A. Details of existing methods for DRE

In this section, we overview examples of DRE methods in the framework of the density ratio matching under BD.

Least Squares Importance Fitting (LSIF). LSIF minimizes the squared error between a density ratio model r and the true density ratio r^* defined as follows (Kanamori et al., 2009):

$$R_{\mathrm{LSIF}}(r) = \mathbb{E}_{\mathrm{de}}[(r(X) - r^*(X))^2] = \mathbb{E}_{\mathrm{de}}[(r^*(X))^2] - 2\mathbb{E}_{\mathrm{nu}}[r(X)] + \mathbb{E}_{\mathrm{de}}[(r(X))^2].$$

In the *unconstrained LSIF* (uLSIF) (Kanamori et al., 2009), we ignore the first term in the above equation and estimate the density ratio by the following minimization problem:

$$\hat{r} = \operatorname*{arg\,min}_{r \in \mathcal{H}} \left[\frac{1}{2}\hat{\mathbb{E}}_{\mathrm{de}}[(r(X))^2] - \hat{\mathbb{E}}_{\mathrm{nu}}[r(X)] + \mathcal{R}(r) \right], \tag{6}$$

where $\mathcal{R}$ is a regularization term. This empirical risk minimization is equal to minimizing the empirical BD defined in (2) with $f(t) = (t-1)^2/2$.

Unnormalized Kullback–Leibler (UKL) divergence and KL Importance Estimation Procedure (KLIEP). The KL importance estimation procedure (KLIEP) is derived from the unnormalized Kullback–Leibler (UKL) divergence objective (Sugiyama et al., 2008; Nguyen et al., 2010; Tsuboi et al., 2009; Yamada & Sugiyama, 2009; Yamada et al., 2010), which uses $f(t) = t\log(t) - t$. Ignoring the terms which are irrelevant for the optimization, we obtain the unnormalized Kullback–Leibler (UKL) divergence objective (Nguyen et al., 2010; Sugiyama et al., 2012) as

$$\mathrm{BD}_{\mathrm{UKL}}(r) = \mathbb{E}_{\mathrm{de}}\left[r(X)\right] - \mathbb{E}_{\mathrm{nu}}\left[\log\left(r(X)\right)\right].$$

Directly minimizing UKL is proposed by Nguyen et al. (2010). The KLIEP also solves the same problem with further imposing a constraint that the ratio model $r(X)$ is non-negative for all X and is normalized as

$$\hat{\mathbb{E}}_{\mathrm{de}}\left[r(X)\right] = 1.$$

Then, following is the optimization criterion of KLIEP (Sugiyama et al., 2008):

$$\max_{r} \hat{\mathbb{E}}_{\mathrm{nu}}\left[\log\left(r(X)\right)\right]$$
$$\text{s.t. } \hat{\mathbb{E}}_{\mathrm{de}}\left[r(X)\right] = 1 \text{ and } r(X) \geq 0 \text{ for all } X.$$

Logistic Regression (LR). By using $f(t) = \log(t) - (1+t)\log(1+t)$, we obtain the following BD called the binary Kullback–Leibler (BKL) divergence:

$$\mathrm{BD}_{\mathrm{BKL}}(r) = -\mathbb{E}_{\mathrm{de}}\left[\log\left(\frac{1}{1+r(X)}\right)\right] - \mathbb{E}_{\mathrm{nu}}\left[\log\left(\frac{r(X)}{1+r(X)}\right)\right].$$

This BD is derived from a formulation based on the logistic regression (Hastie et al., 2001; Sugiyama et al., 2011b).

PU Learning with the log loss. Consider a binary classification problem and let X and $y \in \{\pm 1\}$ be the feature and the label of a sample, respectively. In PU learning, the goal is to train a classifier only using positive data sampled from $p(X \mid y = +1)$, and unlabeled data sampled from $p(X)$ in binary classification (Elkan & Noto, 2008). More precisely, this problem setting of PU learning is called the *case-control scenario* (Elkan & Noto, 2008; Niu et al., 2016). Let $\mathcal{G}$ be the set of measurable functions from $\mathcal{X}$ to $[\epsilon, 1-\epsilon]$, where $\epsilon \in (0, 1/2)$ is a small positive value. For a loss function $\ell : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}^+$, du Plessis et al. (2015) showed that the classification risk of $g \in \mathcal{G}$ in the PU problem setting can be expressed as

$$R_{\mathrm{PU}}(g) = \pi \int \Big(\ell(g(X), +1) - \ell(g(X), -1)\Big) p(X \mid y = +1) dX + \int \ell(g(X), -1)] p(X) dX. \tag{7}$$

According to Kato et al. (2019), we can derive the following risk for DRE from the risk for PU learning (7) as follows:

$$\mathrm{BD}_{\mathrm{PU}}(g) = \frac{1}{R}\mathbb{E}_{\mathrm{nu}}\left[-\log\left(g(X)\right) + \log\left(1 - g(X)\right)\right] - \mathbb{E}_{\mathrm{de}}\left[\log\left(1 - g(X)\right)\right],$$

and Kato et al. (2019) showed that $g^* = \arg\min_{g \in \mathcal{G}} \mathrm{BD}_{\mathrm{PU}}(g)$ satisfies the following:

Proposition 1. It holds almost everywhere that

$$g^*(X) = \begin{cases} 1-\varepsilon & (X \notin D_2), \\ C\frac{p_{\mathrm{nu}}(X)}{p_{\mathrm{de}}(X)} & (X \in D_1 \cap D_2), \\ \varepsilon & (X \notin D_1), \end{cases}$$

where $C = \frac{1}{R}$, $D_1 = \{X \mid Cp_{\mathrm{nu}}(X) \geq \epsilon p_{\mathrm{de}}(X)\}$, and $D_2 = \{X | Cp_{\mathrm{nu}}(X) \leq (1-\epsilon)p_{\mathrm{de}}(X)\}$.

Using this result, we define the empirical version of $\mathrm{BD}_{\mathrm{PU}}(g)$ as follows:

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

To see that this is also a BD minimization method, define $f(t)$ as

$$f(t) = C\log(1-t) + Ct\left(\log(t) - \log(1-t)\right).$$

Then, we have

$$\partial f(t) = -\frac{C}{1-t} + C(\log(t) - \log(1-t)) + Ct\left(\frac{1}{t} + \frac{1}{1-t}\right).$$

Therefore, we have

$$\begin{aligned}
&\mathrm{BD}_f(r) := \mathbb{E}_{\mathrm{de}}\Big[\partial f\big(r(X_i)\big)r(X_i) - f\big(r(X_i)\big)\Big] - \mathbb{E}_{\mathrm{nu}}\Big[\partial f\big(r(X_j)\big)\Big] \\
&= \mathbb{E}_{\mathrm{de}}\Big[-\frac{Cr(X_i)}{1-r(X_i)} + Cr(X_i)(\log(r(X_i)) - \log(1-r(X_i))) + Cr^2(X_i)\left(\frac{1}{r(X_i)} + \frac{1}{1-r(X_i)}\right)\Big] \\
&\quad - \mathbb{E}_{\mathrm{de}}\Big[\log(1-r(X_i)) + Cr(X_i)\left(\log(r(X_i)) - \log(1-r(X_i))\right)\Big] \\
&\quad - \mathbb{E}_{\mathrm{nu}}\Big[-\frac{C}{1-r(X_i)} + C(\log(r(X_i)) - \log(1-r(X_i))) + Cr(X_i)\left(\frac{1}{r(X_i)} + \frac{1}{1-r(X_i)}\right)\Big] \\
&= \mathbb{E}_{\mathrm{de}}\Big[-\frac{Cr(X_i)}{1-r(X_i)} + Cr(X_i)(\log(r(X_i)) - \log(1-r(X_i))) + \frac{Cr(X_i)}{1-r(X_i)}\Big] \\
&\quad - \mathbb{E}_{\mathrm{de}}\Big[\log(1-r(X_i)) + Cr(X_i)\left(\log(r(X_i)) - \log(1-r(X_i))\right)\Big] \\
&\quad - \mathbb{E}_{\mathrm{nu}}\Big[-\frac{C}{1-r(X_i)} + C(\log(r(X_i)) - \log(1-r(X_i))) + \frac{C}{1-r(X_i)}\Big] \\
&= \mathbb{E}_{\mathrm{de}}\Big[\log(1-r(X_i))\Big] - C\mathbb{E}_{\mathrm{nu}}\Big[\log(r(X_i)) - \log(1-r(X_i))\Big].
\end{aligned}$$

Remark 1 (DRE and PU learning)**.** Menon & Ong (2016) showed that minimizing a proper CPE loss is equivalent to minimizing a BD to the true density ratio, and demonstrated the viability of using existing losses from one problem for the other for CPE and DRE. Kato et al. (2019) pointed out the relation between the PU learning and density ratio estimation and leveraged it to solve a sample selection bias problem in PU learning. In this paper, we introduced the BD with $f(t) = \log(1-Ct) + Ct\left(\log(Ct) - \log(1-Ct)\right)$, inspired by the objective function of PU learning with the log loss. In the terminology of Menon & Ong (2016), this f results in a DRE objective without a *link function*. In other words, it yields a direct DRE method.

B. Examples of $\tilde{f}$

Here, we show the examples of $\tilde{f}$ such that $\partial f(t) = C\big(\partial f(t)t - f(t)\big) + \tilde{f}(t)$, where $\tilde{f}(t)$ is bounded from above, and $\partial f(t)t - f(t) + A$ is non-negative.

First, we consider $f(t) = (t-1)^2/2$, which results in the LSIF objective. Because $\partial f(t) = t-1$, we have

$$\begin{aligned}
&t - 1 = C\big((t-1)t - (t-1)^2/2\big) + \tilde{f}(t) \\
&\Leftrightarrow \tilde{f}(t) = -C\big((t-1)t - (t-1)^2/2\big) + t - 1 = -\frac{C}{2}t^2 + \frac{C}{2} + t - 1.
\end{aligned}$$

The function is a concave quadratic function, therefore it is upper bounded.

Second, we consider $f(t) = t\log(t) - t$, which results in the UKL or KLIEP objective. Because $\partial f(t) = \log(t)$, we have

$$\log(t) = C\big(\log(t)t - t\log(t) + t\big) + \tilde{f}(t)$$
$$\Leftrightarrow \tilde{f}(t) = -tC + \log(t).$$

We can easily confirm that the function is upper bounded by taking the derivative and finding that $t = 1/C$ gives the maximum.

Third, we consider $f(t) = t\log(t) - (1+t)\log(1+t)$, which is used for DRE based on LR or BKL. Because $\partial f(t) = \log(t) - \log(1+t)$, we have

$$\log(t) - \log(1+t) = C\big((\log(t) - \log(1+t))t - t\log(t) + (1+t)\log(1+t)\big) + \tilde{f}(t)$$
$$\Leftrightarrow \tilde{f}(t) = -C\big(\log(1+t)\big) + \log(t) - \log(1+t) = \log\left(\frac{C}{1+t}\right) + \log\left(\frac{t}{1+t}\right).$$

We can easily confirm that the function is upper bounded as the terms involving t always add up to be negative.

Fourth, we consider DRE based on PULog. By setting $f(t) = \log(1-t) + Ct\left(\log(t) - \log(1-t)\right)$, we can obtain the same risk functional introduced in Kiryo et al. (2017).

C. Train-loss hacking problem in PU classification

Here, we introduce the train-loss hacking discussed in the PU learning literature (Kiryo et al., 2017). In a standard binary classification problem, we train a classifier ψ by minimizing the following empirical risk:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[y_i = +1]\ell(\psi(X_i)) + \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[y_i = -1]\ell(-\psi(X_i)), \tag{8}$$

where $y_i \in \{\pm 1\}$ is a binary label, X_i is a feature, and ℓ is a loss function. On the other hand, in PU learning formulated by du Plessis et al. (2015), because we only have positive data $\{(y'_i = +1, X'_i)\}_{i=1}^{n'}$ and unlabeled data $\{(\boldsymbol{x}''_j)\}_{j=1}^{n''}$, we minimize the following alternative empirical risk:

$$\frac{\pi}{n'}\sum_{i=1}^{n'} \ell(\psi(X'_i)) \underbrace{- \frac{\pi}{n'}\sum_{i=1}^{n'} \ell(-\psi(X'_i))}_{\text{Cause of train-loss hacking.}} + \frac{1}{n''}\sum_{j=1}^{n''} \ell(-\psi(X''_j)), \tag{9}$$

where π is a hyperparameter representing $p(y = +1)$. Note that the empirical risk (9) is unbiased to the population binary classification risk (8) (du Plessis et al., 2015). While the the empirical risk (8) of the standard binary classification is lower bounded under an appropriate choice of ℓ, the empirical risk (9) of PU learning proposed by du Plessis et al. (2015) is not lower bounded owing to the existence of the second term. Therefore, if a model is sufficiently flexible, we can significantly minimize the empirical risk only by minimizing the second term $-\frac{\pi}{n'}\sum_{i=1}^{n'} \ell(-\psi(X'_i))$ without increasing the other terms. Kiryo et al. (2017) proposed non-negative risk correction for avoiding this problem when using neural networks.

D. Network structure used in Sections 5.1 and 6

We explain the structures of neural networks used in the experiments.

D.1. Network structure used in Sections 5.1

In Section 5.1, we used `CIFAR-10` datasets. The model was a convolutional net (Springenberg et al., 2015): $(32 \times 32 \times 3)$-$C(3 \times 6, 3)$-$C(3 \times 16, 3)$-128-84-1, where the input is a 32×32 RGB image, $C(3 \times 6, 3)$ indicates that 3 channels of 3×6 convolutions followed by ReLU is used. This structure has been adopted from the tutorial of Paszke et al. (2019).

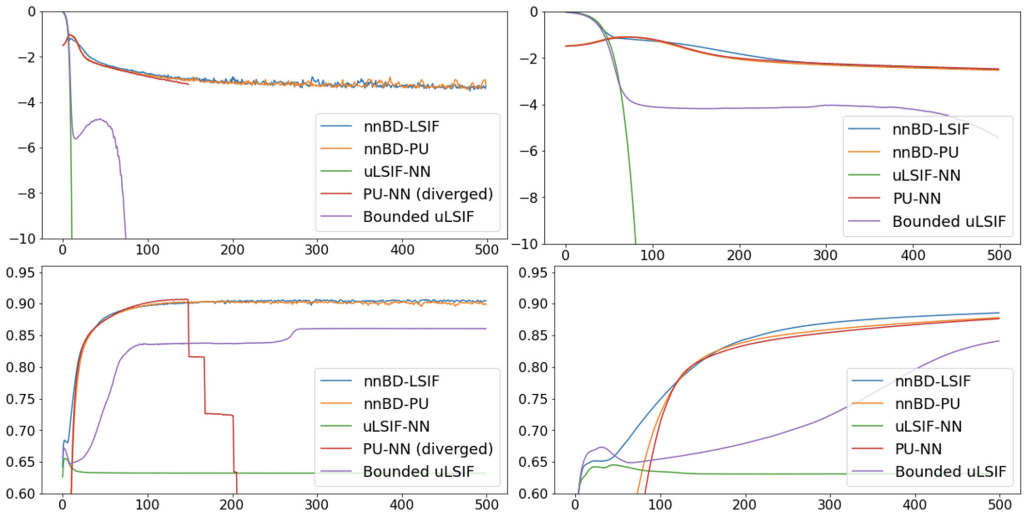

Figure 3. The learning curves of the experiments in Section 5.1. The horizontal axis is epoch. The vertical axes of the top figures indicate the training losses. The vertical axes of the bottom figures show the AUROC for the test data. The bottom figures are identical to the ones displayed in Section 5.1.

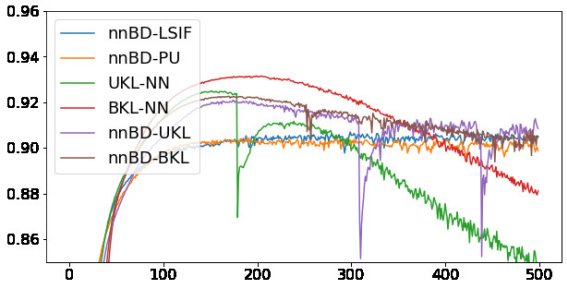

Figure 4. The results of Section F.1.1. The horizontal axis is epoch, and the vertical axis is AUROC.

D.2. Network structure used in Sections 6

Inlier-based Outlier Detection. We used the same LeNet-type CNNs proposed in Ruff et al. (2020). In the CNNs, each convolutional module consists of a convolutional layer followed by leaky ReLU activations with leakiness $\alpha = 0.1$ and (2×2)-max-pooling. For MNIST, we employ a CNN with two modules: $(32 \times 32 \times 3)$-$C(3 \times 32, 5)$-$C(32 \times 64, 5)$-$C(64 \times 128, 5)$-1. For `CIFAR-10` we employ the following architecture: $(32 \times 32 \times 1)$-$C(1 \times 8, 5)$-$C(8 \times 4, 5)$-1 with a batch normalization (Ioffe & Szegedy, 2015) after each convolutional layer.

The WRN architecture was proposed in Zagoruyko & Komodakis (2016) and it is also used in Golan & El-Yaniv (2018). This structure improved the performance of image recognition by decreasing the depth and increasing the width of the residual networks (He et al., 2015). We omit the detailed description of the structure here.

Covariate Shift Adaptation. We used the 5-layer perceptron with ReLU activations. The structure is 10000-1000-1000-1000-1000-1.

E. Existing methods for anomaly detection

This section introduces the existing methods for anomaly detection. DeepSAD is a method for semi-supervised anomaly detection, which tries to take advantage of labeled anomalies (Ruff et al., 2020). GT proposed by Golan & El-Yaniv (2018) trains neural networks based on a self-labeled dataset by performing 72 geometric transformations. The anomaly score based on GT is calculated based on the Dirichlet distribution obtained by maximum likelihood estimation using the softmax

output from the trained network.

In the problem setting of the DeepSAD, we have access to a small pool of labeled samples, e.g. a subset verified by some domain expert as being normal or anomalous. In the experimental results shown in Ruff et al. (2020) indicate that, when we can use such samples, the DeepSAD outperforms the other methods. However, in our experimental results, such samples are not assumed to be available, hence the method does not perform well. The problem setting of Ruff et al. (2020) and ours are both termed *semi-supervised learning* in anomaly detection, but the two settings are different.

F. Details of experiments

The details of experiments are shown in this section. The description of the data is as follows:

`MNIST`**:** The `MNIST` database is one of the most popular benchmark datasets for image classification, which consists of 28×28 pixel handwritten digits from 0 to 9 with $60,000$ train samples and $10,000$ test samples (LeCun et al., 1998). See `http://yann.lecun.com/exdb/mnist/`.

`CIFAR-10`**:** The `CIFAR-10` dataset consists of $60,000$ color images of size 32×32 from 10 classes, each having 6000. There are $50,000$ training images and $10,000$ test images (Krizhevsky et al., 2012). See `https://www.cs.toronto.edu/~kriz/cifar.html`.

`fashion-MNIST`**:** The `fashion-MNIST` dataset consists of $70,000$ grayscale images of size 28×28 from 10 classes. There are $60,000$ training images and $10,000$ test images (Xiao et al., 2017). See `https://github.com/zalandoresearch/fashion-mnist`.

`Amazon Review Dataset`**:** Blitzer et al. (2007) published the text data of Amazon review. The data originally consists of a rating (0-5 stars) for four different genres of products in the electronic commerce site Amazon.com: books, DVDs, electronics, and kitchen appliances. Blitzer et al. (2007) also released the pre-processed and balanced data of the original data. The pre-processed data consists of text data with four labels 1, 2, 4, and 5. We map the text data into $10,000$ dimensional data by the TF-IDF mapping with that vocabulary size. In the experiment, for the pre-processed data, we solve the regression problem where the text data are the inputs and the ratings 1, 2, 4, and 5 are the outputs. When evaluating the performance, following Menon & Ong (2016), we calculate PD (=1-AUROUC) by regarding 4 and 5 ratings as positive labels and 1 and 2 ratings as negative labels.

F.1. Experiments with image data

We show the additional results of Section 5.1. In Figure 3, we show the training loss of LSIF-based methods to demonstrate the train-loss hacking phenomenon caused by the objective function without a lower bound. In Figure 3, even though the training loss of uLSIF-NN and that of Bounded uLSIF decrease more rapidly than that of nnBD-LSIF, the test AUROC score (the higher the better) either drops or fails to increase. These graphs are the manifestations of the severe train-loss hacking in DRE without our proposed device.

F.1.1. Comparison with various estimators using nnBD divergence

Let UKL-NN and BKL-NN be DRE method with the UKL and BKL losses with neural networks without non-negative correction. Finally, we examine the performances of nnBD-LSIF, nnBD-PU, UKL-NN, BKL-NN, nnBD-UKL, and nnBD-BKL. The learning rate was 1×10^{-4}, and the other settings were identical to those in the previous experiments. These results are shown in Figure 4. UKL-NN and BKL-NN also suffer from train-loss hacking although BKL loss seems to be more robust against the train-loss hacking than the other loss functions. Although nnBD-UKL and nnBD-BKL show better performance in earlier epochs, nnBD-LSIF and nnBD-PU appear to be more stable.

F.1.2. Results without gradient ascent

We also show the experimental results without the gradient ascent heuristic. Figure 5 corresponds to the Figure 2 without the gradient ascent heuristic. Figure 6 corresponds to the Figure 3 without the gradient ascent heuristic. Figure 7 corresponds to the Figure 4 without the gradient ascent heuristic. As shown in these experiments, although the gradient ascent/descent heuristic improve the performance, there is no significant difference between empirical performance with and without the heuristic. Therefore, we recommend practitioners to use the gradient ascent/descent heuristic, but if the reader concerns the

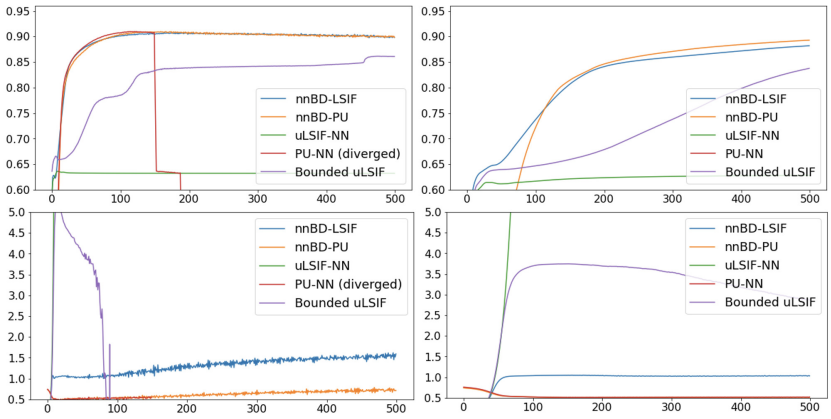

Figure 5. Experimental results of Section 5.1 without gradient ascent/descent heuristic. The horizontal axis is epoch, and the vertical axis is AUROC. The learning rates of the left and right graphs are 1×10^{-4} and 1×10^{-5}, respectively. The upper graphs show the AUROCs and the lower graphs show $\hat{\mathbb{E}}_{\mathrm{de}}[\hat{r}(X)]$, which will approach 1 when we successfully estimate the density ratio.

theoretical guarantee, they can use the plain gradient descent algorithm; that is, naively minimize the proposed empirical nnBD risk.

F.2. Experiments of inlier-based outlier detection

In Table 4, we show the full results of inlier-based outlier detection. In almost all the cases, D3RE for inlier-based outlier detection outperforms the other methods. As explained in Section E, we consider that DeepSAD does not work well because the method assumes the availability of the labeled anomaly data, which is not available in our problem setting.

In Table 5, for different $1/C$ chosen from $\{1, 3, 5, 10\}$, we report the AUROCs of nnBD-LSIF with and without gradient ascent. As shown in the results, loose specification does not significantly decrease the performances. The gradient ascent technique improves the performances, but plain gradient descent still performs well.

Remark 2 (Benchmark Methods)**.** Although GT is outperformed by our proposed method, the problem setting for the comparison is not in favor of GT as it does not assume the access to the test data. Recently proposed methods for semi-supervised anomaly detection by Ruff et al. (2020) did not perform well without using other side information used in Ruff et al. (2020). On the other hand, there is no other competitive methods in this problem setting, to the best of our knowledge.

F.3. Experiments of covariate shift adaptation

In Table 6, we show the detailed results of experiments of covariate shift adaptation. Even when the training data and the test data follow the same distribution, the covariate shift adaptation based on D3RE improves the mean PD. We consider that this is because the importance weighting emphasizes the loss in the empirical higher-density regions of the test examples.

G. Other applications

In this section, we explain other potential applications of the proposed method.

G.1. Covariate shift adaptation by importance weighting

We consider training a model using input distribution different from the test input distribution, which is called *covariate shift*, (Bickel et al., 2009). To solve this problem, the density ratio has been used via importance weighting (IW) (Shimodaira, 2000; Yamada et al., 2010; Reddi et al., 2015).

We use a document dataset of Amazon[4] (Blitzer et al., 2007) for multi-domain sentiment analysis (Blitzer et al., 2007). This data consists of text reviews from four different product domains: book, electronics (elec), dvd, and kitchen. Following Chen et al. (2012) and Menon & Ong (2016), we transform the text data using TF-IDF to map them into the instance

[4] http://john.blitzer.com/software.html

Table 4. Average area under the ROC curve (Mean) of anomaly detection methods averaged over 5 trials with the standard deviation (SD). For all datasets, each model was trained on the single class, and tested against all other classes. The best performing method in each experiment is in bold. *SD*: Standard deviation.

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
5	0.997	0.000	0.996	0.001	0.998	0.000	**1.000**	0.000	**1.000**	0.000	0.502	0.046	0.990	0.001
6	0.997	0.001	0.997	0.001	0.999	0.000	**1.000**	0.000	**1.000**	0.000	0.671	0.027	0.998	0.000
7	0.996	0.001	0.993	0.001	0.998	0.001	**1.000**	0.000	**1.000**	0.001	0.685	0.032	0.927	0.004
8	0.997	0.000	0.994	0.001	0.997	0.000	**0.999**	0.000	**0.999**	0.000	0.654	0.026	0.949	0.002
9	0.993	0.002	0.990	0.002	0.994	0.001	**0.998**	0.001	**0.998**	0.001	0.786	0.021	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
dog	0.728	0.103	0.875	0.004	**0.906**	0.002	0.887	0.005	0.896	0.002	0.490	0.025	0.873	0.002
frog	0.750	0.060	0.944	0.003	**0.958**	0.001	0.948	0.004	0.948	0.001	0.744	0.014	0.879	0.008
horse	0.782	0.048	0.928	0.003	0.948	0.002	0.921	0.007	0.927	0.002	0.519	0.015	**0.953**	0.001
ship	0.780	0.048	0.958	0.003	**0.965**	0.001	0.964	0.002	0.957	0.001	0.430	0.062	0.921	0.009
truck	0.708	0.081	0.939	0.003	**0.955**	0.001	0.952	0.003	0.949	0.001	0.393	0.008	0.911	0.003

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
Sandal	0.968	0.011	0.997	0.001	**0.999**	0.000	**0.999**	0.000	**0.999**	0.000	0.681	0.023	0.949	0.005
Shirt	0.919	0.005	0.952	0.001	**0.958**	0.001	0.944	0.005	0.932	0.001	0.618	0.015	0.842	0.004
Sneaker	0.991	0.001	0.994	0.002	**0.998**	0.000	**0.998**	0.000	**0.998**	0.000	0.802	0.054	0.954	0.006
Bag	0.980	0.005	0.994	0.001	**0.999**	0.000	0.998	0.000	**0.999**	0.000	0.447	0.034	0.973	0.006
Ankle boot	0.992	0.001	0.985	0.015	**0.999**	0.000	0.997	0.000	0.996	0.000	0.583	0.023	0.996	0.000

Table 5. We show average area under the ROC curve (Mean) of anomaly detection methods averaged over 5 trials with the standard deviation (SD) for nnBD-LSIF with LeNet. We choose $1/C$, which represents a guessed upper bound, from $\{1, 3, 5, 10\}$. Each model is trained on the single class, and tested against all other classes. We show both results with and without gradient ascent and ∘ denotes the use of the gradient ascent technique. The best performing method for each inlier class is highlighted in bold. The best performing method for each $1/C$ is highlighted in underline.

CIFAR-10	nnBD-LSIF															
Network	LeNet															
$1/C$ (Guessed upper bound)	1				3				5				10			
With gradient ascent	∘				∘				∘				∘			
Inlier Class	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
plane	0.491	0.009	0.642	0.019	**0.934**	0.002	0.918	0.003	0.920	0.003	0.899	0.002	0.886	0.007	0.839	0.009
car	0.521	0.032	0.644	0.011	**0.957**	0.002	0.950	0.002	0.951	0.003	0.939	0.004	0.920	0.006	0.894	0.013
bird	0.501	0.013	0.622	0.012	**0.850**	0.007	0.832	0.004	0.835	0.005	0.812	0.006	0.818	0.004	0.765	0.010
cat	0.491	0.015	0.616	0.014	**0.820**	0.003	0.807	0.003	0.802	0.007	0.770	0.005	0.773	0.011	0.721	0.006
deer	0.523	0.017	0.658	0.022	**0.886**	0.004	0.879	0.001	0.873	0.005	0.862	0.004	0.852	0.007	0.820	0.009
dog	0.514	0.018	0.621	0.011	**0.875**	0.004	0.855	0.005	0.852	0.008	0.820	0.007	0.821	0.009	0.758	0.017
frog	0.496	0.018	0.671	0.018	**0.944**	0.003	0.932	0.003	0.927	0.003	0.917	0.005	0.886	0.004	0.845	0.014
horse	0.506	0.017	0.631	0.018	**0.928**	0.003	0.910	0.003	0.916	0.005	0.885	0.003	0.880	0.007	0.823	0.020
ship	0.494	0.027	0.680	0.026	**0.958**	0.003	0.949	0.001	0.956	0.002	0.942	0.002	0.933	0.004	0.907	0.006
truck	0.506	0.013	0.660	0.016	**0.939**	0.003	0.930	0.003	0.922	0.003	0.907	0.007	0.885	0.007	0.843	0.018

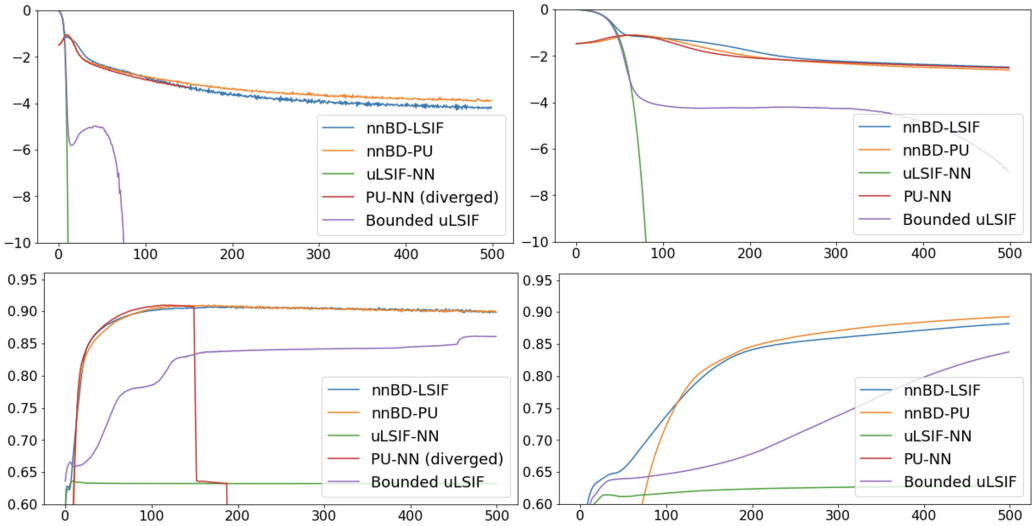

Figure 6. The learning curves of the experiments in Section 5.1 without gradient ascent/descent heuristic. The horizontal axis is epoch. The vertical axes of the top figures indicate the training losses. The vertical axes of the bottom figures show the AUROC for the test data. The bottom figures are identical to the ones displayed in Section 5.1.

space $\mathcal{X} = \mathbb{R}^{10000}$ (Salton & McGill, 1986). Each review is endowed with four labels indicating the positivity of the review, and our goal is to conduct regression for these labels. To achieve this goal, we perform kernel ridge regression with the polynomial kernel. We compare regression without IW (w/o IW) with regression using the density ratio estimated by PU-NN, uLSIF-NN, nnBD-LSIF, nnBD-PU, uLSIF with Gaussian kernels (Kernel uLSIF), and KLIEP with Gaussian kernels (Kernel KLIEP). We conduct experiments on $2,000$ samples from one domain, and test $2,000$ samples. Following Menon & Ong (2016), we reduce the dimension into 100 dimensions by principal component analysis when using Kernel uLSIF, Kernel KLEIP, and regressions. Following Menon & Ong (2016) and Cortes & Mohri (2011), the mean and standard deviation of the pairwise disagreement (PD), $1 - \mathrm{AUROC}$, is reported. A part of results is in Table 7. The full results are in Appendix F.3. The methods with D3RE show preferable performance, but the improvement is not significant compared with the image data. We consider this is owing to the difficulty of the covariate shift problem in this dataset.

f-divergence estimation. f-divergences (Ali & Silvey, 1966; Csiszár, 1967) are the discrepancy measures of probability densities based on the density ratio, hence the proposed method can be used for their estimation. They include the KL divergence (Kullback & Leibler, 1951), the Hellinger distance (Hellinger, 1909), and the Pearson divergence (Pearson, 1900), as examples.

Two-sample homogeneity test. The purpose of a homogeneity test is to determine if two or more datasets come from the same distribution (Loevinger, 1948). For two-sample testing, using a semiparametric f-divergence estimator with nonparametric density ratio models has been studied (Keziou., 2003; Keziou & Leoni-Aubin, 2005). Kanamori et al. (2010) and Sugiyama et al. (2011a) employed direct DRE for the nonparametric DRE.

Generative adversarial networks. Generative adversarial networks (GANs) are successful deep generative models, which learns to generate new data with the same distribution as the training data Goodfellow et al. (2014). Various GAN methods have been proposed, amongst which Nowozin et al. (2016) proposed *f-GAN*, which minimizes the variational estimate of f-divergence. Uehara et al. (2016) extended the idea of Nowozin et al. (2016) to use BD minimization for DRE. The estimator proposed in this paper also has a potential to improve the method of Uehara et al. (2016).

Average treatment effect estimation and off-policy evaluation. One of the goals in causal inference is to estimate the expected treatment effect, which is a *counterfactual* value. Therefore, following the causality formulated by Rubin (1974), we consider estimating the average treatment effect (ATE). Recently, from machine learning community, off-policy evaluation (OPE) is also proposed, which is a generalization of ATE (Dudík et al., 2011; Imai & Ratkovic, 2014; Wang et al., 2017; Narita et al., 2019; Bibaut et al., 2019; Kallus & Uehara, 2019; Oberst & Sontag, 2019). OPE has garnered attention

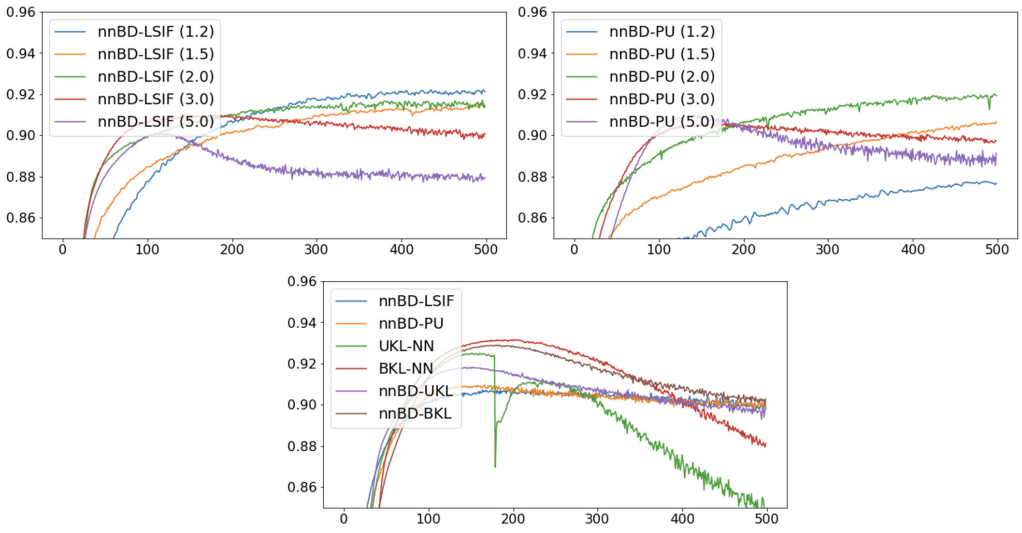

Figure 7. The detailed experimental results for Section F.1.1. The horizontal axis is epoch, and the vertical axis is AUROC.

in applications such as advertisement design selection, personalized medicine, search engines, and recommendation systems (Beygelzimer & Langford, 2009; Li et al., 2010; Athey & Wager, 2017).

The problem in ATE estimation and OPE is sample selection bias. For removing the bias, the density ratio has a critical role. An idea of using the density ratio dates back to (Rosenbaum, 1987), which proposed an inverse probability weighting (IPW) method (Horvitz & Thompson, 1952) for ATE estimation. In the IPW method, we approximate the parameter of interest with the sample average with inverse assignment probability of treatment (action), which is also called propensity score. Here, it is known that using the true assignment probability yields higher variance than the case where we use an estimated assignment probability even if we know the true value (Hirano et al., 2003; Henmi & Eguchi, 2004; Henmi et al., 2007). This property can be explained from the viewpoint of semiparametric efficiency (Bickel et al., 1998). While the asymptotic variance of the IPW estimator with an estimated propensity score can achieve the efficiency bound, that of the IPW estimator with the true propensity score does not.

By extending the IPW estimator, more robust ATE estimators are proposed by Rosenbaum (1983), which is known as a doubly robust (DR) estimator. The doubly robust estimator is not only robust to model misspecification but also useful in showing asymptotic normality. In particular, when using the density ratio and the other nuisance parameters estimated from the machine learning method, the conventional IPW and DR estimators do not have asymptotic normality (Chernozhukov et al., 2018). This is because the nuisance estimators do not satisfy Donsker's condition, which is required for showing the asymptotic normality of semiparametric models. However, by using the sample splitting method proposed by Klaassen (1987), Zheng & van der Laan (2011), and Chernozhukov et al. (2018), we can show the asymptotic normality when using the DR estimator. Note that for the IPW estimator, we cannot show the asymptotic normality even if using sample-splitting.

When using the IPW and DR estimator, we often consider a two-stage approach: in the first stage, we estimate the nuisance parameters, including the density ratio; in the second stage, we construct a semiparametric ATE estimator including the first-stage nuisance estimators. This is also called two-step generalized method of moments (GMM). On the other hand, from the causal inference community, there are also weighting-based covariate balancing methods (Qin & Zhang, 2007; Tan, 2010; Hainmueller, 2012; Imai & Ratkovic, 2014). In particular, Imai & Ratkovic (2014) proposed a covariate balancing propensity score (CBPS), which simultaneously estimates the density ratio and ATE. The idea of CBPS is to construct moment conditions, including the density ratios, and estimate the ATE and density ratio via GMM simultaneously. Although the asymptotic property of the CBPS is the same as other conventional estimators, existing empirical studies report that the CBPS outperforms them (Wyss et al., 2014).

Readers may feel that the CBPS (Imai & Ratkovic, 2014) has a close relationship with the direct DRE, but we consider that it is less relevant to the context of the direct DRE. From the DRE perspective, the method of Imai & Ratkovic (2014) boils down to the method of Gretton et al. (2009), which proposed direct DRE through moment matching. The research motivation of

Table 6. Average PD (Mean) with standard deviation (SD) over 10 trials with different seeds per method. The best performing method in terms of the mean PD is specified by bold face.

Domains (Train → Test)	books → books		dvd → books		dvd → dvd		elec → books		elec → dvd	
DRE method	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
w/o IW	0.093	0.003	0.128	0.008	0.100	0.005	0.212	0.012	0.187	0.008
Kernel uLSIF	0.089	0.002	0.114	0.006	0.094	0.004	0.200	0.009	0.179	0.006
Kernel KLIEP	0.089	0.002	0.116	0.006	0.094	0.004	0.205	0.011	0.184	0.008
uLSIF-NN	0.093	0.003	0.128	0.008	0.100	0.005	0.212	0.012	0.187	0.008
PU-NN	0.093	0.003	0.128	0.008	0.100	0.005	0.212	0.012	0.187	0.008
nnBD-LSIF	**0.086**	0.002	**0.113**	0.005	**0.091**	0.004	**0.199**	0.009	**0.176**	0.005
nnBD-PU	0.090	0.003	**0.113**	0.006	0.096	0.004	**0.199**	0.009	**0.176**	0.006

Domains (Train → Test)	elec → elec		kitchen → books		kitchen → dvd		kitchen → elec		kitchen → kitchen	
DRE method	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
w/o IW	0.079	0.005	0.202	0.013	0.185	0.006	0.073	0.004	0.062	0.002
Kernel uLSIF	0.072	0.003	0.192	0.007	0.178	0.008	0.071	0.003	0.060	0.003
Kernel KLIEP	0.072	0.003	0.195	0.005	0.182	0.007	0.072	0.004	0.060	0.002
uLSIF-NN	0.079	0.005	0.202	0.013	0.185	0.006	0.073	0.004	0.062	0.002
PU-NN	0.079	0.005	0.202	0.013	0.185	0.006	0.073	0.004	0.062	0.002
nnBD-LSIF	**0.071**	0.003	**0.189**	0.008	**0.174**	0.008	**0.068**	0.003	**0.058**	0.003
nnBD-PU	0.074	0.004	0.190	0.008	**0.174**	0.008	**0.068**	0.003	0.062	0.005

Table 7. Average PD (Mean) with standard deviation (SD) over 10 trials with different seeds per method. The best performing method in terms of the mean PD is specified by bold face.

Domains (Train → Test)	book → dvd		book → elec		book → kitchen		dvd → elec		dvd → kitchen		elec → kitchen	
DRE method	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
w/o IW	0.126	0.008	0.174	0.010	0.166	0.009	0.162	0.006	0.146	0.010	0.074	0.005
Kernel uLSIF	0.122	0.009	0.162	0.009	0.159	0.007	0.153	0.006	0.142	0.007	0.068	0.005
Kernel KLIEP	0.130	0.010	0.164	0.009	0.161	0.007	0.154	0.006	0.143	0.006	0.070	0.005
uLSIF-NN	0.126	0.008	0.174	0.010	0.166	0.009	0.162	0.006	0.146	0.010	0.074	0.005
PU-NN	0.126	0.008	0.174	0.010	0.166	0.009	0.162	0.006	0.146	0.010	0.074	0.005
nnBD-LSIF	0.120	0.008	**0.160**	0.008	0.157	0.008	**0.148**	0.006	**0.138**	0.007	**0.066**	0.005
nnBD-PU	**0.119**	0.008	**0.160**	0.008	**0.156**	0.007	**0.148**	0.005	**0.138**	0.007	**0.066**	0.005

Imai & Ratkovic (2014) is to estimate the ATE with estimating a nuisance density ratio estimator simultaneously. Therefore, the density ratio itself is *nuisance* parameter; that is, they are not interested in the estimation performance of the density ratio. Under their motivation, they are interested in a density ratio estimator satisfying the moment condition for estimating the ATE, not in a density ratio estimator predicting the true density ratio well. In addition, while the direct DRE method adopts linear-in-parameter models and neural networks (our work), it is not appropriate to use those methods with the CBPS (Chernozhukov et al., 2018). This is because the density ratio estimator does not satisfy Donsker's condition. Even naive Ridge and Lasso regression estimators do not satisfy the Donsker's condition. Therefore, when using machine learning methods for estimating the density ratio, we cannot show asymptotic normality of an ATE estimator obtained by the CBPS; therefore, we need to use the sample-splitting method by (Chernozhukov et al., 2018). This means that when using the CBPS, we can only use a naive parametric linear model without regularization or classic nonparametric kernel regression. Recently, for GMM with such non-Donsker nuisance estimators, Chernozhukov et al. (2016) also proposed a new GMM method based on the conventional two-step approach. For these reasons, the CBPS is less relevant to the direct DRE context.

Off-policy evaluation with external validity. By the problem setting of combining causal inference and domain adaptation, Uehara et al. (2020) recently proposed using covariate shift adaptation to solve the *external validity* problem in OPE, i.e., the case that the distribution of covariates is the same between the historical and evaluation data (Cole & Stuart, 2010; Pearl & Bareinboim, 2014).

Change point detection. The methods for *change-point detection* try to detect abrupt changes in time-series data (Basseville & Nikiforov, 1993; Brodsky & Darkhovsky, 1993; Gustafsson, 2000; Nguyen et al., 2011). There are two types of problem settings in change-point detection, namely the real-time detection (Adams, 2007; Garnett et al., 2009; Paquet, 2007) and the retrospective detection (Basseville & Nikiforov, 1993; Yamanishi & Takeuchi, 2002). In retrospective detection, which requires longer reaction periods, Liu et al. (2012) proposed using techniques of direct DRE. Whereas the existing methods rely on linear-in-parameter models, our proposed method enables us to employ more complex models for change point detection.

Similarity-based sentiment analysis. Kato (2019) used the density ratio estimated from PU learning for sentiment analysis of text data based on similarity.

H. Generalization error bound

The generalization error bound can be proved by building upon the proof techniques in Kiryo et al. (2017); Lu et al. (2020).

Notations for the theoretical analysis. We denote the set of real values by $\mathbb{R}$ and that of positive integers by $\mathbb{N}$. Let $\mathcal{X} \subset \mathbb{R}^d$. Let $p_{\mathrm{nu}}(x)$ and $p_{\mathrm{de}}(x)$ be probability density functions over $\mathcal{X}$, and assume that the density ratio $r^*(x) := \frac{p_{\mathrm{nu}}(x)}{p_{\mathrm{de}}(x)}$ is existent and bounded: $\overline{R} := \|r^*\|_\infty < \infty$. Assume $0 < C < \frac{1}{\overline{R}}$. Since $\overline{R} \geq 1$ (because $1 = \int p_{\mathrm{de}}(x) r^*(x) dx \leq 1 \cdot \|r^*\|_\infty$), we have $C \in (0, 1]$ and hence $p_{\mathrm{mod}} := p_{\mathrm{de}} - C p_{\mathrm{nu}} > 0$.

Problem Setup. Let the hypothesis class of density ratio be $\mathcal{H} \subset \{r : \mathbb{R}^D \to (b_r, B_r) =: I_r\}$, where $0 \leq b_r < \overline{R} < B_r$. Let $f : I_r \to \mathbb{R}$ be a twice continuously-differentiable convex function with a bounded derivative. Define $\tilde{f}$ by $\partial f(t) = C(\partial f(t) t - f(t)) + \tilde{f}(t)$, where ∂f is the derivative of f continuously extended to 0 and B_r. Recall the definitions $\ell_1(t) := \partial f(t) t - f(t) + A$, $\ell_2(t) := -\tilde{f}(t)$, and

$$
\begin{aligned}
\mathrm{BD}_f(r) &:= \mathbb{E}_{\mathrm{de}} \left[\partial f(r(X)) r(X) - f(r(X)) + A\right] - \mathbb{E}_{\mathrm{nu}} \left[\partial f(r(X))\right] \\
&= \mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}} \left[\partial f(r(X)) r(X) - f(r(X)) + A\right] - \mathbb{E}_{\mathrm{nu}} \left[\tilde{f}(r(X))\right] \\
&= \mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}} \ell_1(r(X)) + \mathbb{E}_{\mathrm{nu}} \ell_2(r(X)) \\
&\left(= (\mathbb{E}_{\mathrm{de}} - C\mathbb{E}_{\mathrm{nu}}) \ell_1(r(X)) + \mathbb{E}_{\mathrm{nu}} \ell_2(r(X))\right), \\
\widehat{\mathrm{nnBD}}_f(r) &:= \rho\left(\hat{\mathbb{E}}_{\mathrm{mod}} \ell_1(r(X))\right) + \hat{\mathbb{E}}_{\mathrm{nu}} \ell_2(r(X)) \\
&\left(= \rho((\hat{\mathbb{E}}_{\mathrm{de}} - C\hat{\mathbb{E}}_{\mathrm{nu}}) \ell_1(r(X)) + \hat{\mathbb{E}}_{\mathrm{nu}} \ell_2(r(X))\right),
\end{aligned}
$$

where we denoted $\hat{\mathbb{E}}_{\mathrm{mod}} = \hat{\mathbb{E}}_{\mathrm{de}} - C\hat{\mathbb{E}}_{\mathrm{nu}}$ and ρ is a consistent correction function with Lipschitz constant L_ρ (Definition 1).

Remark 3. The true density ratio r^* minimizes BD_f.

Definition 1 (Consistent correction function (Lu et al., 2020))**.** A function $f : \mathbb{R} \to \mathbb{R}$ is called a consistent correction function if it is Lipschitz continuous, non-negative and $f(x) = x$ for all $x \geq 0$.

Definition 2 (Rademacher complexity)**.** Given $n \in \mathbb{N}$ and a distribution p, define the Rademacher complexity $\mathcal{R}_n^p(\mathcal{H})$ 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 Rademacher variables (i.e., independent variables following the uniform distribution over $\{-1, +1\}$) and $\{X_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p$.

The theorem in the paper is a special case of Theorem 3 with $\rho(\cdot) := \max\{0, \cdot\}$ (in which case $L_\rho = 1$) and Theorem 4.

Theorem 3 (Generalization error bound)**.** *Assume that $B_\ell := \sup_{t \in I_r} \{\max\{|\ell_1(t)|, |\ell_2(t)|\}\} < \infty$. Assume ℓ_1 is L_{ℓ_1}-Lipschitz and ℓ_2 is L_{ℓ_2}-Lipschitz. Assume that 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 $\bar{r} \in \arg\min_{r \in \mathcal{H}} \mathrm{BD}_f(r)$. Also assume $\inf_{r \in \mathcal{H}} \mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}} \ell_1(r(X)) > 0$ and that $(\rho - \mathrm{Id})$ is $(L_{\rho - \mathrm{Id}})$-Lipschitz. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$
\begin{aligned}
\mathrm{BD}_f(\hat{r}) - \mathrm{BD}_f(\bar{r}) \leq\ & 8 L_\rho L_{\ell_1} \mathcal{R}_{n_{\mathrm{de}}}^{p_{\mathrm{de}}}(\mathcal{H}) + 8(L_\rho C L_{\ell_1} + L_{\ell_2}) \mathcal{R}_{n_{\mathrm{nu}}}^{p_{\mathrm{nu}}}(\mathcal{H}) \\
& + 2\Phi_{(C,f,\rho)}(n_{\mathrm{nu}}, n_{\mathrm{de}}) + B_\ell \sqrt{8 \left(\frac{L_\rho^2}{n_{\mathrm{de}}} + \frac{(1 + L_\rho C)^2}{n_{\mathrm{nu}}} \right) \log \frac{1}{\delta}},
\end{aligned}
$$

where $\Phi_{(C,f,\rho)}(n_{\mathrm{nu}}, n_{\mathrm{de}})$ is defined as in Lemma 2.

Proof. Since $\hat{r}$ minimizes $\widehat{\mathrm{nnBD}}_f$, we have

$$\begin{aligned}
\mathrm{BD}_f(\hat{r}) - \mathrm{BD}_f(\bar{r}) &= \mathrm{BD}_f(\hat{r}) - \widehat{\mathrm{nnBD}}_f(\hat{r}) + \widehat{\mathrm{nnBD}}_f(\hat{r}) - \mathrm{BD}_f(\bar{r}) \\
&\leq \mathrm{BD}_f(\hat{r}) - \widehat{\mathrm{nnBD}}_f(\hat{r}) + \widehat{\mathrm{nnBD}}_f(\bar{r}) - \mathrm{BD}_f(\bar{r}) \\
&\leq 2 \sup_{r \in \mathcal{H}} |\widehat{\mathrm{nnBD}}_f(r) - \mathrm{BD}_f(r)| \\
&\leq \underbrace{2 \sup_{r \in \mathcal{H}} |\widehat{\mathrm{nnBD}}_f(r) - \mathbb{E}\widehat{\mathrm{nnBD}}_f(r)|}_{\text{Maximal deviation}} + \underbrace{2 \sup_{r \in \mathcal{H}} |\mathbb{E}\widehat{\mathrm{nnBD}}_f(r) - \mathrm{BD}_f(r)|}_{\text{Bias}}.
\end{aligned}$$

We apply McDiarmid's inequality (McDiarmid, 1989; Mohri et al., 2018) to the maximal deviation term. The absolute value of the difference caused by altering one data point in the maximal deviation term is bounded from above by $2B_\ell \frac{L_\rho}{n_{\mathrm{de}}}$ if the altered point is a sample from p_{de} and $2B_\ell \frac{1+L_\rho C}{n_{\mathrm{nu}}}$ if it is from p_{nu}. Therefore, McDiarmid's inequality implies, with probability at least $1-\delta$, that we have

$$\begin{aligned}
&\sup_{r \in \mathcal{H}} |\widehat{\mathrm{nnBD}}_f(r) - \mathbb{E}\widehat{\mathrm{nnBD}}_f(r)| \\
&\quad \leq \underbrace{\mathbb{E}\left[\sup_{r \in \mathcal{H}} |\widehat{\mathrm{nnBD}}_f(r) - \mathbb{E}\widehat{\mathrm{nnBD}}_f(r)|\right]}_{\text{Expected maximal deviation}} + B_\ell \sqrt{2\left(\frac{L_\rho^2}{n_{\mathrm{de}}} + \frac{(1+L_\rho C)^2}{n_{\mathrm{nu}}}\right)\log\frac{1}{\delta}}.
\end{aligned}$$

Applying Lemma 1 to the expected maximal deviation term and Lemma 2 to the bias term, we obtain the assertion. □

The following lemma generalizes the symmetrization lemmas proved in Kiryo et al. (2017) and Lu et al. (2020).

Lemma 1 (Symmetrization under Lipschitz-continuous modification)**.** *Let $0 \leq a < b$, $J \in \mathbb{N}$, and $\{K_j\}_{j=1}^J \subset \mathbb{N}$. Given i.i.d. samples $\mathcal{D}_{(j,k)} := \{X_i\}_{i=1}^{n_{(j,k)}}$ each from a distribution $p_{(j,k)}$ over $\mathcal{X}$, consider a stochastic process $\hat{S}$ indexed by $\mathcal{F} \subset (a,b)^{\mathcal{X}}$ of the form*

$$\hat{S}(f) = \sum_{j=1}^{J} \rho_j \left(\sum_{k=1}^{K_j} \hat{\mathbb{E}}_{(i,j)}[\ell_{(j,k)}(f(X))] \right),$$

where each ρ_j is a L_{ρ_j}-Lipschitz function on $\mathbb{R}$, $\ell_{(j,k)}$ is a $L_{\ell_{(j,k)}}$-Lipschitz function on (a,b), and $\hat{\mathbb{E}}_{(i,j)}$ denotes the expectation with respect to the empirical measure of $\mathcal{D}_{(j,k)}$. Denote $S(f) := \mathbb{E}\hat{S}(f)$ where $\mathbb{E}$ is the expectation with respect to the product measure of $\{\mathcal{D}_{(j,k)}\}_{(j,k)}$. Here, the index j denotes the grouping of terms due to ρ_j, and k denotes each sample average term. Then we have

$$\mathbb{E} \sup_{f \in \mathcal{F}} |\hat{S}(f) - S(f)| \leq 4 \sum_{j=1}^{J} \sum_{k=1}^{K_j} L_{\rho_j} L_{\ell_{(j,k)}} \mathcal{R}_{n_{(j,k)}, p_{(j,k)}}(\mathcal{F}).$$

Proof. First, we consider a continuous extension of $\ell_{(j,k)}$ defined on (a,b) to $[0,b)$. Since the functions in $\mathcal{F}$ take values only in (a,b), this extension can be performed without affecting the values of $\hat{S}(f)$ or $S(f)$. We extend the function by defining the values for $x \in [0,a]$ as $\ell_{(j,k)}(x) := \lim_{x' \downarrow a} \ell_{(j,k)}(x')$, where the right-hand side is guaranteed to exist since $\ell_{(j,k)}$ is Lipschitz continuous hence uniformly continuous. Then, $\ell_{(j,k)}$ remains a L_{ρ_j}-Lipschitz continuous function on $[0,b)$. Now we perform symmetrization (Vapnik, 1998), deal with ρ_j's, and then bound the symmetrized process by Rademacher complexity. Denoting independent copies of $\{X_{(j,k)}\}$ by $\{X_{j,k}^{(\mathrm{gh})}\}_{(j,k)}$ and the corresponding expectations as

well as the sample averages with $^{(\mathrm{gh})}$,

$$
\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}}|\hat{S}(f)-S(f)| \\
&\leq \sum_{j=1}^{J}\mathbb{E}\sup_{f\in\mathcal{F}}|\rho_j(\sum_{k=1}^{K_j}\hat{\mathbb{E}}_{(i,j)}\ell_{(j,k)}(f(X)))-\mathbb{E}^{(\mathrm{gh})}\rho_j(\sum_{k=1}^{K_j}\hat{\mathbb{E}}^{(\mathrm{gh})}_{(j,k)}\ell_{(j,k)}(f(X^{(\mathrm{gh})})))| \\
&\leq \sum_{j=1}^{J}\mathbb{E}\mathbb{E}^{(\mathrm{gh})}\sup_{f\in\mathcal{F}}|\rho_j(\sum_{k=1}^{K_j}\hat{\mathbb{E}}_{(i,j)}\ell_{(j,k)}(f(X)))-\rho_j(\sum_{k=1}^{K_j}\hat{\mathbb{E}}^{(\mathrm{gh})}_{(j,k)}\ell_{(j,k)}(f(X^{(\mathrm{gh})})))| \\
&\leq \sum_{j=1}^{J}L_{\rho_j}\sum_{k=1}^{K_j}\mathbb{E}\mathbb{E}^{(\mathrm{gh})}\sup_{f\in\mathcal{F}}|\hat{\mathbb{E}}_{(i,j)}\ell_{(j,k)}(f(X))-\hat{\mathbb{E}}^{(\mathrm{gh})}_{(j,k)}\ell_{(j,k)}(f(X^{(\mathrm{gh})}))| \\
&= \sum_{j=1}^{J}L_{\rho_j}\sum_{k=1}^{K_j}\mathbb{E}\mathbb{E}^{(\mathrm{gh})}\sup_{f\in\mathcal{F}}|\hat{\mathbb{E}}_{(i,j)}(\ell_{(j,k)}(f(X))-\ell_{(j,k)}(0))-\hat{\mathbb{E}}^{(\mathrm{gh})}_{(j,k)}(\ell_{(j,k)}(f(X^{(\mathrm{gh})}))-\ell_{(j,k)}(0))| \\
&\leq \sum_{j=1}^{J}L_{\rho_j}\sum_{k=1}^{K_j}\left(2\mathcal{R}_{n_{(j,k)},p_{(j,k)}}(\{\ell_{(j,k)}\circ f-\ell_{(j,k)}(0):f\in\mathcal{F}\})\right) \\
&\leq \sum_{j=1}^{J}L_{\rho_j}\sum_{k=1}^{K_j}2\cdot 2L_{\ell_{(j,k)}}\mathcal{R}_{n_{(j,k)},p_{(j,k)}}(\mathcal{F}),
\end{aligned}
$$

where we applied Talagrand's contraction lemma for two-sided Rademacher complexity (Ledoux & Talagrand, 1991; Bartlett & Mendelson, 2001) with respect to $(t\mapsto\ell_{(j,k)}(t)-\ell_{(j,k)}(0))$ in the last inequality. □

Lemma 2 (Bias due to risk correction)**.** *Assume* $\inf_{r\in\mathcal{H}}\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))>0$ *and that* $(\rho-\mathrm{Id})$ *is* $(L_{\rho-\mathrm{Id}})$*-Lipschitz on* $\mathbb{R}$*. There exists* $\alpha>0$ *such that*

$$
\begin{aligned}
\sup_{r\in\mathcal{H}}|\widehat{\mathbb{E}\mathrm{nnBD}}_f(r)-\mathrm{BD}_f(r)| &\leq (1+C)B_\ell L_{\rho-\mathrm{Id}}\exp\left(-\frac{2\alpha^2}{({B_\ell}^2/n_{\mathrm{de}})+(C^2{B_\ell}^2/n_{\mathrm{nu}})}\right) \\
&=: \Phi_{(C,f,\rho)}(n_{\mathrm{nu}},n_{\mathrm{de}}).
\end{aligned}
$$

Remark 4. Note that we already have $p_{\mathrm{mod}}\geq 0$ and $\ell_1\geq 0$ and hence $\inf_{r\in\mathcal{H}}\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))\geq 0$. Therefore, the assumption of Lemma 2 is essentially referring to the strict positivity of the infimum. Here, $\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}$ and $\mathbb{P}(\cdot)$ denote the expectation and the probability with respect to the joint distribution of the samples included in $\hat{\mathbb{E}}_{\mathrm{mod}}$.

Proof. Fix an arbitrary $r\in\mathcal{H}$. We have

$$
\begin{aligned}
&|\widehat{\mathbb{E}\mathrm{nnBD}}_f(r)-\mathrm{BD}_f(r)| = |\mathbb{E}[\widehat{\mathrm{nnBD}}_f(r)-\widehat{\mathrm{BD}}_f(r)]| \\
&= |\mathbb{E}[\rho(\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))]| \leq \mathbb{E}\left[|\rho(\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))|\right] \\
&= \mathbb{E}\left[\mathbb{1}\{\rho(\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))\neq\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))\}\cdot|\rho(\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))|\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\{\rho(\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))\neq\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))\}\right]\left(\sup_{s:|s|\leq(1+C)B_\ell}|\rho(s)-s|\right)
\end{aligned}
$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, and we used $|\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))|\leq(1+C)B_\ell$. Further, we have

$$
\begin{aligned}
\sup_{s:|s|\leq(1+C)B_\ell}|\rho(s)-s| &\leq \sup_{s:|s|\leq(1+C)B_\ell}|(\rho-\mathrm{Id})(s)-(\rho-\mathrm{Id})(0)|+|(\rho-\mathrm{Id})(0)| \\
&\leq \sup_{s:|s|\leq(1+C)B_\ell}L_{\rho-\mathrm{Id}}|s-0|+0\leq(1+C)B_\ell L_{\rho-\mathrm{Id}},
\end{aligned}
$$

where Id denotes the identity function. On the other hand, since $\inf_{r\in\mathcal{H}} \mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) > 0$ is assumed, there exists $\alpha > 0$ such that for any $r \in \mathcal{H}$, $\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) > \alpha$. Therefore, denoting the support of a function by $\mathrm{supp}(\cdot)$,

$$\mathbb{E}\left[\mathbb{1}\{\rho(\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))) \neq \hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))\}\right] = \mathbb{P}\left(\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) \in \mathrm{supp}(\rho - \mathrm{Id})\right)$$
$$\leq \mathbb{P}\left(\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) < 0\right) \leq \mathbb{P}\left(\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) < \mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) - \alpha\right)$$

holds. Now we apply McDiarmid's inequality to the right-most quantity. The absolute difference caused by altering one data point in $\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))$ is bounded by $\frac{B_\ell}{n_{\mathrm{de}}}$ if the change is in a sample from p_{de} and $\frac{CB_\ell}{n_{\mathrm{nu}}}$ otherwise. Therefore, McDiarmid's inequality implies

$$\mathbb{P}\left(\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) < \mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) - \alpha\right) \leq \exp\left(-\frac{2\alpha^2}{({B_\ell}^2/n_{\mathrm{de}}) + (C^2{B_\ell}^2/n_{\mathrm{nu}})}\right).$$

□

Theorem 4 (Generalization error bound)**.** *Under Assumption 3, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have* $\mathrm{BD}_f(\hat{r}) - \mathrm{BD}_f(\bar{r}) \leq L_{\ell_1}\mathcal{R}^{p_{\mathrm{de}}}_{n_{\mathrm{de}}}(\mathcal{H}) + 8(CL_{\ell_1} + L_{\ell_2})\mathcal{R}^{p_{\mathrm{nu}}}_{n_{\mathrm{nu}}}(\mathcal{H}) + 2\Phi^f_C(n_{\mathrm{nu}}, n_{\mathrm{de}}) + B_\ell\sqrt{8\left(\frac{1}{n_{\mathrm{de}}} + \frac{(1+C)^2}{n_{\mathrm{nu}}}\right)\log\frac{1}{\delta}}$, *where* $\Phi^f_C(n_{\mathrm{nu}}, n_{\mathrm{de}}) := (1+C)B_\ell \exp\left(-\frac{2\alpha^2}{({B_\ell}^2/n_{\mathrm{de}}) + (C^2{B_\ell}^2/n_{\mathrm{nu}})}\right)$ *and $\alpha > 0$ is a constant determined in the proof of Lemma 2 in Appendix H.*

Remark 5 (Explicit form of the bound in Theorem 1)**.** Here, we show the explicit form of the bound in Theorem 1 as follows:

$$\begin{aligned}
&\mathrm{BD}_f(\hat{r}) - \mathrm{BD}_f(\bar{r}) \\
&\leq \frac{\kappa_1}{\sqrt{n_{\mathrm{de}}}} + \frac{\kappa_2}{\sqrt{n_{\mathrm{nu}}}} + 2\Phi^f_C(n_{\mathrm{nu}}, n_{\mathrm{de}}) + B_\ell\sqrt{8\left(\frac{1}{n_{\mathrm{de}}} + \frac{(1+C)^2}{n_{\mathrm{nu}}}\right)\log\frac{1}{\delta}} \\
&= L_{\ell_1}\frac{B_{p_{\mathrm{de}}}\left(\sqrt{2\log(2)L}+1\right)\prod_{j=1}^L B_{W_j}}{\sqrt{n_{\mathrm{de}}}} \\
&\quad + 8(CL_{\ell_1} + L_{\ell_2})\frac{B_{p_{\mathrm{nu}}}\left(\sqrt{2\log(2)L}+1\right)\prod_{j=1}^L B_{W_j}}{\sqrt{n_{\mathrm{nu}}}} \\
&\quad + 2(1+C)B_\ell \exp\left(-\frac{2\alpha^2}{({B_\ell}^2/n_{\mathrm{de}}) + (C^2{B_\ell}^2/n_{\mathrm{nu}})}\right) \\
&\quad + B_\ell\sqrt{8\left(\frac{1}{n_{\mathrm{de}}} + \frac{(1+C)^2}{n_{\mathrm{nu}}}\right)\log\frac{1}{\delta}}.
\end{aligned}$$

I. Rademacher complexity bound

The following lemma provides an upper-bound on the Rademacher complexity for multi-layer perceptron models in terms of the Frobenius norms of the parameter matrices. Alternatively, other approaches to bound the Rademacher complexity can be employed. The assertion of the lemma follows immediately from the proof of Theorem 1 of Golowich et al. (2019) after a slight modification to incorporate the absolute value function in the definition of Rademacher complexity.

Lemma 3 (Rademacher complexity bound (Golowich et al., 2019, Theorem 1))**.** *Assume the distribution p has a bounded support: $B_p := \sup_{x\in\mathrm{supp}(p)}\|x\| < \infty$. Let $\mathcal{H}$ be the class of real-valued networks of depth L over the domain $\mathcal{X}$, where each parameter matrix W_j has Frobenius norm at most $B_{W_j} \geq 0$, and with 1-Lipschitz activation functions φ_j which are positive-homogeneous (i.e., φ_j is applied element-wise and $\varphi_j(\alpha t) = \alpha\varphi_j(t)$ for all $\alpha \geq 0$). Then*

$$\mathcal{R}^p_n(\mathcal{H}) \leq \frac{B_p\left(\sqrt{2\log(2)L}+1\right)\prod_{j=1}^L B_{W_j}}{\sqrt{n}}.$$

Proof. The assertion immediately follows once we modify the beginning of the proof of Theorem 1 by introducing the absolute value function inside the supremum of the Rademacher complexity as

$$\mathbb{E}_\sigma\left[\sup_{r\in\mathcal{H}}\left|\sum_{i=1}^n \sigma_i r(x_i)\right|\right] \leq \frac{1}{\lambda}\log \mathbb{E}_\sigma \sup_{r\in\mathcal{H}} \exp\left(\lambda\left|\sum_{i=1}\sigma_i r(x_i)\right|\right).$$

for $\lambda > 0$. The rest of the proof is identical to that of Theorem 1 of Golowich et al. (2019). □

J. Proof of Theorem 2

We consider relating the L^2 error bound to the BD generalization error bound in the following lemma.

Lemma 4 (L^2 distance bound)**.** *Let* $\mathcal{H} := \{r : \mathcal{X} \to (b_r, B_r) =: I_r | \int |r(x)|^2 \mathrm{d}\boldsymbol{x} < \infty\}$ *and assume* $r^* \in \mathcal{H}$*. If* $\inf_{t\in I_r} f''(t) > 0$*, then there exists* $\mu > 0$ *such that for all* $r \in \mathcal{H}$*,*

$$\|r - r^*\|^2_{L^2(p_{\mathrm{de}})} \leq \frac{2}{\mu}\left(\mathrm{BD}_f(r) - \mathrm{BD}_f(r^*)\right)$$

holds.

Proof. Since $\mu := \inf_{t\in I_r} f''(t) > 0$, the function f is μ-strongly convex. By the definition of strong convexity,

$$\begin{aligned}
&\mathrm{BD}_f(r) - \mathrm{BD}_f(r^*) = \left(\mathrm{BD}_f(r) - \mathbb{E}_{\mathrm{de}} f(r^*(X))\right) - \underbrace{\left(\mathrm{BD}_f(r^*) + \mathbb{E}_{\mathrm{de}} f(r^*(X))\right)}_{=0} \\
&= \mathbb{E}_{\mathrm{de}}\left[f(r^*(X)) - f(r(X)) + \partial f(r(X))(r^*(X) - r(X))\right] \\
&\geq \mathbb{E}_{\mathrm{de}}\left[\frac{\mu}{2}(r^*(X) - r(X))^2\right] = \frac{\mu}{2}\|r^* - r\|^2_{L^2(p_{\mathrm{de}})}.
\end{aligned}$$

□

Lemma 5 (ℓ_2 distance bound)**.** *Fix* $r \in \mathcal{H}$*. Given* n *samples* $\{x_i\}_{i=1}^n$ *from* p_{de}*, with probability at least* $1-\delta$*, we have*

$$\frac{1}{n}\sum_{i=1}^n (r(x_i) - r^*(x_i))^2 \leq \underbrace{\mathbb{E}\left[(r-r^*)^2(X)\right]}_{=\|r-r^*\|^2_{L^2(p_{\mathrm{de}})}} + (2\overline{R})^2\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Proof. The assertion follows from McDiarmid's inequality after noting that altering one sample results in an absolute change bounded by $\frac{1}{n}(2\overline{R})^2$. □

Thus, a generalization error bound in terms of BD_f can be converted to that of an L^2 distance when the true density ratio and the density ratio model are square-integrable and f is strongly convex. However, when using the result of Theorem 1, the convergence rate shown here is slower than $\mathcal{O}_{\mathbb{P}}\left((\min\{n_{\mathrm{de}}, n_{\mathrm{nu}}\})^{-1/(4)}\right)$. On the other hand, Kanamori et al. (2012) derived $\mathcal{O}_{\mathbb{P}}\left((\min\{n_{\mathrm{de}}, n_{\mathrm{nu}}\})^{-1/(2+\gamma)}\right)$ convergence rate. To derive this bound when using neural network, we need to restrict the neural network models. In the following part, we prove Theorem 2 for the following hypothesis class $\mathcal{H}$.

Definition 3 (ReLU neural networks; Schmidt-Hieber, 2020)**.** For $L \in \mathbb{N}$ and $p = (p_0, \ldots, p_{L+1}) \in \mathbb{N}^{L+2}$,

$$\begin{aligned}
\mathcal{F}(L,p) := \{f : x \mapsto &W_L\sigma_{v_L} W_{L-1}\sigma_{v_{L-1}}\cdots W_1\sigma_{v_1}W_0 x : \\
&W_i \in \mathbb{R}^{p_{i+1}\times p_i}, v_i \in \mathbb{R}^{p_i}(i = 0, \ldots, L)\},
\end{aligned}$$

where $\sigma_v(y) := \sigma(y - v)$, and $\sigma(\cdot) = \max\{\cdot, 0\}$ is applied in an element-wise manner. Then, for $s \in \mathbb{N}, F \geq 0, L \in \mathbb{N}$, and $p \in \mathbb{N}^{L+2}$, define

$$\mathcal{H}(L,p,s,F) := \{f \in \mathcal{F}(L,p) : \sum_{j=0}^L \|W_j\|_0 + \|v_j\|_0 \leq s, \|f\|_\infty \leq F\},$$

where $\|\cdot\|_0$ denotes the number of non-zero entries of the matrix or the vector, and $\|\cdot\|_\infty$ denotes the supremum norm. Now, fixing $\bar{L}, \bar{p}, s \in \mathbb{N}$ as well as $F > 0$, we define

$$\mathrm{Ind}_{\bar{L},\bar{p}} := \{(L,p) : L \in \mathbb{N}, L \leq \bar{L}, p \in [\bar{p}]^{L+2}\},$$

and we consider the hypothesis class

$$\begin{aligned}\bar{\mathcal{H}} &:= \bigcup_{(L,p)\in \mathrm{Ind}_{\bar{L},\bar{p}}} \mathcal{H}(L,p,s,F)\\ \mathcal{H} &:= \{r \in \bar{\mathcal{H}} : \mathrm{Im}(r) \subset (b_r, B_r)\}.\end{aligned}$$

Moreover, we define $I_1 : \mathrm{Ind}_{\bar{L},\bar{p}} \to \mathbb{R}$ and $I : \mathcal{H} \to [0,\infty)$ by

$$\begin{aligned}I_1(L,p) &:= 2|\mathrm{Ind}_{\bar{L},\bar{p}}|^{\frac{1}{s+1}}(L+1)V^2,\\ I(r) &:= \max\left\{\|r\|_\infty, \min_{\substack{(L,p)\in \mathrm{Ind}_{\bar{L},\bar{p}}\\ r\in\mathcal{H}(L,p,s,F)}} I_1(L,p)\right\},\end{aligned}$$

where $V := \prod_{l=0}^{L+1}(p_l+1)$, and we define

$$\mathcal{H}_M := \{r \in \mathcal{H} : I(r) \leq M\}.$$

Note that the requirement for the hypothesis class of Theorem 1 is not as tight as that of Theorem 2. Then, we prove Theorem 2 as follows:

Proof. Thanks to the strong convexity, by Lemma 4, we have

$$\begin{aligned}&\frac{\mu}{2}\|\hat{r} - r^*\|^2_{L^2(p_{\mathrm{de}})} \leq \mathrm{BD}_f(\hat{r}) - \mathrm{BD}_f(r^*)\\ &= \mathrm{BD}_f(\hat{r}) - \mathrm{BD}_f(r^*)\\ &\quad \underbrace{-\widehat{\mathrm{BD}}_f(\hat{r}) + \widehat{\mathrm{BD}}_f(\hat{r})}_{=0}\underbrace{-\widehat{\mathrm{nnBD}}_f(\hat{r}) + \widehat{\mathrm{nnBD}}_f(\hat{r})}_{=0}\underbrace{-\widehat{\mathrm{BD}}_f(r^*) + \widehat{\mathrm{BD}}_f(r^*)}_{=0}\\ &\leq \mathrm{BD}_f(\hat{r}) - \widehat{\mathrm{BD}}_f(\hat{r}) + (\widehat{\mathrm{BD}}_f(\hat{r}) - \widehat{\mathrm{nnBD}}_f(\hat{r}))\\ &\quad + (\widehat{\mathrm{nnBD}}_f(r^*) - \widehat{\mathrm{BD}}_f(r^*)) + \widehat{\mathrm{BD}}_f(r^*) - \mathrm{BD}_f(r^*)\\ &\leq \underbrace{(\mathrm{BD}_f(\hat{r}) - \mathrm{BD}_f(r^*) + \widehat{\mathrm{BD}}_f(r^*) - \widehat{\mathrm{BD}}_f(\hat{r}))}_{=:A} + \underbrace{2\sup_{r\in\mathcal{H}}|\widehat{\mathrm{BD}}_f(r) - \widehat{\mathrm{nnBD}}_f(r)|}_{=:B},\end{aligned}$$

where we used $\widehat{\mathrm{nnBD}}_f(\hat{r}) \leq \widehat{\mathrm{nnBD}}_f(r^*)$. To bound A, for ease of notation, let $\ell_1^r = \ell_1(r(X))$ and $\ell_2^r = \ell_2(r(X))$. Then, since

$$\begin{aligned}\mathrm{BD}_f(r) &= \mathbb{E}_{\mathrm{de}}\ell_1(r(X)) - C\mathbb{E}_{\mathrm{nu}}\ell_1(r(X)) + \mathbb{E}_{\mathrm{nu}}\ell_2(r(X)),\\ \widehat{\mathrm{BD}}_f(r) &= \hat{\mathbb{E}}_{\mathrm{de}}\ell_1(r(X)) - C\hat{\mathbb{E}}_{\mathrm{nu}}\ell_1(r(X)) + \hat{\mathbb{E}}_{\mathrm{nu}}\ell_2(r(X)),\end{aligned}$$

we have

$$\begin{aligned}A &= \mathrm{BD}_f(\hat{r}) - \mathrm{BD}_f(r^*) + \widehat{\mathrm{BD}}_f(r^*) - \widehat{\mathrm{BD}}_f(\hat{r})\\ &= (\mathbb{E}_{\mathrm{de}} - \hat{\mathbb{E}}_{\mathrm{de}})(\ell_1^{\hat{r}} - \ell_1^{r^*}) - C(\mathbb{E}_{\mathrm{nu}} - \hat{\mathbb{E}}_{\mathrm{nu}})(\ell_1^{\hat{r}} - \ell_1^{r^*}) + (\mathbb{E}_{\mathrm{nu}} - \hat{\mathbb{E}}_{\mathrm{nu}})(\ell_2^{\hat{r}} - \ell_2^{r^*})\\ &\leq |(\mathbb{E}_{\mathrm{de}} - \hat{\mathbb{E}}_{\mathrm{de}})(\ell_1^{\hat{r}} - \ell_1^{r^*})| + C|(\mathbb{E}_{\mathrm{nu}} - \hat{\mathbb{E}}_{\mathrm{nu}})(\ell_1^{\hat{r}} - \ell_1^{r^*})| + |(\mathbb{E}_{\mathrm{nu}} - \hat{\mathbb{E}}_{\mathrm{nu}})(\ell_2^{\hat{r}} - \ell_2^{r^*})|\end{aligned}$$

By applying Lemma 10, for any $0 < \gamma < 2$, we have

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

On the other hand, by Lemma 12 and Lemma 7, and the assumption $\inf_{r\in\mathcal{H}} \mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) > 0$, there exists $\alpha > 0$ such that we have $B \leq \mathcal{O}_{\mathbb{P}}\left(\exp\left(-\frac{2\alpha^2}{(B_\ell{}^2/n_{\mathrm{de}})+(C^2B_\ell{}^2/n_{\mathrm{nu}})}\right)\right)$. Combining the above bounds on A and B, for any $0 < \gamma < 2$, we get

$$\begin{aligned}
\|\hat{r} - r^*\|_{L^2(p_{\mathrm{de}})}^2 &\leq \mathcal{O}_{\mathbb{P}}\left(\max\left\{\frac{\|\hat{r} - r^*\|_{L^2(p_{\mathrm{de}})}^{1-\gamma/2}}{\sqrt{\min\{n_{\mathrm{de}}, n_{\mathrm{nu}}\}}}, \frac{1}{(\min\{n_{\mathrm{de}}, n_{\mathrm{nu}}\})^{2/(2+\gamma)}}\right\}\right) \\
&\quad + \mathcal{O}_{\mathbb{P}}\left(\exp\left(-\frac{2\alpha^2}{(B_\ell{}^2/n_{\mathrm{de}}) + (C^2B_\ell{}^2/n_{\mathrm{nu}})}\right)\right) \\
&\leq \mathcal{O}_{\mathbb{P}}\left(\max\left\{\frac{\|\hat{r} - r^*\|_{L^2(p_{\mathrm{de}})}^{1-\gamma/2}}{\sqrt{\min\{n_{\mathrm{de}}, n_{\mathrm{nu}}\}}}, \frac{1}{(\min\{n_{\mathrm{de}}, n_{\mathrm{nu}}\})^{2/(2+\gamma)}}\right\}\right).
\end{aligned}$$

As a result, we have

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

□

Each lemma used in the proof is provided as follows.

J.1. Complexity of the hypothesis class

For the function classes in Definition 3, we have the following evaluations of their complexities.

Lemma 6 (Lemma 5 in Schmidt-Hieber (2020))**.** *For $L \in \mathbb{N}$ and $p \in \mathbb{N}^{L+2}$, let $V := \prod_{l=0}^{L+1}(p_l + 1)$. Then, for any $\delta > 0$,*

$$\log \mathcal{N}(\delta, \mathcal{H}(L, p, s, \infty), \|\cdot\|_\infty) \leq (s+1)\log(2\delta^{-1}(L+1)V^2).$$

Lemma 7. *There exists $c > 0$ such that*

$$\mathcal{R}_{n_{\mathrm{nu}}}^{p_{\mathrm{nu}}}(\mathcal{H}) \leq cn_{\mathrm{nu}}^{-1/2}, \quad \mathcal{R}_{n_{\mathrm{de}}}^{p_{\mathrm{de}}}(\mathcal{H}) \leq cn_{\mathrm{de}}^{-1/2}.$$

Proof. By Dudley's entropy integral bound (Wainwright, 2019, Theorem 5.22) and Lemma 6, we have

$$\begin{aligned}
\mathcal{R}_{n_{\mathrm{nu}}}^{p_{\mathrm{nu}}}(\mathcal{H}(L, p, s, F)) &\leq 32\int_0^{2F}\sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{H}(L, p, s, F), \|\cdot\|_\infty)}{n_{\mathrm{nu}}}}d\delta \\
&= \left(32\int_0^{2F}\left((s+1)\log(2\delta^{-1}(L+1)V^2)\right)^{1/2}d\delta\right)n_{\mathrm{nu}}^{-1/2}.
\end{aligned}$$

Therefore, there exists $c > 0$ such that

$$\mathcal{R}_{n_{\mathrm{nu}}}^{p_{\mathrm{nu}}}(\mathcal{H}) \leq \sum_{(L,p)\in \mathrm{Ind}_{\bar{L},\bar{p}}} \mathcal{R}_{n_{\mathrm{nu}}}^{p_{\mathrm{nu}}}(\mathcal{H}(L, p, s, F)) \leq cn_{\mathrm{nu}}^{-1/2}.$$

The same argument applies to $\mathcal{R}_{n_{\mathrm{de}}}^{p_{\mathrm{de}}}(\mathcal{H})$, and we obtain the assertion. □

Lemma 8. *There exists $c_0 > 0$ such that for any $\gamma > 0$, any $\delta > 0$, and any $M \geq 1$, we have*

$$\log \mathcal{N}\left(\delta, \mathcal{H}_M, \|\cdot\|_\infty\right) \leq \frac{s+1}{\gamma}\left(\frac{M}{\delta}\right)^\gamma.$$

and

$$\sup_{r \in \mathcal{H}_M} \|r - r^*\|_\infty \leq c_0 M.$$

Proof. The first assertion is a result of the following calculation:

$$\begin{aligned}
\log \mathcal{N}\left(\delta, \mathcal{H}_M, \|\cdot\|_\infty\right) &\leq \log \sum_{\substack{(L,p) \in \mathrm{Ind}_{\bar{L},\bar{p}} \\ I_1(L,p) \leq M}} \mathcal{N}\left(\delta, \mathcal{H}(L,p,s,M), \|\cdot\|_\infty\right) \\
&\leq \log \sum_{\substack{(L,p) \in \mathrm{Ind}_{\bar{L},\bar{p}} \\ I_1(L,p) \leq M}} \left(\frac{2}{\delta}(L+1)V^2\right)^{s+1} \\
&\leq \log |\mathrm{Ind}_{\bar{L},\bar{p}}| \left(\frac{1}{\delta} M |\mathrm{Ind}_{\bar{L},\bar{p}}|^{-\frac{1}{s+1}}\right)^{s+1} \\
&= (s+1)\log\left(\frac{M}{\delta}\right) < (s+1)\frac{1}{\gamma}\left(\frac{M}{\delta}\right)^\gamma,
\end{aligned}$$

where the first inequality follows from $\mathcal{H}_M \subset \bigcup_{(L,p) \in \mathrm{Ind}_{\bar{L},\bar{p}} : I_1(L,p) \leq M} \mathcal{H}(L,p,s,F)$, and the last inequality from $\gamma \log x^{\frac{1}{\gamma}} = \log x < x$ that holds for all $x, \gamma > 0$.

The second assertion can be confirmed by noting that for any $r \in \mathcal{H}_M$ with $M \geq 1$,

$$\begin{aligned}
\|r - r^*\|_\infty &\leq \|r\|_\infty + \|r^*\|_\infty \leq M + \|r^*\|_\infty \\
&\leq \left(1 + \frac{\|r^*\|_\infty}{M}\right) M \leq (1 + \|r^*\|_\infty) M
\end{aligned}$$

holds.

□

Definition 4 (Derived function class and bracketing entropy)**.** Given a real-valued function class $\mathcal{F}$, define $\ell \circ \mathcal{F} := \{\ell \circ f : f \in \mathcal{F}\}$. By extension, we define $I : \ell \circ \mathcal{H} \to [1, \infty)$ by $I(\ell \circ r) = I(r)$ and $\ell \circ \mathcal{H}_M := \{\ell \circ r : r \in \mathcal{H}_M\}$. Note that, as a result, $\ell \circ \mathcal{H}_M$ coincides with $\{\ell \circ r \in \ell \circ \mathcal{H} : I(\ell \circ r) \leq M\}$.

Lemma 9. *Let $\ell : (b_r, B_r) \to \mathbb{R}$ be a ν-Lipschitz continuous function. Let $H_B\left(\delta, \mathcal{F}, \|\cdot\|_{L^2(P)}\right)$ denote the bracketing entropy of $\mathcal{F}$ with respect to a distribution P. Then, for any distribution P, any $\gamma > 0$, any $M \geq 1$, and any $\delta > 0$, we have*

$$H_B\left(\delta, \ell \circ \mathcal{H}_M, \|\cdot\|_{L^2(P)}\right) \leq \frac{(s+1)(2\nu)^\gamma}{\gamma}\left(\frac{M}{\delta}\right)^\gamma.$$

Moreover, there exists $c_0 > 0$ such that for any $M \geq 1$ and any distribution P,

$$\begin{aligned}
&\sup_{\ell \circ r \in \ell \circ \mathcal{H}_M} \|\ell \circ r - \ell \circ r^*\|_{L^2(P)} \leq c_0 \nu M, \\
&\sup_{\substack{\ell \circ r \in \ell \circ \mathcal{H}_M \\ \|\ell \circ r - \ell \circ r^*\|_{L^2(P)} \leq \delta}} \|\ell \circ r - \ell \circ r^*\|_\infty \leq c_0 \nu M, \quad \textit{for all } \delta > 0.
\end{aligned}$$

Proof. By combining Lemma 2.1 in van de Geer (2000) with Lemma 6, we have

$$\begin{aligned}
H_B\left(\delta, \ell \circ \mathcal{H}_M, \|\cdot\|_{L^2(P)}\right) &\leq \log \mathcal{N}\left(\frac{\delta}{2}, \ell \circ \mathcal{H}_M, \|\cdot\|_\infty\right), \\
&\leq \log \mathcal{N}\left(\frac{\delta}{2\nu}, \mathcal{H}_M, \|\cdot\|_\infty\right) \leq \frac{s+1}{\gamma}\left(\frac{2\nu M}{\delta}\right)^\gamma.
\end{aligned}$$

For $M \geq 1$, we have

$$\sup_{\ell \circ r \in \ell \circ \mathcal{H}_M} \|\ell \circ r - \ell \circ r^*\|_{L^2(P)} \leq \sup_{\ell \circ r \in \ell \circ \mathcal{H}_M} \|\ell \circ r - \ell \circ r^*\|_\infty$$
$$\sup_{\substack{\ell \circ r \in \ell \circ \mathcal{H}_M \\ \|\ell \circ r - \ell \circ r^*\|_{L^2(P)} \leq \delta}} \|\ell \circ r - \ell \circ r^*\|_\infty \leq \sup_{\ell \circ r \in \ell \circ \mathcal{H}_M} \|\ell \circ r - \ell \circ r^*\|_\infty,$$

and Lemma 6 implies

$$\sup_{\ell \circ r \in \ell \circ \mathcal{H}_M} \|\ell \circ r - \ell \circ r^*\|_\infty \leq \sup_{r \in \mathcal{H}_M} \nu \|r - r^*\|_\infty \leq \nu c_0 M.$$

□

J.2. Bounding the empirical deviations

Lemma 10. *Under the conditions of Theorem 2, for any* $0 < \gamma < 2$*, we have*

$$|(\mathbb{E}_{\mathrm{de}} - \hat{\mathbb{E}}_{\mathrm{de}})(\ell_1^{\hat{r}} - \ell_1^{r^*})| = \mathcal{O}_{\mathbb{P}}\left(\max\left\{\frac{\|\hat{r} - r^*\|_{L^2(p_{\mathrm{de}})}^{1-\gamma/2}}{\sqrt{n_{\mathrm{de}}}}, \frac{1}{n_{\mathrm{de}}^{2/(2+\gamma)}}\right\}\right)$$
$$|(\mathbb{E}_{\mathrm{nu}} - \hat{\mathbb{E}}_{\mathrm{nu}})(\ell_1^{\hat{r}} - \ell_1^{r^*})| = \mathcal{O}_{\mathbb{P}}\left(\max\left\{\frac{\|\hat{r} - r^*\|_{L^2(p_{\mathrm{de}})}^{1-\gamma/2}}{\sqrt{n_{\mathrm{nu}}}}, \frac{1}{n_{\mathrm{nu}}^{2/(2+\gamma)}}\right\}\right)$$
$$|(\mathbb{E}_{\mathrm{nu}} - \hat{\mathbb{E}}_{\mathrm{nu}})(\ell_2^{\hat{r}} - \ell_2^{r^*})| = \mathcal{O}_{\mathbb{P}}\left(\max\left\{\frac{\|\hat{r} - r^*\|_{L^2(p_{\mathrm{de}})}^{1-\gamma/2}}{\sqrt{n_{\mathrm{nu}}}}, \frac{1}{n_{\mathrm{nu}}^{2/(2+\gamma)}}\right\}\right)$$

as $n_{\mathrm{nu}}, n_{\mathrm{de}} \to \infty$.

Proof. Since $0 < \gamma < 2$, we can apply Lemma 11 in combination with Lemma 9 to obtain

$$\sup_{r \in \mathcal{H}} \frac{|(\mathbb{E}_{\mathrm{de}} - \hat{\mathbb{E}}_{\mathrm{de}})(\ell_1^r - \ell_1^{r^*})|}{D_1(r)} = \mathcal{O}_{\mathbb{P}}(1),$$
$$\sup_{r \in \mathcal{H}} \frac{|(\mathbb{E}_{\mathrm{nu}} - \hat{\mathbb{E}}_{\mathrm{nu}})(\ell_1^r - \ell_1^{r^*})|}{D_2(r)} = \mathcal{O}_{\mathbb{P}}(1),$$
$$\sup_{r \in \mathcal{H}} \frac{|(\mathbb{E}_{\mathrm{nu}} - \hat{\mathbb{E}}_{\mathrm{nu}})(\ell_2^r - \ell_2^{r^*})|}{D_3(r)} = \mathcal{O}_{\mathbb{P}}(1),$$

where

$$D_1(r) = \max\left\{\frac{\|\ell_1^r - \ell_1^{r^*}\|_{L^2(p_{\mathrm{de}})}^{1-\gamma/2} I(\ell_1^r)^{\gamma/2}}{\sqrt{n_{\mathrm{de}}}}, \frac{I(\ell_1^r)}{n_{\mathrm{de}}^{2/(2+\gamma)}}\right\},$$
$$D_2(r) = \max\left\{\frac{\|\ell_1^r - \ell_1^{r^*}\|_{L^2(p_{\mathrm{nu}})}^{1-\gamma/2} I(\ell_1^r)^{\gamma/2}}{\sqrt{n_{\mathrm{nu}}}}, \frac{I(\ell_1^r)}{n_{\mathrm{nu}}^{2/(2+\gamma)}}\right\},$$
$$D_3(r) = \max\left\{\frac{\|\ell_2^r - \ell_2^{r^*}\|_{L^2(p_{\mathrm{nu}})}^{1-\gamma/2} I(\ell_2^r)^{\gamma/2}}{\sqrt{n_{\mathrm{nu}}}}, \frac{I(\ell_2^r)}{n_{\mathrm{nu}}^{2/(2+\gamma)}}\right\},$$

Noting that $\sup_{r \in \mathcal{H}} I(r) < \infty$, that ℓ_2, ℓ_1 are Lipschitz continuous, and that $\|\hat{r} - r^*\|_{L^2(p_{\mathrm{nu}})} \leq \left(\sup_{x \in \mathcal{X}} \left|\frac{p_{\mathrm{nu}}(x)}{p_{\mathrm{de}}(x)}\right|\right) \|\hat{r} - r^*\|_{L^2(p_{\mathrm{de}})}$ holds, we have the assertion. □

Following is a proposition originally presented in van de Geer (2000), which was rephrased in Kanamori et al. (2012) in a form that is convenient for our purpose.

Lemma 11 (Lemma 5.14 in van de Geer (2000), Proposition 1 in Kanamori et al. (2012))**.** *Let $\mathcal{F} \subset L^2(P)$ be a function class and the map $I(f)$ be a complexity measure of $f \in \mathcal{F}$, where I is a non-negative function on $\mathcal{F}$ and $I(f_0) < \infty$ for a fixed $f_0 \in \mathcal{F}$. We now define $\mathcal{F}_M = \{f \in \mathcal{F} : I(f) \leq M\}$ satisfying $\mathcal{F} = \bigcup_{M \geq 1} \mathcal{F}_M$. Suppose that there exist $c_0 > 0$ and $0 < \gamma < 2$ such that*

$$\sup_{f \in \mathcal{F}_M} \|f - f_0\| \leq c_0 M, \quad \sup_{\substack{f \in \mathcal{F}_M \\ \|f - f_0\|_{L^2(P)} \leq \delta}} \|f - f_0\|_\infty \leq c_0 M, \quad \text{for all } \delta > 0,$$

and that $H_B(\delta, \mathcal{F}_M, P) = \mathcal{O}\left(M/\delta\right)^\gamma$. Then, we have

$$\sup_{f \in \mathcal{F}} \frac{\left|\int (f - f_0) d(P - P_n)\right|}{D(f)} = \mathcal{O}_{\mathbb{P}}\left(1\right), \ (n \to \infty),$$

where $D(f)$ is defined by

$$D(f) = \max\left\{ \frac{\|f - f_0\|_{L^2(P)}^{1-\gamma/2} I(f)^{\gamma/2}}{\sqrt{n}}, \frac{I(f)}{n^{2/(2+\gamma)}} \right\}.$$

J.3. Bounding the difference of the BD estimators

Lemma 12. *Assume $\mathcal{R}_{n_{\mathrm{de}}}^{p_{\mathrm{de}}}(\mathcal{H}) = o(1)(n_{\mathrm{de}} \to \infty)$ and $\mathcal{R}_{n_{\mathrm{nu}}}^{p_{\mathrm{nu}}}(\mathcal{H}) = o(1)(n_{\mathrm{nu}} \to \infty)$. Also assume the same conditions as Theorem 3. Then,*

$$\sup_{r \in \mathcal{H}} |\widehat{\mathrm{nnBD}}_f(r) - \widehat{\mathrm{BD}}_f(r)| = \mathcal{O}_{\mathbb{P}}\left(\exp\left(-\frac{2\alpha^2}{({B_\ell}^2/n_{\mathrm{de}}) + (C^2 {B_\ell}^2/n_{\mathrm{nu}})}\right)\right)$$

as $n_{\mathrm{nu}}, n_{\mathrm{de}} \to \infty$.

Proof. First, by combining Lemma 13, the assumption on the Rademacher complexities, and Markov's inequality, there exist $\alpha > 0$ and $n_{\mathrm{de}}^0, n_{\mathrm{nu}}^0 \in \mathbb{N}$ such that for any $n_{\mathrm{de}} \geq n_{\mathrm{de}}^0$ and $n_{\mathrm{nu}} \geq n_{\mathrm{nu}}^0$ and any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$,

$$\sup_{r \in \mathcal{H}} |\widehat{\mathrm{nnBD}}_f(r) - \widehat{\mathrm{BD}}_f(r)| \leq \frac{(1 + C) B_\ell L_{\rho - \mathrm{Id}}}{\delta} \exp\left(-\frac{2\alpha^2}{({B_\ell}^2/n_{\mathrm{de}}) + (C^2 {B_\ell}^2/n_{\mathrm{nu}})}\right).$$

Therefore, we have the assertion. □

Lemma 13. *Assume $\mathcal{R}_{n_{\mathrm{de}}}^{p_{\mathrm{de}}}(\mathcal{H}) = o(1)(n_{\mathrm{de}} \to \infty)$ and $\mathcal{R}_{n_{\mathrm{nu}}}^{p_{\mathrm{nu}}}(\mathcal{H}) = o(1)(n_{\mathrm{nu}} \to \infty)$. Also assume the same conditions as Theorem 3. Then, there exist $\alpha > 0$ and $n_{\mathrm{de}}^0, n_{\mathrm{nu}}^0 \in \mathbb{N}$ such that for any $n_{\mathrm{de}} \geq n_{\mathrm{de}}^0$ and $n_{\mathrm{nu}} \geq n_{\mathrm{nu}}^0$,*

$$\mathbb{E}\left[\sup_{r \in \mathcal{H}} |\widehat{\mathrm{nnBD}}_f(r) - \widehat{\mathrm{BD}}_f(r)|\right] \leq (1 + C) B_\ell L_{\rho - \mathrm{Id}} \exp\left(-\frac{2\alpha^2}{({B_\ell}^2/n_{\mathrm{de}}) + (C^2 {B_\ell}^2/n_{\mathrm{nu}})}\right)$$

holds.

Proof. First, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{r \in \mathcal{H}} |\widehat{\mathrm{nnBD}}_f(r) - \widehat{\mathrm{BD}}_f(r)|\right] \\
&= \mathbb{E}\left[\sup_{r \in \mathcal{H}} \left|\rho(\hat{\mathbb{E}}_{\mathrm{mod}} \ell_1(r(X))) - \hat{\mathbb{E}}_{\mathrm{mod}} \ell_1(r(X))\right|\right] \\
&= \mathbb{E}\left[\sup_{r \in \mathcal{H}} \mathbb{1}\{\rho(\hat{\mathbb{E}}_{\mathrm{mod}} \ell_1(r(X))) \neq \hat{\mathbb{E}}_{\mathrm{mod}} \ell_1(r(X))\} \cdot |\rho(\hat{\mathbb{E}}_{\mathrm{mod}} \ell_1(r(X))) - \hat{\mathbb{E}}_{\mathrm{mod}} \ell_1(r(X))|\right] \\
&\leq \mathbb{E}\left[\sup_{r \in \mathcal{H}} \mathbb{1}\{\rho(\hat{\mathbb{E}}_{\mathrm{mod}} \ell_1(r(X))) \neq \hat{\mathbb{E}}_{\mathrm{mod}} \ell_1(r(X))\}\right] \left(\sup_{s : |s| \leq (1+C) B_\ell} |\rho(s) - s|\right),
\end{aligned}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, and we used $|\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))| \leq (1+C)B_\ell$. Further, we have

$$\begin{aligned}&\sup_{s:|s|\leq(1+C)B_\ell}|\rho(s)-s| \leq \sup_{s:|s|\leq(1+C)B_\ell}|(\rho-\mathrm{Id})(s)-(\rho-\mathrm{Id})(0)|+|(\rho-\mathrm{Id})(0)|\\ &\leq \sup_{s:|s|\leq(1+C)B_\ell} L_{\rho-\mathrm{Id}}|s-0|+0 \leq (1+C)B_\ell L_{\rho-\mathrm{Id}},\end{aligned}$$

where Id denotes the identity function. On the other hand, since $\inf_{r\in\mathcal{H}}\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) > 0$ is assumed, there exists $\beta > 0$ such that for any $r \in \mathcal{H}$, $\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) > \beta$. Therefore, denoting the support of a function by $\mathrm{supp}(\cdot)$,

$$\begin{aligned}&\mathbb{E}\left[\sup_{r\in\mathcal{H}}\mathbb{1}\{\rho(\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))) \neq \hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))\}\right]\\ &= \mathbb{E}\left[\sup_{r\in\mathcal{H}}\mathbb{1}\{\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) \in \mathrm{supp}(\rho-\mathrm{Id})\}\right]\\ &= \mathbb{E}\left[\sup_{r\in\mathcal{H}}\mathbb{1}\{\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) < 0\}\right]\\ &= \mathbb{E}\left[\mathbb{1}\{\exists r\in\mathcal{H}:\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) < 0\}\right]\\ &= \mathbb{P}\left(\exists r\in\mathcal{H}:\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) < 0\right)\\ &\leq \mathbb{P}\left(\exists r\in\mathcal{H}:\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) < \mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\beta\right)\\ &\leq \mathbb{P}\left(\beta < \sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))\right).\end{aligned}$$

Take an arbitrary $\alpha \in (0,\beta)$. Since $\mathcal{R}^{p_{\mathrm{de}}}_{n_{\mathrm{de}}}(\mathcal{H}) \to 0 (n_{\mathrm{de}}\to\infty)$ and $\mathcal{R}^{p_{\mathrm{nu}}}_{n_{\mathrm{nu}}}(\mathcal{H}) \to 0 (n_{\mathrm{nu}}\to\infty)$, we can apply Lemma 14 and obtain the assertion. □

Lemma 14. *Let $\beta > \alpha > 0$. Assume that there exist $n^0_{\mathrm{de}}, n^0_{\mathrm{nu}} \in \mathbb{N}$ such that for any $n_{\mathrm{de}} \geq n^0_{\mathrm{de}}$ and $n_{\mathrm{nu}} \geq n^0_{\mathrm{nu}}$,*

$$4L_{\ell_1}\mathcal{R}^{p_{\mathrm{de}}}_{n_{\mathrm{de}}}(\mathcal{H}) + 4CL_{\ell_1}\mathcal{R}^{p_{\mathrm{nu}}}_{n_{\mathrm{nu}}}(\mathcal{H}) < \beta-\alpha.$$

Then, for any $n_{\mathrm{de}} \geq n^0_{\mathrm{de}}$ and $n_{\mathrm{nu}} \geq n^0_{\mathrm{nu}}$, we have

$$\begin{aligned}&\mathbb{P}\left(\beta < \sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))\right)\\ &\qquad\leq \exp\left(-\frac{2\alpha^2}{({B_\ell}^2/n_{\mathrm{de}})+(C^2{B_\ell}^2/n_{\mathrm{nu}})}\right).\end{aligned}$$

Proof. First, we will apply McDiarmid's inequality. The absolute difference caused by altering one data point in $\sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)) - \hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))$ is bounded by $\frac{B_\ell}{n_{\mathrm{de}}}$ if the change is in a sample from p_{de} and $\frac{CB_\ell}{n_{\mathrm{nu}}}$ otherwise. This can be confirmed by letting $\hat{\mathbb{E}}'_{\mathrm{mod}}$ denote the sample averaging operator obtained by altering one data point in $\hat{\mathbb{E}}_{\mathrm{mod}}$ and observing

$$\begin{aligned}&\sup_{r\in\mathcal{H}}\{\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))\}-\sup_{r\in\mathcal{H}}\{\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}'_{\mathrm{mod}}\ell_1(r(X))\}\\ &\leq \sup_{r\in\mathcal{H}}\{\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}'_{\mathrm{mod}}\ell_1(r(X)))\}\\ &\leq \sup_{r\in\mathcal{H}}\{\hat{\mathbb{E}}'_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))\}.\end{aligned}$$

The right-most expression can be bounded by $\frac{B_\ell}{n_{\mathrm{de}}}$ if the change is in a sample from p_{de} and $\frac{CB_\ell}{n_{\mathrm{nu}}}$ otherwise. Likewise, $\sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}'_{\mathrm{mod}}\ell_1(r(X)))-\sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))$ can be bounded by one of these

quantities. Therefore, we have

$$\begin{aligned}&\left|\sup_{r\in\mathcal{H}}\{\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))\}-\sup_{r\in\mathcal{H}}\{\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}'_{\mathrm{mod}}\ell_1(r(X))\}\right|\\&\leq\frac{B_\ell}{n_{\mathrm{de}}}+\frac{CB_\ell}{n_{\mathrm{nu}}},\end{aligned}$$

and McDiarmid's inequality implies, for any $\epsilon>0$,

$$\begin{aligned}&\mathbb{P}\left(\epsilon<\sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))-\mathbb{E}\left[\sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))\right]\right)\\&\leq\exp\left(-\frac{2\epsilon^2}{({B_\ell}^2/n_{\mathrm{de}})+(C^2{B_\ell}^2/n_{\mathrm{nu}})}\right).\end{aligned}\tag{10}$$

Now, applying Lemma 1, we have

$$\begin{aligned}&\mathbb{E}\left[\sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))\right]\\&\leq\mathbb{E}\left[\sup_{r\in\mathcal{H}}|\mathbb{E}_{\mathrm{de}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{de}}\ell_1(r(X))|\right]+C\mathbb{E}\left[\sup_{r\in\mathcal{H}}|\mathbb{E}_{\mathrm{nu}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{nu}}\ell_1(r(X))|\right]\\&\leq 4L_{\ell_1}\mathcal{R}^{p_{\mathrm{de}}}_{n_{\mathrm{de}}}(\mathcal{H})+4CL_{\ell_1}\mathcal{R}^{p_{\mathrm{nu}}}_{n_{\mathrm{nu}}}(\mathcal{H})=:\mathcal{R}.\end{aligned}$$

By the assumption, if $n_{\mathrm{de}}\geq n^0_{\mathrm{de}}$ and $n_{\mathrm{nu}}\geq n^0_{\mathrm{nu}}$, we have $\mathcal{R}<\beta-\alpha$. Therefore,

$$\mathbb{E}\left[\sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))\right]<\beta-\alpha<\beta,$$

hence $\beta-\mathbb{E}\left[\sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))\right]>0$. Therefore, we can take $\epsilon=\beta-\mathbb{E}\left[\sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))\right]$ in Equation (10) to obtain

$$\begin{aligned}&\mathbb{P}\left(\beta<\sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))\right)\\&\leq\exp\left(-\frac{2(\beta-\mathbb{E}\left[\sup_{r\in\mathcal{H}}(\mathbb{E}\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X))-\hat{\mathbb{E}}_{\mathrm{mod}}\ell_1(r(X)))\right])^2}{({B_\ell}^2/n_{\mathrm{de}})+(C^2{B_\ell}^2/n_{\mathrm{nu}})}\right)\\&\leq\exp\left(-\frac{2(\beta-\mathcal{R})^2}{({B_\ell}^2/n_{\mathrm{de}})+(C^2{B_\ell}^2/n_{\mathrm{nu}})}\right)\leq\exp\left(-\frac{2\alpha^2}{({B_\ell}^2/n_{\mathrm{de}})+(C^2{B_\ell}^2/n_{\mathrm{nu}})}\right),\end{aligned}$$

where we used $0<\alpha<\beta-\mathcal{R}$. □