto, 2008; Kiryo et al., 2017). In PU learning experiments, we mainly consider a binary classification problem, and the class-prior $p(y=+1)$ is given; that is, the goals and the presence of the information are the differences between the experimental settings of inlier-based outlier detection and PU learning. In this paper, we successfully related PU learning methods to DRE. The class-prior in PU learning plays a similar role to the upper bound of r^* in DRE.

5.2. Experiments on L^2 error

We empirically investigate the L^2 error in the proposed D3RE. We compare our method with the uLSIF. For uLSIF (Kanamori et al., 2009), we use an open-source implementation[3], which uses a linear-in-parameter model with the Gaussian kernel (Kanamori et al., 2012). For D3RE, we use nnBD-LSIF and 3-layer perceptron with a ReLU activation function, where the number of the nodes in the middle layer is 100. We conducted nnBD-LSIF for all $C \in \{0.8, 1, 2, 3, 4, 5, 10, 15, 20\}$. We also compare these methods with a naively implemented LSIF with a 3-layer perceptron. Let the dimensions of the domain be d and

$$p_{\mathrm{nu}}(X) = \mathcal{N}(X; \mu^{\mathrm{nu}}, I_d),\ p_{\mathrm{de}}(X) = \mathcal{N}(X; \mu^{\mathrm{de}}, I_d),$$

where $\mathcal{N}(X; \mu, \Sigma)$ denotes the multivariate normal distribution with mean μ and Σ, μ^{nu} and μ^{de} are d-dimensional vectors $\mu^{\mathrm{nu}} = (1, 0, \ldots, 0)^\top$ and $\mu^{\mathrm{de}} = (0, 0, \ldots, 0)^\top$, and I_d is a d-dimensional identity matrix. We fix the sample sizes at $n_{\mathrm{nu}} = n_{\mathrm{de}} = 1,000$ and estimate the density ratio using uLSIF, LSIF, and D3RE (nnBD-LSIF). To measure the performance, we use the mean squared error (MSE) and the standard deviation (SD) averaged over 50 trials. Note that in this setting, we know the true density ratio r^*. The results are shown in Table 2. The proposed nnBD-LSIF method estimates the density ratio more accurately than the other methods with a lower MSE. In many cases of the results, nnBD-LSIF achieves the best performance at approximately $C = 2$. This result implies that we do not need to know the exact C to achieve a high level performance in D3RE.

6. Inlier-based outlier detection

As an application of D3RE, we perform *inlier-based outlier detection* experiments with benchmark datasets. In addition to `CIFAR-10`, we use `MNIST` (LeCun et al., 1998) and `fashion-MNIST` (`FMNIST`) (Xiao et al., 2017), both of which have 10 classes. Hido et al. (2008; 2011) applied the a direct DRE for inlier-based outlier detection; that is, finding outliers in a test set based on a training set consisting only of inliers by using the ratio of training and test data densities as an outlier score. Nam & Sugiyama (2015) and Abe & Sugiyama (2019) proposed using shallow neural networks with DRE to deal with this problem. In relation to the experimental setting of Section 5.1, the problem setting can be seen as a transductive variant of PU learning (Kato et al., 2019).

We follow the setting proposed by Golan & El-Yaniv (2018). There are ten classes in each dataset, `MNIST`, `CIFAR-10`, and `FMNIST`. We use one class as an inlier class and treat

[3] https://github.com/hoxo-m/densratio_py.

Table 3. Average AUROC curve (Mean) with the standard deviation (SD) over 5 trials of anomaly detection methods. For all datasets, each model was trained on a single class and tested against all other classes. The best result is in bold.

MNIST	uLSIF-NN		nnBD-LSIF		nnBD-PU		nnBD-LSIF		nnBD-PU		Deep SAD		GT	
Network	LeNet		LeNet		LeNet		WRN		WRN		LeNet		WRN	
Inlier Class	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	0.999	0.000	0.997	0.000	0.999	0.000	**1.000**	0.000	**1.000**	0.000	0.592	0.051	0.963	0.002
1	**1.000**	0.000	0.999	0.000	**1.000**	0.000	**1.000**	0.000	**1.000**	0.000	0.942	0.016	0.517	0.039
2	0.997	0.001	0.994	0.000	0.997	0.001	**1.000**	0.000	**1.000**	0.001	0.447	0.027	0.992	0.001
3	0.997	0.000	0.995	0.001	0.998	0.000	**1.000**	0.000	**1.000**	0.000	0.562	0.035	0.974	0.001
4	0.998	0.000	0.997	0.001	0.999	0.000	**1.000**	0.000	**1.000**	0.000	0.646	0.015	0.989	0.001
CIFAR-10	uLSIF-NN		nnBD-LSIF		nnBD-PU		nnBD-LSIF		nnBD-PU		Deep SAD		GT	
Network	LeNet		LeNet		LeNet		WRN		WRN		LeNet		WRN	
Inlier Class	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
plane	0.745	0.056	0.934	0.002	**0.943**	0.001	0.925	0.004	0.923	0.001	0.627	0.066	0.697	0.009
car	0.758	0.078	0.957	0.002	**0.968**	0.001	0.965	0.002	0.960	0.001	0.606	0.018	0.962	0.003
bird	0.768	0.012	0.850	0.007	**0.878**	0.004	0.844	0.004	0.858	0.004	0.404	0.006	0.752	0.002
cat	0.745	0.037	0.820	0.003	**0.856**	0.002	0.810	0.009	0.841	0.002	0.517	0.018	0.727	0.014
deer	0.758	0.036	0.886	0.004	**0.909**	0.002	0.864	0.008	0.872	0.002	0.704	0.052	0.863	0.014
FMNIST	uLSIF-NN		nnBD-LSIF		nnBD-PU		nnBD-LSIF		nnBD-PU		Deep SAD		GT	
Network	LeNet		LeNet		LeNet		WRN		WRN		LeNet		WRN	
Inlier Class	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
T-shirt/top	0.960	0.005	0.981	0.001	**0.985**	0.000	0.984	0.001	0.982	0.000	0.558	0.031	0.890	0.007
Trouser	0.961	0.010	0.998	0.000	**1.000**	0.000	0.998	0.000	0.998	0.000	0.758	0.022	0.974	0.004
Pullover	0.944	0.012	0.976	0.001	0.980	0.001	**0.983**	0.002	0.972	0.001	0.617	0.046	0.902	0.005
Dress	0.973	0.006	0.986	0.001	**0.992**	0.000	0.991	0.001	0.986	0.000	0.525	0.038	0.843	0.014
Coat	0.958	0.006	0.978	0.001	**0.983**	0.000	0.981	0.002	0.974	0.000	0.627	0.029	0.885	0.003

all other classes as outliers. For example, in the case of CIFAR-10, there are $5,000$ train data per class. On the other hand, there are $1,000$ test data for each class, which amounts to $1,000$ inlier samples and $9,000$ outlier samples. The AUROC is used as a metric to evaluate whether the outlier class can be detected in the outlier samples. We compare the proposed methods with the benchmark methods of deep semi-supervised anomaly detection (DeepSAD) (Ruff et al., 2020) and geometric transformation (GT) (Golan & El-Yaniv, 2018). The details of each method are shown in Appendix E. To make a fair comparison, we use LeNet and Wide ResNet for D3RE, which are the same neural network architectures as those used in Golan & El-Yaniv (2018) and Ruff et al. (2020). The detailed structures are shown in Appendix D. Owing to the space limitation, some of the experimental results with MNIST, CIFAR-10 and FMNIST is shown in Table 3. The full results are shown in Table 4 in Appendix F.2. In almost all cases, the average AUROCs of the proposed methods are better than those of the existing methods. The largest performance gain is seen in the CIFAR-10, where the mean AUROC is improved by 0.157 on average between the uLSIF-NN and nnBD-LSIF. Although GT and DeepSAD are designed for different problem setups, to the best of our knowledge, there are no other appropriate state-of-the-art alternatives to these algorithms under this setting.

In Appendix G, we also introduce other applications such as covariate shift adaptation.

Togashi et al. (2021) applied our proposed method to personalized ranking from implicit feedback in a recommender systems. For this task, there are two approaches, pointwise and pairwise, and the former of which is known to be computationally efficient, whereas the latter shows better accuracy than the former. In that study, they reformulated a pointwise approach using the density ratio and also added the essence of the pairwise approach.

7. Conclusion

We proposed a non-negative correction to the empirical BD for DRE. Using the prior knowledge of the upper bound of the density ratio, we can prevent train-loss hacking from occurring when using flexible models. In our theoretical analyses, we provided generalization error bounds for the proposed method. In our experiments, we empirically confirmed the effectiveness of our proposed approach.

Acknowledgments

The authors would like to thank Hirono Okamoto for his constructive advice.
TT was supported by Masason Foundation.

References

Abe, M. and Sugiyama, M. Anomaly detection by deep direct density ratio estimation. *openreview*, 2019.

Adams, R. P. Bayesian online changepoint detection, 2007.

Ali, S. M. and Silvey, S. D. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society*, Series B(28):131–142, 1966.

Ansari, A. F., Ang, M. L., and Soh, H. Refining deep generative models via wasserstein gradient flows. In *ICLR*, 2020.

Athey, S. and Wager, S. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. In *Computational Learning Theory*, volume 2111, pp. 224–240. Springer Berlin Heidelberg, 2001.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

Basseville, M. and Nikiforov, I. V. *Detection of abrupt changes: theory and application*. Prentice Hall information and system sciences. Prentice Hall, 1993.

Bengio, Y., Ducharme, R., and Vincent, P. A neural probabilistic language model. In *NeurIPS*, pp. 932–938. MIT Press, 2001.

Beygelzimer, A. and Langford, J. The offset tree for learning with partial labels. In *KDD*, pp. 129–138, 2009.

Bibaut, A., Malenica, I., Vlassis, N., and Van Der Laan, M. More efficient off-policy evaluation through regularized targeted learning. In *ICML*. PMLR, 2019.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning under covariate shift. *J. Mach. Learn. Res.*, 10: 2137–2155, December 2009. ISSN 1532-4435.

Blitzer, J., Dredze, M., and Pereira, F. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, June 2007.

Bregman, L. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200 – 217, 1967. ISSN 0041-5553.

Brodsky, E. and Darkhovsky, B. *Nonparametric Methods in Change Point Problems*. Mathematics and Its Applications. Springer Netherlands, 1993.

Chen, M., Xu, Z., Weinberger, K. Q., and Sha, F. Marginalized denoising autoencoders for domain adaptation. In *ICML*, ICML'12, pp. 1627–1634, Madison, WI, USA, 2012. Omnipress.

Cheng, k.-F. and Chu, C. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10, 08 2004.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. Locally robust semiparametric estimation, 2016.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68, 2018.

Cole, S. R. and Stuart, E. A. Generalizing evidence from randomized clinical trials to target populations. *American Journal of Epidemiology*, 172(1):107–115, 2010.

Cortes, C. and Mohri, M. Domain adaptation in regression. In *Algorithmic Learning Theory*, pp. 308–323, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 1967.

du Plessis, M. C., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In *ICML*, pp. 1386–1394, 2015.

Dudík, M., Langford, J., and Li, L. Doubly Robust Policy Evaluation and Learning. In *ICML*, pp. 1097–1104, 2011.

Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In *ICDM*, pp. 213–220, 2008.

Fang, T., Lu, N., Niu, G., and Sugiyama, M. Rethinking importance weighting for deep learning under distribution shift. In *NeurIPS*, 2020.

Garnett, R., Osborne, M. A., and Roberts, S. J. Sequential bayesian prediction in the presence of changepoints. In *ICML*, pp. 345–352, New York, NY, USA, 2009. Association for Computing Machinery.

Golan, I. and El-Yaniv, R. Deep anomaly detection using geometric transformations. In *NeurIPS*, pp. 9758–9769. Curran Associates, Inc., 2018.

Golowich, N., Rakhlin, A., and Shamir, O. Size-Independent Sample Complexity of Neural Networks. *arXiv:1712.06541 [cs, stat]*, November 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NeurIPS*, pp. 2672–2680. Curran Associates, Inc., 2014.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning, 131-160 (2009)*, 01 2009.

Gustafsson, M. G. L. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. *Journal of Microscopy*, 198(2):82–87, 2000.

Hainmueller, J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 2012.

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2001.

He, K., Xiangyu Zhang, S. R., and Sun, J. Deep residual learning for image recognition. In *CoRR*, 2015.

Hellinger, E. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909.

Henmi, M. and Eguchi, S. A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*, 2004.

Henmi, M., Yoshida, R., and Eguchi, S. Importance Sampling Via the Estimated Sampler. *Biometrika*, 2007.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., and Kanamori, T. Inlier-based outlier detection via direct density ratio estimation. In *ICDM*, 2008.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., and Kanamori, T. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, 26(2):309–336, Feb 2011.

Hirano, K., Imbens, G. W., and Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. J. Correcting sample selection bias by unlabeled data. In *NeurIPS*, pp. 601–608. MIT Press, 2007.

Imai, K. and Ratkovic, M. Covariate balancing propensity score. *J. R. Statist. Soc. B*, 76(1):243–263, 2014.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pp. 448–456, 2015.

Kallus, N. and Uehara, M. Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In *NeurIPS*, 2019.

Kanamori, T., Hido, S., and Sugiyama, M. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul.):1391–1445, 2009.

Kanamori, T., Suzuki, T., and Sugiyama, M. f -divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58, 10 2010.

Kanamori, T., Suzuki, T., and Sugiyama, M. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3), March 2012.

Kato, M. Identifying different definitions of future in the assessment of future economic conditions: Application of pu learning and text mining. *arXiv*, 2019.

Kato, M., Teshima, T., and Honda, J. Learning from positive and unlabeled data with a selection bias. In *ICLR*, 2019.

Kawahara, Y. and Sugiyama, M. Change-point detection in time-series data by direct density-ratio estimation. In *ICDM*, 2009.

Keziou., A. Utilisation des divergences entre mesures en statistique inferentielle. *PhD thesis*, 2003.

Keziou, A. and Leoni-Aubin, S. Test of homogeneity in semiparametric two-sample density ratio models. *Comptes Rendus Mathematique - C R MATH*, 340:905–910, 06 2005.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. In *NeurIPS*, 2017.

Klaassen, C. A. J. Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.*, 1987.

Krizhevsky, A. Learning multiple layers of features from tiny images, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pp. 1097–1105. Curran Associates, Inc., 2012.

Kullback, S. and Leibler, R. A. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 1998.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin, 1991.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pp. 661–670, 2010.

Liu, S., Yamada, M., Collier, N., and Sugiyama, M. Change-point detection in time-series data by relative density-ratio estimation. In *Structural, Syntactic, and Statistical Pattern Recognition*, pp. 363–372, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting, 2014.

Loevinger, J. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, 45(6):507–529, 1948. ISSN 0033-2909.

Lu, N., Zhang, T., Niu, G., and Sugiyama, M. Mitigating overfitting in supervised classification from two unlabeled datasets: A consistent risk correction approach. *arXiv:1910.08974 [cs, stat]*, March 2020.

McDiarmid, C. On the method of bounded differences. In *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, London Mathematical Society Lecture Note Series, pp. 148–188. Cambridge University Press, 1989.

Menon, A. and Ong, C. S. Linking losses for density ratio and class-probability estimation. In *ICML*, volume 48, pp. 304–313, New York, New York, USA, 2016.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts, second edition, 2018.

Nam, H. and Sugiyama, M. Direct density ratio estimation with convolutional neural networks with application in outlier detection. *IEICE Transactions on Information and Systems*, E98.D(5):1073–1079, 2015.

Narita, Y., Yasui, S., and Yata, K. Efficient counterfactual learning from bandit feedback. *AAAI*, 2019.

Nguyen, M. N., Li, X.-L., and Ng, S.-K. Positive unlabeled leaning for time series classification. In *IJCAI*, pp. 1421–1426, 2011.

Nguyen, X., Wainwright, M., and Jordan, M. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE*, 2010.

Niu, G., du Plessis, M. C., Sakai, T., Ma, Y., and Sugiyama, M. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *NeurIPS*, pp. 1199–1207, 2016.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In *NeurIPS*, pp. 271–279. Curran Associates, Inc., 2016.

Oberst, M. and Sontag, D. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *ICML*, volume 97, pp. 4881–4890, 2019.

Paquet, U. Empirical bayesian change point detection. *Graphical Models*, 1995, 01 2007.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035. Curran Associates, Inc., 2019.

Pearl, J. and Bareinboim, E. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29, 2014.

Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Magazine*, 5(50):157–175, 1900.

Qin, J. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

Qin, J. and Zhang, B. Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society*, 2007.

Reddi, S. J., Póczos, B., and Smola, A. J. Doubly robust covariate shift correction. In *AAAI*, pp. 2949–2955. AAAI Press, 2015.

Rhodes, B., Xu, K., and Gutmann, M. Telescoping density-ratio estimation. In *NeurIPS*, 2020.

Rosenbaum, P. R. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

Rosenbaum, P. R. Model-based direct adjustment. *Journal of the American Statistical Association*, 82:387–394, 1987.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., Müller, K.-R., and Kloft, M. Deep semi-supervised anomaly detection. In *ICLR*, 2020.

Salton, G. and McGill, M. J. *Introduction to modern information retrieval*. McGraw-Hill, Inc., 1986.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, 48(4):1875–1897, 2020.

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

Smola, A., Song, L., and Teo, C. H. Relative novelty detection. In *AISTATS*, volume 5 of *Proceedings of Machine Learning Research*, pp. 536–543, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 2009. PMLR.

Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 02 2008.

Sugiyama, M., Suzuki, T., Itoh, Y., Kanamori, T., and Kimura, M. Least-squares two-sample test. *Neural networks : the official journal of the International Neural Network Society*, 24:735–51, 04 2011a.

Sugiyama, M., Suzuki, T., and Kanamori, T. Density ratio matching under the bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64, 10 2011b.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, New York, NY, USA, 1st edition, 2012.

Tan, Z. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 2010.

Togashi, R., Kato, M., Otani, M., and Satoh, S. Density-ratio based personalised ranking from implicit feedback. In *The World Wide Web Conference*, 2021.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., and Sugiyama, M. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009. doi: 10.2197/ipsjjip.17.138.

Uehara, M., Sato, I., Suzuki, M., Nakayama, K., and Matsuo, Y. Generative adversarial tets from a density ratio estimation perspective, 2016.

Uehara, M., Kato, M., and Yasui, S. Off-policy evaluation and learning for external validity under a covariate shift. In *NeurIPS*, 2020.

van de Geer, S. *Empirical Processes in M-Estimation*, volume 6. Cambridge university press, 2000.

Vapnik, V. N. *Statistical Learning Theory*. Wiley, September 1998.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 1st edition, 2019.

Wang, Y.-X., Agarwal, A., and Dudik, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *ICML*, pp. 3589–3597, 2017.

Wyss, R., Ellis, A. R., Brookhart, M. A., Girman, C. J., Jonsson Funk, M., LoCasale, R., and Stürmer, T. The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-Balancing Propensity Score. *American Journal of Epidemiology*, 2014.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv*, abs/1708.07747, 2017.

Yamada, M. and Sugiyama, M. Direct importance estimation with gaussian mixture models. *IEICE Transactions on Information and Systems*, E92.D(10):2159–2162, 2009. doi: 10.1587/transinf.E92.D.2159.

Yamada, M., Sugiyama, M., Wichern, G., and Simm, J. Direct importance estimation with a mixture of probabilistic principal component analyzers. *IEICE Transactions*, 93-D:2846–2849, 10 2010.

Yamanishi, K. and Takeuchi, J. A unifying framework for detecting outliers and change points from non-stationary time series data. In *KDD*, 2002.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12. BMVA Press, September 2016.

Zheng, W. and van der Laan, M. J. Cross-validated targeted minimum-loss-based estimation. *Targeted Learning: Causal Inference for Observational and Experimental Data*, 2011.