Mediated Uncoupled Learning: Learning Functions without Direct Input-output Correspondences

Ikko Yamane [1 2] **Junya Honda** [3 2] **Florian Yger** [1 2] **Masashi Sugiyama** [2 4]

Abstract

Ordinary supervised learning is useful when we have paired training data of input X and output Y. However, such paired data can be difficult to collect in practice. In this paper, we consider the task of predicting Y from X when we have no paired data of them, but we have two separate, independent datasets of X and Y each observed with some mediating variable U, that is, we have two datasets $S_X = \{(X_i, U_i)\}$ and $S_Y = \{(U'_j, Y'_j)\}$. A naive approach is to predict U from X using S_X and then Y from U using S_Y, but we show that this is not statistically consistent. Moreover, predicting U can be more difficult than predicting Y in practice, e.g., when U has higher dimensionality. To circumvent the difficulty, we propose a new method that avoids predicting U but directly learns $Y = f(X)$ by training $f(X)$ with S_X to predict $h(U)$ which is trained with S_Y to approximate Y. We prove statistical consistency and error bounds of our method and experimentally confirm its practical usefulness.

1. Introduction

Supervised learning methods have been popular as powerful tools for many prediction tasks when we have training data consisting of direct correspondences between the output variable Y to be predicted and the input variable X to be used for the prediction (Murphy, 2012; Mohri et al., 2012; Shalev-Shwartz & Ben-David, 2014).

However, in some applications, it is difficult or expensive to collect training data consisting of (X, Y)-pairs (Chapelle et al., 2006; Zhu, 2005; van Engelen & Hoos, 2020). For example, consider the case where we want to learn a function predicting sentiment of an image (Mittal et al., 2018). Even if we do not have image data labeled with sentiment information, we might be able to find separate datasets consisting of images with text captions (Xu et al., 2015) and texts with sentiment labels (Medhat et al., 2014). Another example is translation between minor languages. If there are bilingual corpora in those languages, we could apply supervised learning techniques. However, it can be hard to obtain such training data since there may not be many speakers bilingual in minor languages. Instead, we may have a better chance to find separate translation corpora in each language with a major one such as English.

In this paper, we consider the situation in which we do not have access to direct correspondences between X and Y, but we only have two separate datasets of X and Y, each observed with some *mediating variable* U, $S_X = \{(X_i, U_i)\}$ and $S_Y = \{(U'_j, Y'_j)\}$, where (X_i, U_i) and (U'_j, Y'_j) are independent and thus we have no paired data of X and Y. Note that U_i and U'_j are generally different samples. In the example of image sentiment prediction, text captions can be used as the mediating variable U; S_X corresponds to image data with text captions, and S_Y corresponds to text data with sentiment labels. We call this framework *mediated uncoupled learning*.

A naive approach is to separately learn the function $U = g(X)$ using (X, U)-data and the function $Y = h(U)$ using (U, Y)-data. Then, one can predict Y from X by chaining the estimated functions as $\widehat{Y} := \widehat{h}(\widehat{U})$ with $\widehat{U} := \widehat{g}(X)$, where $\widehat{h}$ and $\widehat{g}$ are estimates of h and g, respectively. However, we show that this method is not statistically consistent since the point prediction $\widehat{U}$ does not carry enough information for predicting Y, unless Y and U have linear relationship or U is a deterministic function of X (see the detailed discussion in Section 3).

One can fix the inconsistency by (implicitly or explicitly) estimating the conditional probability density function (p.d.f.) $p(u \mid x)$ of U given X in place of the deterministic function g. Then, one can predict Y given $X = x$ by calculating $\int h(u)\widehat{p}(u \mid x)\mathrm{d}u$ with the estimated conditional p.d.f. $\widehat{p}(u \mid x)$. However, this approach involves the task of conditional density estimation or learning generative models, which needs delicate modeling and training (Salimans et al.,

[1]LAMSADE, CNRS, Université Paris-Dauphine, PSL Research University, 75016 PARIS, FRANCE [2]RIKEN AIP, Tokyo, Japan [3]Kyoto University, Kyoto, Japan [4]The University of Tokyo, Tokyo, Japan. Correspondence to: Ikko Yamane <ikko.yamane@dauphine.psl.eu>.

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

2016; Gulrajani et al., 2017; Kingma et al., 2016). Moreover, it requires integrating the estimated function at the prediction time, which can be computationally inefficient.

A cause of the weaknesses of these naive methods is that they try to predict U, which is unnecessary in order to solve the original task of predicting Y. To circumvent this issue, we propose a method that learns a function f directly predicting Y from X without attempting to predict U. Our proposed method first learns the correspondence $Y = h(U)$, and then train f so that $f(X)$ will best predict the output of $h(U)$. This simple approach allows us to use state-of-the-art supervised learning methods as building blocks out of the box while providing excellent theoretical and practical properties. Our theoretical analysis shows the statistical consistency and provides an excess error bound for our method. Finally, we demonstrate the practical usefulness of the proposed method through experiments.

2. Problem Setup

Our goal is to estimate a function that predicts a $\mathcal{Y}$-valued output variable Y from an $\mathcal{X}$-valued input variable X, where $\mathcal{X} \subseteq \mathbb{R}^{d_X}$ ($d_X \in \mathbb{N}$) and $\mathcal{Y} \subseteq \mathbb{R}$ are measurable spaces,[*1] and (X, Y) follows an unknown probability distribution with density $p(x, y)$. More specifically, the function that we want to estimate is the conditional expectation of Y given X, which is characterized as the minimizer of the mean squared error (MSE): $f^* := \mathbf{E}[Y \mid X] = \arg\min_{f \in L^2_X} \mathbf{E}[(f(X) - Y)^2]$, where $L^2_X := \{f : \mathcal{X} \to \mathbb{R} \mid \|f\|_2^2 := \mathbf{E}[f(X)^2] < \infty\}$ and $\mathbf{E}[\cdot]$ denotes the expectation over all involved variables. Note that the minimizer is unique in the sense that any minimizer f has distance zero from f^* in the L^2_X-norm: $\|f - f^*\|_2^2 = \mathbf{E}[(f(X) - Y)^2] - \mathbf{E}[(f^*(X) - Y)^2] = 0$. The MSE can be used for classification too, which corresponds to adopting the squared loss as a surrogate loss.[*2]

Unlike standard supervised problems, we have no access to direct supervision provided by joint samples of (X, Y). Instead, we assume that there exists a $\mathcal{U}$-valued *mediating variable* U for which there is a joint density function $p(x, u, y)$ of (X, U, Y), where $\mathcal{U} \subseteq \mathbb{R}^{d_U}$ ($d_u \in \mathbb{N}$) is a measurable space, and we are given two sets of i.i.d. samples, $\{(X_i, U_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p(x, u)$ and $\{(U'_i, Y'_i)\}_{i=1}^{n'} \overset{\text{i.i.d.}}{\sim} p(u, y)$. Here, $p(x, u)$ and $p(u, y)$ are the marginal p.d.f.-s of (X, U) and (U, Y), respectively, that are compatible with $p(x, u, y)$. We call these data *mediated uncoupled data* since X and Y are observed separately with a common variable U that mediates between them.

We assume the *conditional mean independence* given by

$$\mathbf{E}[Y \mid U] = \mathbf{E}[Y \mid U, X = x] \tag{1}$$

for every $x \in \mathcal{X}$. The condition ensures that (U, Y) has enough information to learn $\mathbf{E}[Y \mid X]$. Note that the conditional independence of Y and X given U, i.e., $Y \perp\!\!\!\perp X \mid U$, implies the conditional mean independence, Eq. (1), but the converse is not true. We discuss the case where this assumption is not satisfied in Section 5.5.

A related but different problem setting of learning without input-output correspondences was studied in Zhang et al. (2019). They considered the situation in which, using our notation, $U \perp\!\!\!\perp X \mid Y$ and the conditional probability of Y given U is given instead of (U, Y)-pairs, which is a typical scenario in learning with noisy labels (Angluin & Laird, 1988; Blanchard et al., 2016; Natarajan et al., 2013). Also, these methods focus on the case where Y is discrete. Yamane et al. (2018) considered estimation of causal effect using data separately labeled with either treatment or outcome but not both at the same time. The type of training data is similar to ours, but the problem setup is essentially different.

Dhir & Lee (2020) proposed a method for causal discovery under a similar setup in which not all combinations of the variables of interest are jointly observed. The form of data that they assumed includes ours as a special case, and they provided some real-world examples in which such data arise. However, their goal is to infer causal directions betweeen variables but not to predict their values, and thus their method is not applicable to our problem.

3. Naive Approach Based on Separate Estimators

A naive approach to this problem is to estimate $g^*(x) := \mathbf{E}[U \mid X = x]$ and $h^*(u) := \mathbf{E}[Y \mid U = u]$ as $g(x)$ and $h(u)$, respectively, and then combine them as $f_{\text{combine}}(x) := h(g(x))$ to estimate $\mathbf{E}[Y \mid X = x]$. The combined estimator is consistent when h^* is a linear function, or U is a deterministic function of X so that

$$h^*(g^*(x)) = h^*(\mathbf{E}[U \mid X = x]) = \mathbf{E}[h^*(U) \mid X = x],$$

in which case, we have

$$\begin{aligned} h^*(g^*(x)) &= \mathbf{E}[h^*(U) \mid X = x] \\ &= \mathbf{E}[\mathbf{E}[Y \mid U] \mid X = x] \\ &= \mathbf{E}[\mathbf{E}[Y \mid U, X = x] \mid X = x] \quad \text{(from Eq. (1))} \\ &= \mathbf{E}[Y \mid X = x]. \end{aligned} \tag{2}$$

Hence, the consistency of estimators of h^* and g^* will guarantee the consistency of their composite function to $\mathbf{E}[Y \mid X]$.

[*1] We can easily extend all the results to the case where $\mathcal{Y}$ is multi-dimensional. This manuscript focuses on the one-dimensional case for the ease of notation.

[*2] In the case of multi-class classification, we can use the squared loss with the one-hot representation for class labels.

However, it fails to consistently estimate the target function $\mathbf{E}[Y \mid X = x]$ in many important non-linear cases. For example, when h^* is strictly convex, and U is a stochastic function of X, Jensen's inequality implies

$$\begin{aligned} h^*(g^*(x)) &= h^*(\mathbf{E}[U \mid X = x]) \\ &< \mathbf{E}[h^*(U) \mid X = x] \\ &= \mathbf{E}[Y \mid X = x], \end{aligned}$$

where the last line follows from Eq. (2). Thus, the estimator under-estimates the target, and it is not consistent.

One can develop a consistent version of the naive method by estimating the conditional density function $p(u \mid x)$ of U given X instead of the conditional expectation $\mathbf{E}[U \mid X = x]$. Once $p(u \mid x)$ is estimated as $q(u \mid x)$ and $\mathbf{E}[Y \mid U = u]$ as $h(u)$, one can estimate $\mathbf{E}[Y \mid X = x]$ as

$$f_{\text{integral}}(x) \coloneqq \int h(u) q(u \mid x) \mathrm{d}u. \tag{3}$$

Then, f_{integral} is a consistent estimator of $\mathbf{E}[Y \mid X = x]$ as long as $q(u \mid x)$ and $h(u)$ are consistent because it converges to

$$\begin{aligned} &\int h^*(u) p(u \mid x) \mathrm{d}u \\ &= \int \mathbf{E}[Y \mid U = u] p(u \mid x) \mathrm{d}u \\ &= \int \mathbf{E}[Y \mid U = u, X = x] p(u \mid x) \mathrm{d}u \quad \text{(by Eq. (1))} \\ &= \mathbf{E}[Y \mid X = x]. \end{aligned}$$

However, this modified method solves the hard intermediate problem of estimating the conditional probability density function $p(u \mid x)$. This can be particularly problematic when we use neural networks because it has been reported that they tend to be overconfident and show poor performance in predicting the conditional probability of the output (Hein et al., 2019). Moreover, one needs to be able to accurately calculate the integral in Eq. (3), e.g., by sampling from $q(u \mid x)$, at each prediction, which is often computationally demanding and prohibitive when we need real-time responses in prediction. Although the efficient belief-propagation based algorithm proposed by Song et al. (2010) can be used when distributions are represented by reproducing kernel Hilbert space (RKHS) embeddings, it is not generally applicable when we use function classes other than RKHSs, such as neural networks.

In fact, for several problems that are solvable by performing density estimation as intermediate tasks, directly solving the target task without solving density estimation reportedly improves performance (Sugiyama et al., 2012; 2013; Sasaki et al., 2014). Vapnik (1995) also argued that it is preferable to avoid solving intermediate tasks that are more general than the target task.

Another higher-level criticism of the naive approach above from a statistical point of view is that the intermediate step of estimating $\mathbf{E}[U \mid X]$ (or $p(u \mid x)$) is performed without any attention to the target task of predicting Y. This means that those estimators are not designed in a way that the resulting prediction for Y will be accurate.

4. Proposed Methods

Our approach learns a function $f \colon \mathcal{X} \to \mathcal{Y}$ that directly predicts Y from X without predicting U. It does not try to solve the hard intermediate problem of predicting U from X, and it is free from the issues that the naive approach suffers. Below, we describe our approach to the problem more precisely and propose two methods based on it.

4.1. Two-step Regressed Regression (2Step-RR)

The first proposed method consists of two steps. The first step is for training a function $h : \mathcal{U} \to \mathcal{Y}$ to predict Y from U using S_Y. Because each sample of U in S_Y is labeled with the corresponding sample of Y, this step is no more than an ordinary supervised learning task. Now, h can predict Y, but its input is U, not X. To obtain a function f that takes X as input and predicts Y, we train f so that the output of $f(X)$ will mimic that of $h(U)$, for which we only need S_X consisting of samples of (X, U).

More specifically, we first train a function $\widetilde{h} \colon \mathcal{U} \to \mathcal{Y}$ for predicting Y from U:

$$\widetilde{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}} \frac{1}{n'} \sum_{i=1}^{n'} [(h(U'_i) - Y'_i)^2].$$

Then, we train another function $\widetilde{f} \colon \mathcal{X} \to \mathcal{Y}$ for predicting $\widetilde{h}(U)$ from X:

$$\widetilde{f} = \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} [(f(X_i) - \widetilde{h}(U_i))^2].$$

In the above, $\mathcal{H}$ and $\mathcal{F}$ are function classes for $\widetilde{h}$ and $\widetilde{f}$, respectively. Because we train $\widetilde{h}$ so that $\widetilde{h}(U)$ will predict Y well, $\widetilde{f}(X)$ is expected to predict Y well by predicting $\widetilde{h}(U)$. We call this method *Two-Step Regressed Regression (2Step-RR)* since the predictive function f is learned by regressing the regression of Y.

Note that the objective functional in each step uses samples of either (X, U) or (U, Y), not both at the same time. Thus, we can compute it with our mediated uncoupled data, $\{(X_i, U_i)\}_{i=1}^{n}$ and $\{(U'_i, Y'_i)\}_{i=1}^{n'}$. We summarize the algorithm in Algorithm 1. In Section 5, we will show that

Algorithm 1 Two-Step Regressed Regression (2Step-RR)

$\widetilde{h} \leftarrow \arg\min_{h\in\mathcal{H}} \frac{1}{n'}\sum_{i=1}^{n'}(h(U'_i) - Y'_i)^2$.
$\widetilde{f} \leftarrow \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \widetilde{h}(U_i))^2$.
Return: $\widetilde{f}$

Algorithm 2 Joint Regressed Regression (Joint-RR)

$(\widehat{f}_w, \widehat{h}_w) \coloneqq \arg\min_{f\in\mathcal{F},h\in\mathcal{H}} \widehat{J}_w(f,h)$ (see Eq. (4).)
Return: $\widehat{f}_w$

this method has statistical consistency and admits a nice non-asymptotic error bound.

4.2. Jointly Performing the Two Steps

2Step-RR described above has nice theoretical properties (see Section 5), but there may be room for improvement on practical, finite-sample performance. More specifically, while 2Step-RR uses no information for training f when h is trained, it may be advantageous to let h adapt to the second step in a way that it will be easier for $f(X)$ to fit $h(U)$. Here, we are going to combine the two steps of 2Step-RR to develop a variant called *Joint Regressed Regression (Joint-RR)* that trains f and h at the same time. This allows h to incorporate how well $f(X)$ can fit $h(U)$ and adjust itself in favor of the training of f.

4.2.1. Joint Regressed Regression (Joint-RR)

The procedure of Joint-RR itself is simple (see Algorithm 2). We additively combine the two objective functionals used for training f and h in 2Step-RR:

$$\widehat{J}_w(f,h) \coloneqq \frac{1}{wn}\sum_{i=1}^{n}(f(X_i) - h(U_i))^2 + \frac{1}{(1-w)n'}\sum_{i=1}^{n'}(h(U'_i) - Y'_i)^2, \quad (4)$$

where $w \in (0,1)$ is a weight parameter. Then, we minimize Eq. (4) with respect to f and h jointly:

$$(\widehat{f}_w, \widehat{h}_w) \coloneqq \arg\min_{f\in\mathcal{F},h\in\mathcal{H}} \widehat{J}_w(f,h).$$

In the rest of this section, we will give more detailed justification of Joint-RR as upper bound minimization.

4.2.2. Joint-RR as Upper Bound Minimization

We start by constructing an upper bound of the population version of the MSE that can be approximated with our mediated uncoupled data, S_X and S_Y.

Theorem 4.1. *The MSE can be bounded as*

$$\mathbf{E}[(f(X) - Y)^2] \le J_w(f,h), \quad (5)$$

where

$$J_w(f,h) \coloneqq \frac{1}{w}\mathbf{E}[(f(X) - h(U))^2] + \frac{1}{1-w}\mathbf{E}[(h(U) - Y)^2]$$

for any $w \in (0,1)$ *and any* $h \in L_U^2 \coloneqq \{h\colon \mathcal{U} \to \mathcal{Y} \mid \mathbf{E}[h(U)^2] < \infty\}$.

A proof is in Appendix I in the supplementary material. The functional $J_w(f,h)$ has terms each involving either (X,U) or (U,Y) but not both at the same time. This convenient property allows us to approximate it with the mediated uncoupled data, $S_X = \{(X_i,U_i)\}_{i=1}^n$ and $S_Y = \{(U'_i,Y'_i)\}_{i=1}^{n'}$. Among the upper bounds of the form in Eq. (5), we take the tightest one with respect to h:

$$\mathbf{E}[(f(X) - Y)^2] \le \min_{h\in L_X^2} J_w(f,h). \quad (6)$$

The minimization of the right-hand side of Eq. (6) yields the population version of our optimization problem:

$$\arg\min_{f\in L_X^2} \min_{h\in L_U^2} J_w(f,h). \quad (7)$$

In practice, we solve its empirical version with some hypothesis classes $\mathcal{F}$ and $\mathcal{H}$ for f and h, respectively, to obtain our estimator:

$$\widehat{f}_w \coloneqq \arg\min_{f\in\mathcal{F}} \min_{h\in\mathcal{H}} \widehat{J}_w(f,h), \quad (8)$$

where $\widehat{J}_w(f,h)$ is defined by Eq. (4).

The argument above is valid for any $w \in (0,1)$. How to optimize w so as to minimize the test MSE is not trivial since we do not have (X,Y)-data for validation, and we leave this as an open question. In our experiments, we simply fixed w to the balanced value $1/2$, which performs well in many cases and tends to show more stable performance compared to 2Step-RR. In Section 5.1, we will show that 2Step-RR is the limit of Joint-RR with $w \to 1$.

4.2.3. Closed-form Solution for Linear-in-parameter Models

When interpretation or fast prediction is required, linear-in-parameter-models would be useful. Fortunately, our methods admit closed-form solutions for those models. Due to the limited space, we only present the solution to Joint-RR here. The derivation for 2Step-RR is more straightforward.

Theorem 4.2. *Let* $f_{\boldsymbol{\alpha}}(\boldsymbol{x}) \coloneqq \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x})$, $h_{\boldsymbol{\beta}}(\boldsymbol{u}) \coloneqq \boldsymbol{\beta}^\top \boldsymbol{\psi}(\boldsymbol{u})$, $\boldsymbol{\theta} \coloneqq (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$, *and* $\lambda \in (0, \infty)$. *Then, the* ℓ_2*-regularized solution*

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) \coloneqq \operatorname*{arg\,min}_{(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{R}^{b_{\mathcal{F}}} \times \mathbb{R}^{b_{\mathcal{H}}}} \left[\widehat{J}_w(f_{\boldsymbol{\alpha}}, h_{\boldsymbol{\beta}}) + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta} \right]$$

is given by

$$\widehat{\boldsymbol{\alpha}} \coloneqq \boldsymbol{M}_1^{-1} \boldsymbol{M}_2 \widehat{\boldsymbol{\beta}}, \quad \textit{and} \tag{9}$$

$$\widehat{\boldsymbol{\beta}} \coloneqq (\boldsymbol{M}_3 - \boldsymbol{M}_2^\top \boldsymbol{M}_1^{-1} \boldsymbol{M}_2)^{-1} \boldsymbol{b}_1, \tag{10}$$

where

$$\boldsymbol{M}_1 \coloneqq \frac{1}{nw} \sum_{i=1}^{n} \boldsymbol{\varphi}(X_i) \boldsymbol{\varphi}(X_i)^\top + \lambda \boldsymbol{I}_{b_{\mathcal{F}}},$$

$$\boldsymbol{M}_2 \coloneqq \frac{1}{nw} \sum_{i=1}^{n} \boldsymbol{\varphi}(X_i) \boldsymbol{\psi}(U_i)^\top,$$

$$\boldsymbol{M}_3 \coloneqq \frac{1}{nw} \sum_{i=1}^{n} \boldsymbol{\psi}(U_i) \boldsymbol{\psi}(U_i)^\top + \frac{1}{n'(1-w)} \sum_{i=1}^{n'} \boldsymbol{\psi}(U'_i) \boldsymbol{\psi}(U'_i)^\top + \lambda \boldsymbol{I}_{b_{\mathcal{H}}},$$

$$\boldsymbol{b}_1 \coloneqq \frac{1}{n'(1-w)} \sum_{i=1}^{n'} Y'_i \boldsymbol{\psi}(U'_i).$$

Eqs. (9) and (10) involve matrices of size at most $\max(b_{\mathcal{F}}, b_{\mathcal{H}})$-by-$\max(b_{\mathcal{F}}, b_{\mathcal{H}})$, which requires less computational resources in terms of both space and time compared to a naive solution involving the inversion of a $(b_{\mathcal{F}} + b_{\mathcal{H}})$-by-$(b_{\mathcal{F}} + b_{\mathcal{H}})$ matrix. Details and a proof are presented in Appendix J in the supplementary material.

5. Theoretical Analysis

In this section, we present several theoretical results.

5.1. Connection between 2Step-RR and Joint-RR

2Step-RR and Joint-RR have an interesting connection that helps us better understand them. Briefly speaking, 2Step-RR can be seen as a special case of Joint-RR in the sense that we can obtain the former by taking the limit of the latter with $w \to 1$. The following theorem provides a more formal statement on the connection.

Theorem 5.1. *Suppose that* $\mathcal{F} \subseteq L_X^2$ *and* $\mathcal{H} \subseteq L_U^2$ *satisfy* $h^*(u) \coloneqq \mathbf{E}[Y \mid U = u] \in \mathcal{H}$ *and* $f^*(x) \coloneqq \mathbf{E}[h^*(U) \mid X = x] \in \mathcal{F}$. *Then,*

$$(f^*, h^*) \in \lim_{w \uparrow 1} \operatorname*{arg\,min}_{(f,h) \in \mathcal{F} \times \mathcal{H}} J_w(f, h). \tag{11}$$

We present a proof in Appendix C in the supplementary material. Note that (f^*, h^*) is the solution pair to the optimization problem of 2Step-RR with the population-level objective functionals:

$$h^* \in \operatorname*{arg\,min}_{h \in \mathcal{H}} \mathbf{E}[(h(U) - Y)^2]$$
$$\text{and} \quad f^* \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \mathbf{E}[(f(X) - h^*(U))^2].$$

The theorem states that (f^*, h^*) is also equal to the limit of the solution pair to the Joint-RR optimization problem with the population-level objective functional in Eq. (7).

5.2. Statistical Consistency of 2Step-RR

Let us informally confirm the statistical consistency of 2Step-RR under the conditional mean independence (Eq. (1)). When the models are correctly specified, and Eq. (1) and appropriate convergence conditions hold, the 2Step-RR estimator $\widetilde{f}$ converges as $n \to \infty$ and $n' \to \infty$ to

$$\begin{aligned} f^*(x) &= \mathbf{E}[h^*(U) \mid X = x] \\ &= \mathbf{E}[\mathbf{E}[Y \mid U] \mid X = x] \\ &= \mathbf{E}[\mathbf{E}[Y \mid U, X] \mid X = x] \quad \text{(from Eq. (1))} \\ &= \mathbf{E}[Y \mid X = x]. \end{aligned}$$

Thus, it is consistent under the condition of Eq. (1). We can formally confirm this as a corollary of Theorem 5.2 given later.

5.3. Joint-RR is a Regularized Method

Unlike 2Step-RR, Joint-RR is not statistically consistent, but it can be seen as a nice regularized counterpart in the sense that its objective function is the MSE plus the deviation of $f(X)$ from the conditional mean $\mathbf{E}[f(X) \mid U]$.

Under the assumption in Eq. (1), we have

$$\begin{aligned} &\mathrm{MSE}(f) \coloneqq \mathbf{E}[(Y - f(X))^2] \\ &= \mathbf{E}[(\mathbf{E}[Y \mid U] - f(X))^2] + \mathbf{E}[(Y - \mathbf{E}[Y \mid U])^2], \end{aligned}$$

where the last term is a constant that does not depend on f. Hence,

$$\begin{aligned} &\min_{h \in L_U^2} J_w(f, h) \\ &= \mathbf{E}[(\mathbf{E}[Y \mid U] - f(X))^2] \\ &\quad + \frac{1-w}{w} \mathbf{E}[(f(X) - \mathbf{E}[f(X) \mid U])^2] + \text{ const.} \\ &= \mathrm{MSE}(f) + \text{const.} \\ &\quad + \underbrace{\frac{1-w}{w} \mathbf{E}[(f(X) - \mathbf{E}[f(X) \mid U])^2]}_{\text{The shrinkage regularizer.}} \end{aligned} \tag{12}$$

for any $w \in (0, 1)$. See Appendix G in the supplementary material for a more detailed calculation.

This shows that Joint-RR with $w < 1$ minimizes a biased objective functional. However, it is often favorable to trade off some bias for smaller variance in practice. We can also see that the amplitude of the shrinkage term can be controlled by w and it vanishes at the limit of $w \to 1$. In our experiments, we simply fixed w to the balanced value $1/2$, which performs well in many cases.

5.4. Excess Error Bound

2Step-RR solves a simple least-squares problem twice. This allows us to derive a non-asymptotic bound of the excess error $\mathrm{MSE}(\widetilde{f}) - \mathrm{MSE}(f^\dagger)$, where $f^\dagger(x) \coloneqq \mathbf{E}[Y \mid X = x]$, in terms of the Rademacher complexities of function classes.

Theorem 5.2. *Let $C_{\mathcal{F}} \coloneqq \sup_{f \in \mathcal{F}, x \in \mathcal{X}} f(x) < \infty$, $C_{\mathcal{H}} \coloneqq \sup_{h \in \mathcal{H}, u \in \mathcal{U}} f(u) < \infty$, and $C_{\mathcal{Y}} \coloneqq \sup \mathcal{Y} < \infty$. Let $\mathfrak{R}_n(\mathcal{F})$ denote the Rademacher complexity of $\mathcal{F}$ over $\{(X_i, U_i)\}_{i=1}^n$ and $\mathfrak{R}_{n'}(\mathcal{H})$ denote that of $\mathcal{H}$ over $\{(U'_i, Y'_i)\}_{i=1}^{n'}$ (see the exact definitions in Appendix D in the supplementary material). Suppose that $h^* \in \mathcal{H}$, $f^* \in \mathcal{F}$, and $f^\dagger \in \mathcal{F}$. Then, the excess error can be bounded as*

$$
\begin{aligned}
&\mathbf{E}[(\widetilde{f}(X) - f^\dagger(X))^2] = \mathrm{MSE}(\widetilde{f}) - \mathrm{MSE}(f^\dagger) \\
&\leq 8(C_{\mathcal{F}} + C_{\mathcal{H}})(\mathfrak{R}_n(\mathcal{F}) + \mathfrak{R}_n(\mathcal{H})) \\
&\quad + 8(C_{\mathcal{H}} + C_{\mathcal{Y}})\mathfrak{R}_{n'}(\mathcal{H}) \\
&\quad + 4(C_{\mathcal{F}} + C_{\mathcal{H}})^2 \sqrt{\frac{2}{n} \log \frac{1}{\delta}} \\
&\quad + 2(C_{\mathcal{H}} + C_{\mathcal{Y}})^2 \sqrt{\frac{2}{n'} \log \frac{1}{\delta}} \\
&\leq \mathcal{O}_p\Bigg(\mathfrak{R}_n(\mathcal{F}) + \mathfrak{R}_n(\mathcal{H}) + \mathfrak{R}_{n'}(\mathcal{H}) \\
&\qquad + \sqrt{(\frac{1}{n} + \frac{1}{n'}) \log \frac{1}{\delta}} \Bigg).
\end{aligned}
$$

For instance, when $\mathcal{F}$ and $\mathcal{H}$ are bounded linear-in-parameter models, $\mathfrak{R}_n(\mathcal{F}) = \mathcal{O}(1/\sqrt{n})$, $\mathfrak{R}_n(\mathcal{H}) = \mathcal{O}(1/\sqrt{n})$, and $\mathfrak{R}_{n'}(\mathcal{H}) = \mathcal{O}(1/\sqrt{n'})$ (Mohri et al., 2012). Using Theorem 5.2, we can bound the excess error by $\mathcal{O}_p(1/\sqrt{n} + 1/\sqrt{n'})$. A proof is in Appendix F in the supplementary material.

5.5. Discussion on the Assumption

So far, we have focused on the ideal case in which the conditional mean independence (Eq. (1)) holds. However, it may be difficult to exactly ensure the condition in practice.

Here, we relax Eq. (1) by allowing the gap between the left-hand and right-hand sides to be potentially larger than zero but bounded by c^2 for some constant $c \in (0, \infty)$. We will show that (i) even the best possible method suffers an MSE of at least $c^2/2$ in the worse-case within this scenario while (ii) 2Step-RR suffers an MSE of at most $c^2 + o(1)$.

To see the claim (ii), notice that we have already shown that $\widetilde{f}$ converges to $\mathbf{E}[\mathbf{E}[Y \mid U] \mid X = (\cdot)]$. This implies that 2Step-RR suffers an MSE of

$$
\begin{aligned}
&\mathbf{E}[(\mathbf{E}[\mathbf{E}[Y \mid U] \mid X] - \mathbf{E}[Y \mid X])^2] + o(1) \\
&= \mathbf{E}[(\mathbf{E}[\mathbf{E}[Y \mid U] - \mathbf{E}[Y \mid U, X] \mid X])^2] + o(1) \\
&\leq \mathbf{E}[(\mathbf{E}[Y \mid U] - \mathbf{E}[Y \mid U, X])^2] + o(1) \\
&\leq c^2 + o(1)
\end{aligned}
$$

under the relaxed assumption.

The following proposition is a formal statement of the claim (i). A proof is in Appendix H in the supplementary material.

Proposition 5.1. *For any estimator $\widehat{f}_{(\cdot)}$ that takes mediated uncoupled data S_X and S_Y as input and produces a function from $\mathcal{X}$ to $\mathcal{Y}$, we have*

$$
\sup_{p^* \in \mathcal{P}_c} \mathbf{E}\left[\left(\widehat{f}_{S_X, S_Y}(X) - \mathbf{E}[Y \mid X] \right)^2 \right] \geq \frac{1}{2} c^2,
$$

where $(X, Y) \sim p^*(x, y)$, and $\mathcal{P}_c$ is the class of p.d.f.-s $\widetilde{p}(x, u, y)$ for which $\mathbf{E}[(\mathbf{E}[\widetilde{Y} \mid \widetilde{X}] - \mathbf{E}[\widetilde{Y} \mid \widetilde{X}, \widetilde{U}])^2] \leq c^2$, $\widetilde{p}(x) = \widetilde{p}(-x)$, and $\mathbf{E}[\widetilde{U}] = 0$ with $(\widetilde{X}, \widetilde{U}, \widetilde{Y}) \sim \widetilde{p}(x, u, y)$.

Our bound relies on Le Cam's method, which yields the $1/2$ factor (see Appendix H in the supplementary material for details). Whether one can eliminate the factor remains as future work.

6. Experiments

In this section, we present experimental results.

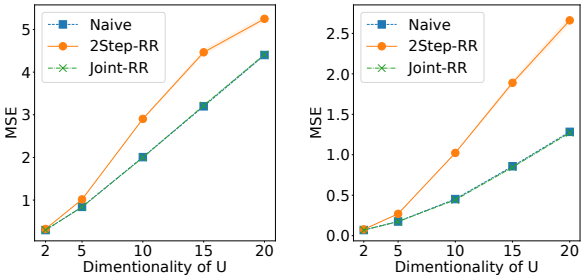
6.1. Experiments with Synthetic Data

First, we present experiments with synthetic data. Because neural networks are becoming the gold standard in many tasks, we test the methods using neural networks as follows.

- The naive method using multi-layer perceptrons with four layers, 20 hidden units in each layer, and ReLU activations. We refer to this method as "Naive".
- 2Step-RR using multi-layer perceptrons with four layers, 20 hidden units in each layer, and ReLU activations.
- Joint-RR using multi-layer perceptrons with four layers, 20 hidden units in each layer, and ReLU activations. We set $w = 1/2$.

We train all models with Adam (Kingma & Ba, 2017) for 200 epochs. We implemented the methods using PyTorch (Paszke et al., 2019)[*3] We use the default values

[*3]The code will be available on https://github.com/i-yamane/mediated_uncoupled_learning.



(a) The setting satisfying the conditional mean independence (Eq. (1)).

(b) The setting violating the conditional mean independence (Eq. (1)).

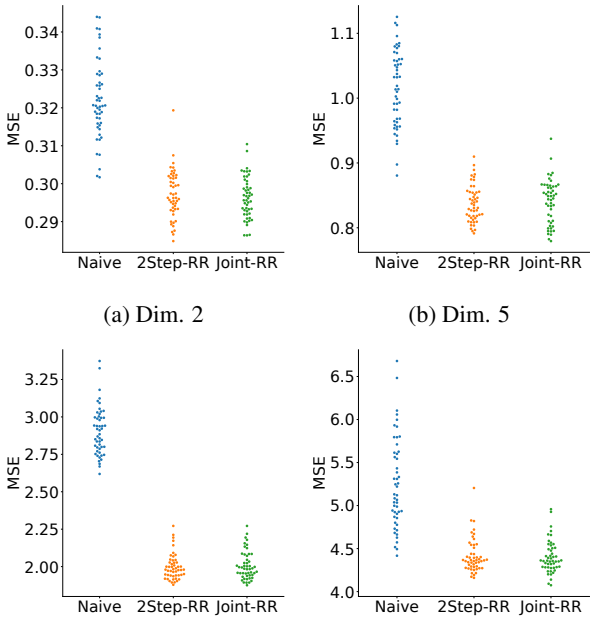
Figure 1. Results for the synthetic data experiments. The plots show the average MSEs and the shaded areas show the standard errors. Note that the standard errors are so small that the shaded area are almost unnoticable.

of the implementation provided by PyTorch (Paszke et al., 2019) for all the parameters of Adam: the learning rate is 0.001, and β is $(0.9, 0.999)$. We prepare mediated uncoupled data (defined in Section 2) for training but ordinary coupled (X, Y)-data for test evaluation. The task here is regression, and we use the MSE as the evaluation metric.

For our synthetic data, we can surely confirm whether the conditional mean independence of Eq. (1) holds or not. We will test the methods with varying dimensionality in both cases in which Eq. (1) is satisfied and violated. For the setting satisfying the condition, we define the data distribution as follows. X is distributed uniformly over $[-1, 1]^d$. $U_j := X_j^3 + \varepsilon_u$, where U_j is the j-th element of U, X_j is the j-th element of X, and ε_u is a uniform noise over $[-0.5, 0.5]^d$. $Y := \|U\|^2 + \varepsilon_y$, where ε_y is a Gaussian noise with mean zero and variance 0.1. This satisfies the condition because Y and X are independent after conditioning on U. For the setting violating the condition, X and U are the same as in the case with the condition satisfied, but Y depends on X rather than U: $Y := \|X\|^2 + \varepsilon_y$, where ε_y is a Gaussian noise with mean zero and variance 0.1. This violates the condition because U lacks some information that X has in predicting Y due to the noise ε_u. In both cases, we use $1{,}000 \times 2$ mediated uncoupled data for training and $10{,}000$ coupled (X, Y)-data for test evaluation.

Results are summarized in Figure 1; Figure 1(a) and 1(b) are for the settings satisfying and violating Eq. (1), respectively. The plots show that the proposed methods outperform the naive method. 2Step-RR and Joint-RR gave similar performances and their plots are almost indistinguishable in the figures.

Figure 2 shows more detailed results for each configuration of data dimensionality, showing that the proposed methods gave consistently lower MSEs than the naive method. Notably, Joint-RR tends to be more stable than 2Step-RR in the sense that the deviation of the MSE is smaller.



(a) Dim. 2

(b) Dim. 5

(c) Dim. 10

(d) Dim. 20

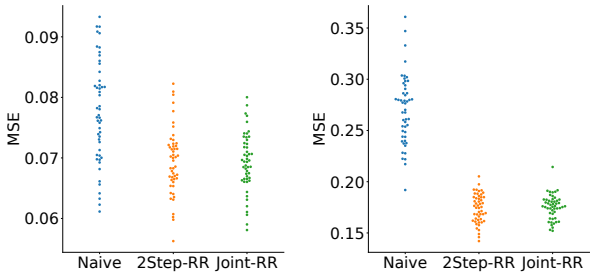
Figure 2. Experiments on synthetic data under the setting satisfying the conditional mean independence (Eq. (1)).

6.2. Classification of Low-quality Images

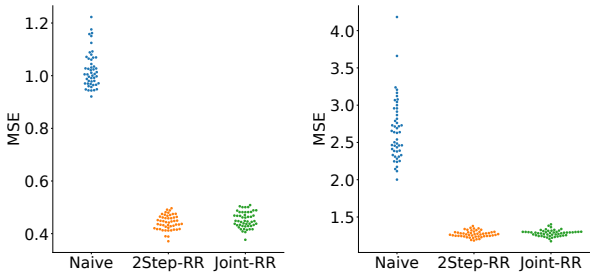
In this section, we test our methods in a more realistic scenario using image benchmark datasets, MNIST (LeCun et al., 1994), Fashion-MNIST (Xiao et al., 2017), CIFAR-10, and CIFAR-100 (Krizhevsky et al., 2009). The task here is to train a model that classifies *low-quality* images using mediated uncoupled data generated from those benchmark data, where the mediating variable is high-quality images.

The motivation behind the setup is that in some applications such as on-board electronics and the Internet of Things (Madakam et al., 2015), the prediction is often done by uploading low quality images to a remote server to fit limited network bandwidths. However, low-quality images can be difficult for human labelers to accurately label in the phase of training data collection. Moreover, the quality of images may depend on the device, the network, and the required response time. Instead of directly labeling low-quality images for different cases each time, we may prepare labeled high-quality images in advance only once or reuse existing data of this kind and collect pairs of high- and low-quality images in an ad hoc manner to adapt the data to each specific case.

In our experiment, we created low-quality images by downsampling images of the benchmark datasets with average pooling with stride $(2, 2)$ and kernel $(2, 2)$, and we took

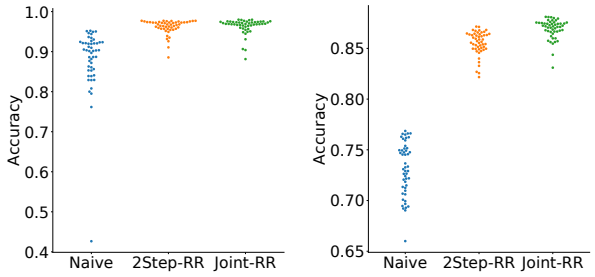


(a) Dim. 2 (b) Dim. 5

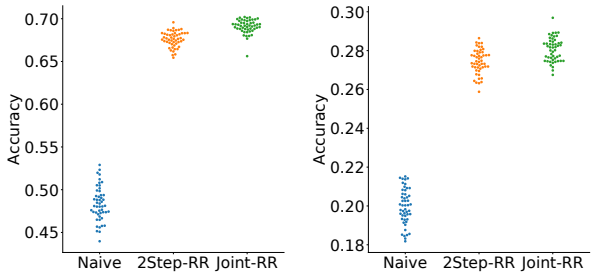


(c) Dim. 10 (d) Dim. 20

Figure 3. Experiments on synthetic data under the setting violating the conditional mean independence (Eq. (1)).



(a) MNIST. (b) Fashion-MNIST.



(c) CIFAR-10. (d) CIFAR-100.

Figure 4. Accuracy rates for the experiments on low-quality image classification.

other original images as high-quality images, i.e., X is a down-sampled image, U is an image of the original resolution, and Y is a class label. Let $K \in \mathbb{N}$ be the number of class labels. We use the one-hot representation for class labels, i.e., Y is K-dimensional vector with all elements being zero except for the dimension corresponding to the represented class. In order to make models f and h output K-dimensional probability vectors, we use the following "square-softmax" function for the last layers of f and h:

$$\mathbb{R}^K \ni (a_1, \dots, a_K) \mapsto \frac{a_y^2}{\sum_{y' \in [K]} a_{y'}^2} \in \mathbb{R}^K,$$

where $a_1, \dots, a_K \in \mathbb{R}$ are the outputs of the second last layer. It is similar to the standard softmax function, but it uses the square function instead of the exponential function.[*4] We use this architecture because the proposed methods are based on the squared loss whereas the softmax function typically uses cross-entropy loss. We evaluate MSEs and the objective functions with the squared ℓ_2-norms $\|f(x) - y\|^2$, $\|f(x) - h(u)\|^2$, and $\|h(u) - y\|^2$ for any $x \in \mathcal{X}$, $f \colon \mathcal{X} \to \mathbb{R}^K$, $h \colon \mathcal{U} \to \mathbb{R}^K$, and $y \in \mathbb{R}^K$, in place of their one-dimensional versions proposed in Section 2.

We train models using mediated uncoupled data consisting of samples of (X, U) and (U, Y). In the test evaluation phase, we use coupled (X, Y)-data, and we let the trained models to classify each low-quality image and compare the prediction with the true class label with the zero-one loss.

As in the synthetic experiments, we use the three methods but with the following configurations.

- For the naive method, we use a U-Net (Ronneberger et al., 2015) for predicting U from X and a ResNet (He et al., 2016) implemented by Idelbayev (2020) for predicting Y from U. These are considered to be state-of-the-art deep neural network architectures for image-to-image translation (X to U) and image classification (U to Y), respectively.
- For 2Step-RR, we use ResNets. Note that both predicting Y from U and Y from X are image classification.
- For Joint-RR, we again use ResNets since the trained models essentially are image classifiers. We set $w = 0.5$.

We train all models with Adam (Kingma & Ba, 2017) for 200 epochs. We turn off the weight decay and set the other tuning parameters of Adam as in PyTorch (Paszke et al., 2019): the learning rate is 0.001, the β is $(0.9, 0.999)$. We use randomly sampled $10{,}000 \times 2$ mediated uncoupled data for training and $10{,}000$ coupled (X, Y)-data for test evaluation. We repeat the experiment for 50 times.

[*4] In our experiments, we found that applying the square-softmax function makes training much easier than the standard softmax function.

Table 1. Accuracy rates and MSEs for the experiment on classification of low-quality images with the image benchmark datasets. The numbers outside of parentheses are means, and those in parentheses are standard errors calculated from 50 repetitions of the experiments. The scores comparable to the best in terms of Wilcoxon's signed rank test are emphasized in bold fonts.

	Accuracy			MSE		
Dataset	Naive	2Step-RR	Joint-RR	Naive	2Step-RR	Joint-RR
MNIST	88.06% (1.13)	**96.19**% (0.26)	**96.34**% (0.28)	0.184 (0.018)	**0.059** (0.003)	**0.056** (0.003)
Fashion-MNIST	73.12% (0.37)	85.53% (0.16)	**86.93**% (0.14)	0.417 (0.005)	0.213 (0.002)	**0.194** (0.002)
CIFAR-10	48.35% (0.28)	67.60% (0.13)	**69.10**% (0.11)	0.778 (0.005)	0.444 (0.002)	**0.424** (0.001)
CIFAR-100	19.97% (0.12)	27.43% (0.08)	**28.05**% (0.08)	0.935 (0.001)	0.850 (0.000)	**0.846** (0.000)

Table 1 shows the averages and the standard errors of the accuracy rates and the MSEs obtained by each method. In terms of accuracy, the two proposed methods outperformed the naive method. Joint-RR improved the accuracy compared to 2Step-RR for all datasets but MNIST. We can see the same tendency for the MSEs. Note that these accuracy rates are far from those of state-of-the-art supervised methods (e.g., Kowsari et al. (2018); Foret et al. (2021); Tanveer et al. (2021)) not only due to the uncoupled setting but also due to the down-sampling that significantly reduces the amount of information contained in images.

Figure 4 shows more details of the results with scatter plots of the accuracy rates. The figure indicates that the accuracy rates for the proposed methods and the naive method are clearly isolated except for MNIST, meaning that the proposed ones consistently performed better than the naive one for those datasets. Similar results are observed for the MSEs (see Figure 5 in Appendix K in the supplementary material). We can also see that the performance of the naive method highly deviates over the trials while those of the proposed methods tend to be more concentrated and show stable performances. 2Step-RR and Joint-RR gave comparable performance with each other, but Joint-RR showed slightly better performance. This performance gain may come from the regularization effect of Joint-RR (see Section 5.3).

7. Conclusion

In this paper, we considered learning from mediated uncoupled data. We proposed a method that learns a function directly predicts the target variable. We showed an excess error bound for the proposed method and demonstrated its practical usefulness through experiments. This paper focused on the squared loss, which is a standard choice for regression problems but not necessarily popular for classification. In future work, we investigate other loss functions for classification such as the logistic loss.

Acknowledgements

We thank Han Bao, Naoto Yokoya, Nontawat Charoenphakdee, Takashi Ishida, Yann Chevaleyre, and Yivan Zhang for the valuable discussions. IY and MS were supported by JST CREST Grant Number JPMJCR18A2. JH was supported by KAKENHI 21K11747. IY and FY acknowledge the support of the ANR as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

Angluin, D. and Laird, P. Learning from noisy examples. *Machine Learning*, 2(4):343–370, April 1988.

Blanchard, G., Flaska, M., Handy, G., Pozzi, S., and Scott, C. Classification with Asymmetric Label Noise: Consistency and Maximal Denoising. *arXiv:1303.1208 [cs, stat]*, August 2016. arXiv: 1303.1208.

Chapelle, O., Schölkopf, B., and Zien, A. (eds.). *Semi-supervised learning*. Adaptive computation and machine learning. MIT Press, 2006.

Dhir, A. and Lee, C. M. Integrating Overlapping Datasets Using Bivariate Causal Discovery. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pp. 3781–3790, 2020.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5767–5777. Curran Associates, Inc., 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Hein, M., Andriushchenko, M., and Bitterwolf, J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41–50, June 2019.

Idelbayev, Y. Proper ResNet implementation for CIFAR10/CIFAR100 in PyTorch. https://github.com/akamaster/pytorch_resnet_cifar10, 2020. Accessed: 2020-7-16.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017. arXiv: 1412.6980.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4743–4751. Curran Associates, Inc., 2016.

Kowsari, K., Heidarysafa, M., Brown, D. E., Meimandi, K. J., and Barnes, L. E. Rmdl: Random multimodel deep learning for classification. In *Proceedings of the 2nd International Conference on Information System and Data Mining*, ICISDM '18, pp. 19–28, New York, NY, USA, 2018. Association for Computing Machinery.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

LeCun, Y., Cortes, C., and Burges, C. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1994.

Ledoux, M. and Talagrand, M. *Probability in Banach spaces: isoperimetry and processes*. Classics in mathematics. Springer, Berlin ; London, 2011.

Madakam, S., Ramaswamy, R., and Tripathi, S. Internet of Things (IoT): A Literature Review. *Journal of Computer and Communications*, 03(05):164–173, 2015.

McDiarmid, C. *On the method of bounded differences*, pp. 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

Medhat, W., Hassan, A., and Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, December 2014.

Mittal, N., Sharma, D., and Joshi, M. L. Image sentiment analysis using deep learning. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 684–687, 2018.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, 2012.

Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. MIT Press, 2012.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with Noisy Labels. In *Advances in Neural Information Processing Systems 26*, pp. 1196–1204. Curran Associates, Inc., 2013.

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 *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer International Publishing, 2015.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2234–2242. Curran Associates, Inc., 2016.

Sasaki, H., Hyvärinen, A., and Sugiyama, M. Clustering via mode seeking by direct estimation of the gradient of a log-density. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2014)*, pp. 19–34, 2014.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Song, L., Gretton, A., and Guestrin, C. Nonparametric tree graphical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 765–772, Chia Laguna Resort, Sardinia, Italy, 2010. PMLR.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Sugiyama, M., Kanamori, T., Suzuki, T., Plessis, M. C. d., Liu, S., and Takeuchi, I. Density-Difference Estimation. *Neural Computation*, 25(10):2734–2775, October 2013.

Tanveer, M. S., Karim Khan, M. U., and Kyung, C.-M. Fine-tuning DARTS for image classification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 4789–4796, 2021.

van der, V. A. W. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

van Engelen, J. E. and Hoos, H. H. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, February 2020.

Vapnik, V. N. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2048–2057. PMLR, 07–09 Jul 2015.

Yamane, I., Yger, F., Atif, J., and Sugiyama, M. Uplift modeling from separate labels. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9927–9937. Curran Associates, Inc., 2018.

Zhang, Y., Charoenphakdee, N., and Sugiyama, M. Learning from indirect observations. *arXiv preprint arXiv:1910.04394*, 2019.

Zhu, X. J. Semi-supervised learning literature survey. Technical Report 1530, University of Wisconsin-Madison Department of Computer Sciences, 2005.

Appendix

A. Proof of Theorem 4.1

We prove the following theorem.

Theorem 4.1. The MSE can be bounded as

$$\mathbf{E}[(f(X)-Y)^2] \leq J_w(f,h), \tag{13}$$

where

$$J_w(f,h) \coloneqq \frac{1}{w}\mathbf{E}[(f(X)-h(U))^2] + \frac{1}{1-w}\mathbf{E}[(h(U)-Y)^2]$$

for any $w \in (0,1)$ and any $h \in L_U^2 \coloneqq \{h\colon \mathcal{U} \to \mathcal{Y} \mid \mathbf{E}[h(U)^2] < \infty\}$.

Proof. First, from Jensen's bound, we have

$$\begin{aligned}(a+b)^2 &= \left(w\frac{a}{w} + (1-w)\frac{a}{1-w}\right)^2 \\ &= w\left(\frac{a}{w}\right)^2 + (1-w)\left(\frac{a}{1-w}\right)^2 \\ &= \frac{a^2}{w} + \frac{a^2}{1-w},\end{aligned}$$

for any $a \in \mathbb{R}$, $b \in \mathbb{R}$, and $w \in (0,1)$. Using the inequality, we obtain

$$\begin{aligned}(f(X)-Y)^2 &= (f(X)-h(U)+h(U)-Y)^2 \\ &\leq \frac{(f(X)-h(U))^2}{w} + \frac{(h(U)-Y)^2}{1-w}.\end{aligned}$$

Taking the expectations of both sides, we complete the proof. □

B. Le Cam's Method

We suppose all involved probability distributions have density functions. For any density function $p(x,u,y)$ over $\mathcal{X}\times\mathcal{U}\times\mathcal{Y}$, we denote its marginal distributions by $p(x)$, $p(u)$, $p(y)$, $p(x,u)$, $p(x,y)$, and $p(u,y)$ and conditional distributions by $p(u \mid x)$, $p(y \mid x)$, $p(y \mid u)$, $p(y \mid u,x)$, and so on following the usual convention.

Definition 7.1. Fix any probability density function $p(x)$ over $\mathcal{X}$. For any $c \in [0,\infty)$, let $\mathbb{P}_{p,c}$ denote the set of density functions over $\mathcal{X}\times\mathcal{U}\times\mathcal{Y}$ defined as follows. Any density function $q(x,u,y)$ over $\mathcal{X}\times\mathcal{U}\times\mathcal{Y}$ is a member of $\mathbb{P}_{p,c}$ if and only if for $(X,U,Y) \sim q(x,u,y)$,

- X follows $p(x)$, and
- the difference between $\mathbf{E}[Y \mid X]$ and $\mathbf{E}[\mathbf{E}[Y \mid U] \mid X]$ is bounded by c from above in the sense of $L^2(p)$-distance:

$$\mathbf{E}[(\mathbf{E}[Y \mid X] - \mathbf{E}[\mathbf{E}[Y \mid U] \mid X])^2] \leq c^2.$$

Definition 7.2. For an underlying density function $p^*(x,u,y) \in \mathbb{P}_{p,c}$, *separate samples with proxy* is a set of random variables of the form $S_{n,\widetilde{n}} \coloneqq ((X_1,U_1),\dots,(X_n,U_n),(\widetilde{U}_1,\widetilde{Y}_1),\dots,(\widetilde{U}_{\widetilde{n}},\widetilde{Y}_{\widetilde{n}}))$, where $(X_1,U_1),\dots,(X_n,U_n),(\widetilde{U}_1,\widetilde{Y}_1),\dots,(\widetilde{U}_{\widetilde{n}},\widetilde{Y}_{\widetilde{n}})$ are independent, $(X_i,U_i) \sim p^*(x,u)$, and $(\widetilde{U}_i,\widetilde{Y}_i) \sim p^*(u,y)$. We denote the set of all possible realizations of $S_{n,\widetilde{n}}$ by $\mathcal{S}_{n,\widetilde{n}}$ and the density function of $S_{n,\widetilde{n}}$ by $p^*_{n,\widetilde{n}}(s)$, where

$$\begin{aligned}s \equiv{}& ((x_1,u_1),\dots,(x_n,u_n),(\widetilde{x}_1,\widetilde{u}_1),\dots,(\widetilde{x}_{\widetilde{n}},\widetilde{u}_{\widetilde{n}})) \\ &\in (\mathcal{X}\times\mathcal{U})^{n+\widetilde{n}}.\end{aligned}$$

In this section, we will obtain a lower bound of the expected error that the best learner has to suffer for the worst-case instance of $p^*(x,u,y) \in \mathbb{P}_{p,c}$ for a fixed $c \in [0,\infty)$ and a density function $p(x)$ over $\mathcal{X}$:

$$\begin{aligned}&E_{\text{minimax}} \\ &\coloneqq \inf_{\widehat{f}_{(\cdot)}\colon \mathcal{S}_{n,\widetilde{n}} \to \{f:\mathcal{X}\to\mathcal{Y}\}} \sup_{p^* \in \mathbb{P}_{p,c}} \mathbf{E}[(\widehat{f}_{S_{n,\widetilde{n}}}(X) - \mathbf{E}[Y \mid X])^2],\end{aligned}$$

where $(X,Y) \sim p^*(x,y)$, and the expectation is taken over $S_{n,\widetilde{n}}$ and (X,Y). $\widehat{f}_{(\cdot)}$ represents a learning algorithm ranging over all mappings that input $S_{n,\widetilde{n}}$ and output a function from $\mathcal{X}$ to $\mathcal{Y}$, which include computationally intractable ones.

Definition 7.3. Define a semi-distance metric on $\mathbb{P}_{p,c}$, $\rho\colon \mathbb{P}_{p,c}^2 \to [0,\infty)$, by

$$\begin{aligned}&\rho(q_1(x,u,y),q_2(x,u,y)) \\ &\coloneqq \mathbf{E}[(\mathbf{E}[Y_1 \mid X] - \mathbf{E}[\mathbf{E}[Y_2 \mid U_2] \mid X])^2]\end{aligned}$$

for any $(q_1(x,u,y),q_2(x,u,y)) \in \mathbb{P}_{p,c}^2$, where $(X,U_1,Y_1) \sim q_1(x,u,y)$ and $(X,U_2,Y_2) \sim q_2(x,u,y)$. Note that these two tuples share the common variable X in the definition.

Take any 2δ-separated density functions, $(p_1,p_2) \in \mathbb{P}_{p,c}^2$, in terms of ρ: $\rho(p_1,p_2) > 2\delta$. Le Cam's method states that

$$\begin{aligned}E_{\text{minimax}} &\geq \frac{1}{2}\delta^2\left(1 - \mathrm{TV}(p_1,p_2)\right) \\ &\geq \frac{1}{2}\delta^2\left(1 - \sqrt{\frac{1}{2}\mathrm{KL}(p_1,p_2)}\right),\end{aligned}$$

where $\mathrm{TV}(\cdot,\cdot)$ is the total variation distance, and $\mathrm{KL}(\cdot,\cdot)$ is the Kullback-Leibler divergence.

C. Proof of Theorem 5.1

The goal of this subsection is to show the following theorem.

Theorem 5.1. Suppose that $h^*(u) \coloneqq \mathbf{E}[Y \mid U=u] \in \mathcal{H}$ and $f^*(x) \coloneqq \mathbf{E}[h^*(U) \mid X=x] \in \mathcal{F}$ for some $\mathcal{F} \subseteq L_X^2$

and $\mathcal{H} \subseteq L_U^2$. Then,

$$(f^*, h^*) \in \lim_{w \uparrow 1} \operatorname*{arg\,min}_{(f,h) \in \mathcal{F} \times \mathcal{H}} J_w(f, h),$$

where $C > 0$ is a constant that does not depend on f or h.

Note that (f^*, h^*) is the solution pair to the optimization problem of 2Step-RR with the population-level objective functionals:

$$h^* \in \operatorname*{arg\,min}_{h \in \mathcal{H}} \mathbf{E}[(h(U) - Y)^2]$$
$$\text{and} \quad f^* \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \mathbf{E}[(f(X) - h^*(U))^2].$$

On the other hand, the constant in Eq. (11) is subtracted merely to prevent the objective value from diverging and make the solution well-defined in the limit. Thus, the theorem states that (f^*, h^*) is the limit of the solution pair to the optimization problem of the proposed method with the population-level objective functional.

Definition 7.4. For $w \in (0, 1)$, $f \in L_X^2$, and $h \in L_U^2$, define

$$Q(w, f, h) \coloneqq J_w(f, h) - \frac{1}{1-w} \mathbf{E}[(Y - h^*(U))^2],$$
$$R(w, f) \coloneqq \inf_{h \in \mathcal{H}} Q(w, f, h),$$
$$R(1, f) \coloneqq \mathbf{E}[(f(X) - h^*(U))^2].$$

Recall that

$$J_w(f, h) \equiv \frac{1}{w} \mathbf{E}[(f(X) - h(U))^2] + \frac{1}{1-w} \mathbf{E}[(h(U) - Y)^2],$$

Using the identity

$$\begin{aligned}
&\mathbf{E}[(h(U) - Y)^2] \\
&= \mathbf{E}[(h(U) - \mathbf{E}[Y \mid X])^2] + \mathbf{E}[(\mathbf{E}[Y \mid X] - Y)^2] \\
&= \mathbf{E}[(h(U) - h^*(U))^2] + \mathbf{E}[(h^*(U) - Y)^2],
\end{aligned}$$

we obtain

$$\begin{aligned}
&Q(w, f, h) \\
&= \frac{1}{w} \mathbf{E}[(f(X) - h(U))^2] + \frac{1}{1-w} \mathbf{E}[(h(U) - h^*(U))^2] \\
&\quad + \frac{1}{1-w} \mathbf{E}[(h^*(U) - Y)^2] - \frac{1}{1-w} \mathbf{E}[(h^*(U) - Y)^2] \\
&= \frac{1}{w} \mathbf{E}[(f(X) - h(U))^2] + \frac{1}{1-w} \mathbf{E}[(h(U) - h^*(U))^2].
\end{aligned}$$

Moreover, note that

$$f^* = \operatorname*{arg\,min}_{f \in \mathcal{F}} R(1, f).$$

Lemma 7.1. *$(w, f) \mapsto R(w, f)$ is continuous on $\{1\} \times \mathcal{F}$, and $w \mapsto \inf_{f \in \mathcal{F}} R(w, f)$ is continuous at $w = 1$.*

Proof. For any $f_0, f_1 \in \mathcal{F}$ and any $w \in (0, 1)$, we have

$$\begin{aligned}
&R(w, f_1) - R(1, f_0) \\
&= \inf_{h \in \mathcal{H}} \left[\frac{1}{w} \mathbf{E}[(f_1(X) - h(U))^2] \right. \\
&\qquad \left. + \frac{1}{1-w} \mathbf{E}[(h(U) - h^*(U))^2] \right] \\
&\quad - \mathbf{E}[(f_0(X) - h^*(U))^2] \\
&\le \frac{1}{w} \mathbf{E}[(f_1(X) - h^*(U))^2] - \mathbf{E}[(f_0(X) - h^*(U))^2].
\end{aligned}$$

Since $h = h^*$ cannot go below the infimum, we have

$$\begin{aligned}
&R(w, f_1) - R(1, f_0) \\
&\le \mathbf{E}[(f_1(X) - h^*(U))^2] - \mathbf{E}[(f_0(X) - h^*(U))^2] \\
&\quad + \frac{1-w}{w} \mathbf{E}[(f_1(X) - h^*(U))^2] \\
&= \mathbf{E}[(f_1(X) - f_0(X))(f_1(X) + f_0(X) - 2h^*(U))] \\
&\quad + \frac{1-w}{w} \mathbf{E}[(f_1(X) - h^*(U))^2] \\
&= \mathbf{E}[(f_1(X) - f_0(X))(f_1(X) - f_0(X) + 2f_0(X) - 2h^*(U))] \\
&\quad + \frac{1-w}{w} \mathbf{E}[3^2(\{f_1(X) - f_0(X)\}/3 + f_0(X)/3 - h^*(U)/3)^2] \\
&\le \|f_1 - f_0\|_2 (\|f_1 - f_0\|_2 + 2\|f_0\|_2 + 2\|h^*\|_2) \\
&\quad + \frac{3(1-w)}{w} (\|f_1 - f_0\|_2^2 + \|f_0\|_2^2 + \|h^*\|_2^2) \\
&\qquad \text{(from the CauchySchwarz and Jensen's inequality)} \\
&\to 0 \quad \text{(as } w \uparrow 1 \text{ and } f_1 \to f_0\text{)}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&R(w, f_1) - R(1, f_0) \\
&= \inf_{h \in \mathcal{H}} \left[\frac{w}{w^2} \mathbf{E}[(f_1(X) - h(U))^2] \right. \\
&\qquad \left. + \frac{1-w}{(1-w)^2} \mathbf{E}[(h(U) - h^*(U))^2] \right] \\
&\quad - \mathbf{E}[(f_0(X) - h^*(U))^2] \\
&\ge \inf_{h \in \mathcal{H}} \left[\mathbf{E} \left[\left(w \cdot \frac{f_1(X) - h(U)}{w} \right. \right. \right. \\
&\qquad \left. \left. \left. + (1-w) \cdot \frac{h(U) - h^*(U)}{1-w} \right)^2 \right] \right] \\
&\quad - \mathbf{E}[(f_0(X) - h^*(U))^2]
\end{aligned}$$

from Jensen's inequality. Hence,

$$
\begin{aligned}
&R(w, f_1) - R(1, f_0) \\
&= \mathbf{E}[(f_1(X) - h^*(U))^2] - \mathbf{E}[(f_0(X) - h^*(U))^2] \\
&= \mathbf{E}[(f_1(X) - f_0(X))(f_1(X) + f_0(X) - 2h^*(U))] \\
&\geq -\|f_1 - f_0\|(\|f_1 - f_0\| + 2\|f_0\| + 2\|h^*\|) \\
&\quad \text{(from the CauchySchwarz inequality)} \\
&\to 0 \quad (\text{as } f_1 \to f_0).
\end{aligned}
$$

By the squeeze theorem, $R(w, f_1) \to R(1, f_0)$ as $w \uparrow 1$ and $f_1 \to f_0$.

Similarly, for any $w \in (0, 1)$, we have

$$
\begin{aligned}
&\inf_{f \in \mathcal{F}} R(w, f) - \inf_{f \in \mathcal{F}} R(1, f) \\
&= \inf_{f \in \mathcal{F}} R(w, f) - R(1, f^*) \\
&= \inf_{(f,h) \in \mathcal{F} \times \mathcal{H}} \left[\frac{1}{w} \mathbf{E}[(f(X) - h(U))^2] \right. \\
&\qquad \left. + \frac{1}{1-w} \mathbf{E}[(h(U) - h^*(U))^2] \right] \\
&\quad - \mathbf{E}[(f^*(X) - h^*(U))^2] \\
&\leq \frac{1}{w} \mathbf{E}[(f^*(X) - h^*(U))^2] - \mathbf{E}[(f^*(X) - h^*(U))^2] \\
&\quad \text{(since } (f, h) = (f^*, h^*) \text{ cannot go below the infimum),} \\
&\leq \left(\frac{1}{w} - 1 \right) \mathbf{E}[(f^*(X) - h^*(U))^2] \\
&\to 0 \quad (\text{as } w \uparrow 1).
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
&\inf_{f \in \mathcal{F}} R(w, f) - \inf_{f \in \mathcal{F}} R(1, f) \\
&\geq \inf_{f \in \mathcal{F}} R(w, f) - R(1, f^*) \\
&= \inf_{(f,h) \in \mathcal{F} \times \mathcal{H}} \left[\frac{w}{w^2} \mathbf{E}[(f(X) - h(U))^2] \right. \\
&\qquad \left. + \frac{1-w}{(1-w)^2} \mathbf{E}[(h(U) - h^*(U))^2] \right] \\
&\quad - \mathbf{E}[(f^*(X) - h^*(U))^2] \\
&\geq \inf_{(f,h) \in \mathcal{F} \times \mathcal{H}} \mathbf{E} \left[\Big(w \cdot \frac{f(X) - h(U)}{w} \right. \\
&\qquad \left. + (1-w) \cdot \frac{h(U) - h^*(U)}{1-w} \Big)^2 \right] \\
&\quad - \mathbf{E}[(f^*(X) - h^*(U))^2] \quad \text{(from Jensen's inequality)} \\
&\geq \inf_{(f,h) \in \mathcal{F} \times \mathcal{H}} \mathbf{E} \left[(f(X) - h^*(U))^2 \right] \\
&\qquad - \mathbf{E}[(f^*(X) - h^*(U))^2] \\
&= 0.
\end{aligned}
$$

By the squeeze theorem,

$$
\inf_{f \in \mathcal{F}} R(w, f) \to \inf_{f \in \mathcal{F}} R(1, f) \tag{14}
$$

as $w \uparrow 1$. □

Lemma 7.2. *$f \mapsto R(1, f)$ has a* well-separated *minimum (van der, 1998), i.e., there exists a minimizer*

$$
f^\dagger \in \arg\min_{f \in \mathcal{F}} R(1, f) \tag{15}
$$

that satisfies

$$
R(1, f^\dagger) < \inf_{f \in \mathcal{F}, \|f - f^\dagger\|_2 \geq \delta} R(1, f) \tag{16}
$$

for every $\delta > 0$.

The minimizer turns out to be unique when it is well-separated, so $f^\dagger = f^*$.

Proof. Let

$$
\begin{aligned}
f^\dagger &:= f^* = \mathbf{E}[h^*(U) \mid X = x] \\
&\in \arg\min_{f \in \mathcal{F}} \mathbf{E}[(f(X) - h^*(U))^2].
\end{aligned}
$$

For any $\delta > 0$ and any $f \in \mathcal{F}$ such that $\|f - f^\dagger\| \geq \delta$, we have

$$
\begin{aligned}
&R(1, f) - R(1, f^\dagger) \\
&= \mathbf{E}[(f(X) - h^*(U))^2] - \mathbf{E}[(f^\dagger(X) - h^*(U))^2] \\
&= \mathbf{E}[(f(X) - \mathbf{E}[h^*(U) \mid X])^2] \\
&\quad + \mathbf{E}[(h^*(U) - \mathbf{E}[h^*(U) \mid X])^2] \\
&\quad - \mathbf{E}[(\mathbf{E}[h^*(U) \mid X] - h^*(U))^2] \\
&= \mathbf{E}[(f(X) - f^\dagger(X))^2] \\
&\geq \delta^2.
\end{aligned}
$$

Hence, for any $\delta > 0$, it holds that

$$
\inf_{f \in \mathcal{F}, \|f - f^\dagger\| \geq \delta} R(1, f) \geq R(1, f^\dagger) + \delta^2 > R(1, f^\dagger). \tag{17}
$$

□

Proof of Theorem 5.1. Let

$$
h_{f,w} := \arg\min_{h \in \mathcal{H}} Q(w, f, h). \tag{18}
$$

First, we show that $h_{f,w} \to h^*$ as $w \uparrow 1$ for any $f \in \mathcal{F}$. Since $Q(w, f, h_{w,f}) \leq Q(w, f, h^*)$, we have

$$
\begin{aligned}
&\frac{1}{w} \mathbf{E}[(f(X) - h_{f,w}(U))^2] \\
&\quad + \frac{1}{1-w} \mathbf{E}[(h_{f,w}(U) - h^*(U))^2] \\
&\leq \frac{1}{w} \mathbf{E}[(f(X) - h^*(U))^2],
\end{aligned}
$$

which implies

$$\begin{aligned}&\frac{1}{1-w}\,\mathbf{E}[(h_{f,w}(U)-h^*(U))^2]\\&\le \frac{1}{w}\,\mathbf{E}[(f(X)-h^*(U))^2]-\frac{1}{w}\,\mathbf{E}[(f(X)-h_{f,w}(U))^2]\\&\le \frac{1}{w}\,\mathbf{E}[(f(X)-h^*(U))^2].\end{aligned}$$

Thus,

$$\begin{aligned}&\|h_{f,w}-h^*\|_2^2 \equiv \mathbf{E}[(h_{f,w}(U)-h^*(U))^2]\\&\le \frac{1-w}{w}\,\mathbf{E}[(f(X)-h^*(U))^2]\to 0 \quad (\text{as } w\uparrow 1).\end{aligned}$$

Next, we show

$$f_w \coloneqq \arg\min_{f\in\mathcal{F}} R(w,f)\to f^* \text{ as } w\uparrow 1. \tag{19}$$

From the continuity of $(w,f)\mapsto R(w,f)$ and $w\mapsto \inf_{f\in\mathcal{F}} R(w,f)$ at $w=1$ (Lemma 7.1), for every $\varepsilon>0$, there exists $\delta>0$ such that for every $\widetilde{w}\in(0,1)$,

$$\begin{aligned}&|\widetilde{w}-1|<\delta\\&\implies |R(1,f_{\widetilde{w}})-R(\widetilde{w},f_{\widetilde{w}})|<\varepsilon/2\\&\quad \text{and } |\inf_{f\in\mathcal{F}} R(\widetilde{w},f)-\inf_{f\in\mathcal{F}} R(1,f)|<\varepsilon/2\\&\implies |R(1,f_{\widetilde{w}})-R(1,f^*)|\\&\quad <|R(1,f_{\widetilde{w}})-R(\widetilde{w},f_{\widetilde{w}})|+|R(\widetilde{w},f_w)-R(1,f^*)|\\&\quad <|R(1,f_{\widetilde{w}})-R(\widetilde{w},f_{\widetilde{w}})|\\&\quad +|\inf_{f\in\mathcal{F}} R(\widetilde{w},f)-\inf_{f\in\mathcal{F}} R(1,f)|\\&\quad <\varepsilon\\&\implies R(1,f_{\widetilde{w}})-R(1,f^*)<\varepsilon.\end{aligned}$$

On the other hand, suppose that

$$\exists\eta>0,\forall\varepsilon>0,\exists f\in\mathcal{F}, \tag{20}$$
$$[\|f-f^*\|_2\ge\eta \text{ and } R(1,f)-R(1,f^*)<\varepsilon]. \tag{21}$$

Then,

$$\exists\eta>0,\forall\varepsilon>0,\inf_{f\in\mathcal{F},\|f-f^*\|\ge\eta} R(1,f)<R(1,f^*)+\varepsilon;$$

hence

$$\exists\eta>0,\inf_{f\in\mathcal{F},\|f-f^*\|_2\ge\eta} R(1,f)=R(1,f^*),$$

which contradicts the fact that f^* is well-separated as a minimizer of $f\mapsto R(1,f)$. This confirms that the negation of (21) holds:

$$\forall\eta>0,\exists\varepsilon>0,\forall f\in\mathcal{F},[R(1,f)-R(1,f^*)<\varepsilon \tag{22}$$
$$\implies \|f-f^*\|_2<\eta]. \tag{23}$$

Combining (20) and (23), for every $\eta>0$, there exist $\varepsilon>0$ and $\delta>0$ such that for every $\widetilde{w}\in(0,1)$,

$$\begin{aligned}\|\widetilde{w}-1\|<\delta &\implies R(1,f_{\widetilde{w}})-R(1,f^*)<\varepsilon\\&\implies \|\widetilde{f}-f^*\|_2<\eta,\end{aligned}$$

which implies that $w\mapsto \arg\min_{f\in\mathcal{F}} R(w,f)$ is continuous at $w=1$.

Combining the results, we conclude that

$$(f^*,h^*)\in\lim_{w\uparrow 1}\arg\min_{f\in\mathcal{F},h\in\mathcal{H}} Q(w,f,h).$$

□

D. Rademacher complexity

Definition 7.5 (Rademacher Complexity)**.** For any set of functions H and any probability density function p over the domain of functions of H, we define the *Rademacher complexity* of H under p as

$$\mathfrak{R}_p^N(H)=\mathop{\mathbf{E}}_{v_1,\dots,v_N,\sigma_1,\dots,\sigma_N}\left[\sup_{h\in H}\frac{1}{N}\sum_{i=1}^N\sigma_i h(v_i)\right],$$

where $v_1,\dots,v_N\sim p$, $\sigma_1,\dots,\sigma_N$ are $\{-1,1\}$-valued uniform random variables, and they are all independent.

E. McDiarmid's Inequality

To derive a uniform deviation bound of our empirical process, we use the following theorem called McDiarmid's inequality.

Theorem 7.1 (McDiarmid's inequality)**.** *Let $\varphi:\mathcal{D}^N\to\mathbb{R}$ be a measurable function. Assume that there exists a real number $B_\varphi>0$ such that*

$$|\varphi(v_1,\dots,v_N)-\varphi(v_1',\dots,v_N')|\le B_\varphi, \tag{24}$$

for any $v_i,\dots,v_N,v_1,\dots,v_N'\in\mathcal{D}$ where $v_i=v_i'$ for all but one $i\in\{1,\dots,N\}$. Then, for any $\mathcal{D}$-valued independent random variables $V_1,\dots,V_N$ and any $\delta>0$ the following holds with probability at least $1-\delta$:

$$\varphi(V_1,\dots,V_N)\le\mathbf{E}[\varphi(V_1,\dots,V_N)]+\sqrt{\frac{B_\varphi^2N}{2}\log\frac{1}{\delta}}.$$

F. Excess error bound for 2Step-RR

Proof. Let $\bar{g}(x)\coloneqq\mathbf{E}[\widetilde{h}(U)\mid X=x]$. Then,

$$\begin{aligned}&\mathbf{E}[(\widetilde{f}(X)-f^*(X))^2]\\&\le 2\,\mathbf{E}[(\widetilde{f}(X)-\bar{g}(X))^2]\\&\quad +2\,\mathbf{E}[(\bar{g}(X)-f^*(X))^2].\end{aligned}$$

We are going to bound each of the terms on the right hand side.

Bounding $\mathbf{E}[(\widetilde{f}(X)-\bar{g}(X))^2]$: First, we have

$$\begin{aligned}&\mathbf{E}[(\widetilde{f}(X)-\widetilde{h}(U))^2]\\&=\mathbf{E}[(\widetilde{f}(X)-\bar{g}(X))^2]\\&+\mathbf{E}[(\bar{g}(X)-\widetilde{h}(U))^2]\\&+2\,\mathbf{E}[(\widetilde{f}(X)-\bar{g}(X))\underbrace{\mathbf{E}[\bar{g}(X)-\widetilde{h}(U)\mid X]}_{=0}].\end{aligned}$$

Hence,

$$\begin{aligned}&\mathbf{E}[(\widetilde{f}(X)-\bar{g}(X))^2]\\&=\mathbf{E}[(\widetilde{f}(X)-\widetilde{h}(U))^2]-\mathbf{E}[(\bar{g}(X)-\widetilde{h}(U))^2].\end{aligned}$$

Observe that

$$\begin{aligned}&\mathbf{E}[(\widetilde{f}(X)-\bar{g}(X))^2]\\&=\mathbf{E}[(\widetilde{f}(X)-\widetilde{h}(U))^2]-\mathbf{E}[(\bar{g}(X)-\widetilde{h}(U))^2]\\&=\mathbf{E}[(\widetilde{f}(X)-\widetilde{h}(U))^2]-\frac{1}{n}\sum_{i=1}^{n}(\bar{g}(X_i)-\widetilde{h}(U_i))^2\\&\quad+\underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{f}(X_i)-\widetilde{h}(U_i))^2-\frac{1}{n}\sum_{i=1}^{n}(\bar{g}(X_i)-\widetilde{h}(U_i))^2}_{\leq 0\ (\text{from }\widetilde{f}\text{'s optimality})}\\&\quad+\frac{1}{n}\sum_{i=1}^{n}(\bar{g}(X_i)-\widetilde{h}(U_i))^2-\mathbf{E}[(\bar{g}(X)-\widetilde{h}(U))^2]\\&\leq 2\sup_{\phi\in\mathcal{F}-\mathcal{H}}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\phi(X_i,U_i)^2-\mathbf{E}[\phi(X,U)^2]\right|\right],\end{aligned}$$

where $\mathcal{F}-\mathcal{H}=\{(x,u)\mapsto f(x)+h(u)\mid f\in\mathcal{F},h\in\mathcal{H}\}$. Let

$$\begin{aligned}&\psi(x_1,u_1,\ldots,x_n,u_n)\\&\coloneqq\sup_{\phi\in\mathcal{F}-\mathcal{H}}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\phi(x_i,u_i)^2-\mathbf{E}[\phi(X,U)^2]\right|\right],\end{aligned}$$

Then, ψ is a function with bounded differences:

$$\begin{aligned}&|\psi(x_1,u_1,\ldots,x_j,u_j,\ldots,x_n,u_n)\\&\quad-\psi(x_1,u_1,\ldots,x'_j,u'_j,\ldots,x_n,u_n)|\\&\leq\sup_{\phi\in\mathcal{F}-\mathcal{H}}\left[\left|\phi(x_j,u_j)^2-\phi(x'_j,u'_j)^2\right|\right]\\&\leq 2(C_{\mathcal{F}}+C_{\mathcal{H}})^2/n.\end{aligned}$$

From McDiarmid's inequality for functions with bounded differences (McDiarmid, 1989), with probability at least $1-\delta$, it holds that

$$\begin{aligned}&\psi(X_1,U_1,\ldots,X_n,U_n)\\&\leq\mathbf{E}[\psi(X_1,U_1,\ldots,X_n,U_n)]\\&\quad+\sqrt{\frac{4(C_{\mathcal{F}}+C_{\mathcal{H}})^4}{2n}\log\frac{1}{\delta}}\\&=\mathbf{E}[\psi(X_1,U_1,\ldots,X_n,U_n)]\\&\quad+(C_{\mathcal{F}}+C_{\mathcal{H}})^2\sqrt{\frac{2}{n}\log\frac{1}{\delta}}.\end{aligned}$$

Here,

$$\begin{aligned}&\mathbf{E}[\psi(X_1,Y_1,\ldots,X_n,Y_n)\}_{i=1}^{n})]\\&\leq\mathbf{E}\left[\sup_{\phi\in\mathcal{F}-\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^{n}\phi(X_i,U_i)^2-\sum_{i=1}^{n}\phi(X'_i,U'_i)^2\right|\right]\\&\leq\mathbf{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\phi(X_i,U_i))^2\right|\right]\\&\leq\mathfrak{R}_n(\{\phi^2\mid\phi\in\mathcal{F}-\mathcal{H}\})\\&\leq 2(C_{\mathcal{F}}+C_{\mathcal{H}})(\mathfrak{R}_n(\mathcal{F})+\mathfrak{R}_n(\mathcal{H})).\end{aligned}$$

The last inequality follows from Talagrand's contraction lemma (Ledoux & Talagrand, 2011) Combining what we have obtained, we confirm that $\mathbf{E}[(\widetilde{f}(X)-\bar{g}(X))^2]$ can be controlled by the Rademacher complexities of $\mathcal{F}$ and $\mathcal{H}$:

$$\begin{aligned}&\mathbf{E}[(\widetilde{f}(X)-\bar{g}(X))^2]\\&\leq 2\psi(X_1,U_1,\ldots,X_n,U_n)\\&\leq 2\,\mathbf{E}[\psi(X_1,U_1,\ldots,X_n,U_n)]\\&\quad+2(C_{\mathcal{F}}+C_{\mathcal{H}})^2\sqrt{\frac{2}{n}\log\frac{1}{\delta}}\\&\leq 4(C_{\mathcal{F}}+C_{\mathcal{H}})(\mathfrak{R}_n(\mathcal{F})+\mathfrak{R}_n(\mathcal{H}))\\&\quad+2(C_{\mathcal{F}}+C_{\mathcal{H}})^2\sqrt{\frac{2}{n}\log\frac{1}{\delta}}.\end{aligned}$$

Bounding $\mathbf{E}[(\bar{g}(X)-f^*(X))^2]$: First, we are going to show that $\mathbf{E}[(\bar{g}(X)-f^*(X))^2]$ can be bounded in terms of $\mathbf{E}[(\widetilde{h}(U)-h^*(U))^2]$. From the optimality of $\bar{g}$, we have

$$\mathbf{E}[(\bar{g}(X)-\widetilde{h}(U))^2]\leq\mathbf{E}[(f^*(X)-\widetilde{h}(U))^2]. \tag{25}$$

By re-arranging equations, we get

$$\begin{aligned}&\mathbf{E}[(\bar{g}(X)-f^*(X))^2]\\&\leq 2\,\mathbf{E}[(\widetilde{h}(U)-h^*(U))(\bar{g}(X)-f^*(X))]\\&\leq 2\sqrt{\mathbf{E}[(\widetilde{h}(U)-h^*(U))^2]\,\mathbf{E}[(\bar{g}(X)-f^*(X))^2]},\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. This implies

$$\mathbf{E}[(\bar{g}(X)-f^*(X))^2]\leq 4\,\mathbf{E}[(\widetilde{h}(U)-h^*(U))^2]. \tag{26}$$

Next, we bound $\mathbf{E}[(\widetilde{h}(U) - h^*(U))^2]$ using a standard generalization error bound using a uniform deviation bound and the Rademacher complexity. Let

$$\begin{aligned}&\varphi(\{(u_i, y_i)\}_{i=1}^{n'}; \mathcal{H})\\ &:= \sup_{h\in\mathcal{H}} \left| \frac{1}{n'} \sum_{i=1}^{n'} (h(u_i) - y_i)^2 - \mathbf{E}[(h(U) - Y)^2] \right|.\end{aligned}$$

Let $\{(u_i, y_i)\}_{i=1}^{n'} \subseteq (\mathcal{U} \times \mathcal{Y})^{n'}$ and $\{(u'_i, y'_i)\}_{i=1}^{n'} \subseteq (\mathcal{U} \times \mathcal{Y})^{n'}$ be any two sets of size n' that differ from each their only by one pair of elements, $((u_\iota, y_\iota), (u'_\iota, y'_\iota))$. One can show that

$$\begin{aligned}&\left|\varphi(\{(u_i, y_i)\}_{i=1}^{n'}; \mathcal{H}) - \varphi(\{(u'_i, y'_i)\}_{i=1}^{n'}; \mathcal{H})\right|\\ &\leq \sup_{h\in\mathcal{H}} |(h(u_\iota) - y_\iota)^2 - (h(u'_\iota) - y'_\iota)^2|\\ &\leq 2(C_\mathcal{H} + C_\mathcal{Y})^2/n'.\end{aligned}$$

From McDiarmid's inequality for functions with bounded differences (McDiarmid, 1989) and the union bound, with probability at least $1 - \delta$, it holds that

$$\begin{aligned}&\varphi(\{(U'_i, Y'_i)\}_{i=1}^{n'}; \mathcal{H})\\ &\leq \mathbf{E}[\varphi(\{(U'_i, U'_i)\}_{i=1}^{n'}; \mathcal{H})]\\ &\quad + \sqrt{\frac{4(C_\mathcal{H} + C_\mathcal{Y})^4}{2n'} \log\frac{1}{\delta}}\\ &= \mathbf{E}[\varphi(\{(U'_i, U'_i)\}_{i=1}^{n'}; \mathcal{H})]\\ &\quad + (C_\mathcal{H} + C_\mathcal{Y})^2 \sqrt{\frac{2}{n'} \log\frac{1}{\delta}}.\end{aligned}$$

Here,

$$\begin{aligned}&\mathbf{E}[\varphi(\{(U'_i, U'_i)\}_{i=1}^{n'}; \mathcal{H})]\\ &\leq \mathbf{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n'} \left| \sum_{i=1}^{n'} (h(U'_i) - Y'_i)^2 - \sum_{i=1}^{n'} (h(U'_i) - Y'_i)^2 \right|\right]\\ &\leq \mathbf{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n'} \left| \sum_{i=1}^{n'} \sigma_i (h(U'_i) - Y'_i)^2 \right|\right]\\ &\leq \mathbf{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n'} \left| \sum_{i=1}^{n'} \sigma_i (h(U'_i) - Y'_i)^2 \right|\right]\\ &\leq \mathfrak{R}_{n'}(\{(u, y) \mapsto (h(u) - y)^2 \mid h \in \mathcal{H}\})\\ &\leq 2(C_\mathcal{H} + C_\mathcal{Y})\mathfrak{R}_{n'}(\mathcal{H}).\end{aligned}$$

The last inequality follows from Talagrand's contraction lemma (Ledoux & Talagrand, 2011). Thus, we have

$$\begin{aligned}&\varphi(\{(U'_i, Y'_i)\}_{i=1}^{n'}; \mathcal{H})\\ &\equiv \sup_{h\in\mathcal{H}} \left| \sum_{i=1}^{n'} (h(u_i) - y_i)^2 - \mathbf{E}[(h(U) - Y)^2] \right|\\ &\leq 2(C_\mathcal{H} + C_\mathcal{Y})\mathfrak{R}_{n'}(\mathcal{H}) + (C_\mathcal{H} + C_\mathcal{Y})^2 \sqrt{2n' \log\frac{1}{\delta}}.\end{aligned}$$

Also, because

$$\begin{aligned}&\mathbf{E}[(\widetilde{h}(U) - Y)^2]\\ &= \mathbf{E}[(\widetilde{h}(U) - h^*(U))^2] + \mathbf{E}[(h^*(U) - Y)^2],\end{aligned}$$

we get

$$\begin{aligned}&\mathbf{E}[(\widetilde{h}(U) - h^*(U))^2]\\ &= \mathbf{E}[(\widetilde{h}(U) - Y)^2] - \mathbf{E}[(h^*(U) - Y)^2].\end{aligned}$$

Using the results above, we have

$$\begin{aligned}&\mathbf{E}[(\widetilde{h}(U) - h^*(U))^2]\\ &= \mathbf{E}[(\widetilde{h}(U) - Y)^2] - \mathbf{E}[(h^*(U) - Y)^2]\\ &= \mathbf{E}[(\widetilde{h}(U) - Y)^2] - \frac{1}{n'} \sum_{i=1}^{n'} [(h(U'_i) - Y'_i)^2]\\ &+ \frac{1}{n'} \sum_{i=1}^{n'} [(\widetilde{h}(U'_i) - Y'_i)^2] - \frac{1}{n'} \sum_{i=1}^{n'} [(h^*(U'_i) - Y'_i)^2]\\ &+ \frac{1}{n'} \sum_{i=1}^{n'} [(h^*(U'_i) - Y'_i)^2] - \mathbf{E}[(h^*(U) - Y)^2]\\ &\leq 2(C_\mathcal{H} + C_\mathcal{Y})\mathfrak{R}_{n'}(\mathcal{H}) + (C_\mathcal{H} + C_\mathcal{Y})^2 \sqrt{\frac{2}{n'} \log\frac{1}{\delta}}\\ &+ 0\\ &+ 2(C_\mathcal{H} + C_\mathcal{Y})\mathfrak{R}_{n'}(\mathcal{H}) + (C_\mathcal{H} + C_\mathcal{Y})^2 \sqrt{\frac{2}{n'} \log\frac{1}{\delta}}\\ &\leq 4(C_\mathcal{H} + C_\mathcal{Y})\mathfrak{R}_{n'}(\mathcal{H}) + (C_\mathcal{H} + C_\mathcal{Y})^2 \sqrt{\frac{2}{n'} \log\frac{1}{\delta}}.\end{aligned}$$

Bounding $\mathbf{E}[(\widetilde{f}(X) - f^*(X))^2]$: Finally, we summarizing the results above to obtain

$$\begin{aligned}
&\mathbf{E}[(\widetilde{f}(X) - f^*(X))^2] \\
&\leq 2\,\mathbf{E}[(\widetilde{f}(X) - \bar{g}(X))^2] \\
&\quad + 2\,\mathbf{E}[(\bar{g}(X) - f^*(X))^2] \\
&\leq 8(C_{\mathcal{F}} + C_{\mathcal{H}})(\mathfrak{R}_n(\mathcal{F}) + \mathfrak{R}_n(\mathcal{H})) \\
&\quad + 4(C_{\mathcal{F}} + C_{\mathcal{H}})^2\sqrt{\frac{2}{n}\log\frac{1}{\delta}} \\
&\quad + 8(C_{\mathcal{H}} + C_{\mathcal{Y}})\mathfrak{R}_{n'}(\mathcal{H}) \\
&\quad + 2(C_{\mathcal{H}} + C_{\mathcal{Y}})^2\sqrt{\frac{2}{n'}\log\frac{1}{\delta}} \\
&\leq \mathcal{O}_p\Bigg(\mathfrak{R}_n(\mathcal{F}) + \mathfrak{R}_n(\mathcal{H}) + \mathfrak{R}_{n'}(\mathcal{H}) \\
&\qquad + \sqrt{(\frac{1}{n} + \frac{1}{n'})\log\frac{1}{\delta}}\Bigg).
\end{aligned}$$

□

F.1. Excess error bound for Joint-RR

Theorem 7.2. *Let*

$$V_w(f^*) \coloneqq \frac{1-w}{w}\,\mathbf{E}[(f^*(X) - \mathbf{E}[f^*(X) \mid U])^2],$$

and

$$f_w \coloneqq \operatorname*{arg\,min}_{f \in L^2_X} \mathbf{E}[(f(X) - f^*(X))^2] + \frac{1-w}{w}\,\mathbf{E}[(f(X) - \mathbf{E}[f(X) \mid U])^2].$$

Proof.

$$\begin{aligned}
&\mathbf{E}[(\widehat{f}_w(X) - f^*(X))^2] \\
&\leq \mathbf{E}[(\widehat{f}_w(X) - f^*(X))^2] \\
&+ \frac{w}{1-w}\,\mathbf{E}[(\widehat{f}_w(X) - \mathbf{E}[\widehat{f}_w(X) \mid U])^2] \\
&\mathbf{E}[(\widehat{f}_w(X) - Y)^2] + \mathbf{E}[(f^*(X) - Y)^2] \\
&\leq J_w(f_w) - \mathbf{E}[(f^*(X) - Y)^2] \\
&\coloneqq \mathbf{E}[(f(X) - Y)^2] + \gamma\|f - Df\|^2 \\
&\coloneqq \mathbf{E}[(f(X) - Y)^2] + \gamma\,\mathbf{E}[(f(X) - \mathbf{E}[f(X) \mid U])^2] \\
&= (1-\gamma)\,\mathbf{E}[(f(X) - \frac{1}{1-\gamma}f^*(X) - \frac{\gamma}{1-\gamma}l(X))^2] \\
&- \frac{\gamma^2}{1-\gamma}\,\mathbf{E}[l(X)^2] + \text{ const.}
\end{aligned}$$

Let $\gamma \coloneqq \frac{1-w}{w}$.

$$\begin{aligned}
&J_w(f) \\
&\coloneqq \mathbf{E}[(f(X) - Y)^2] + \gamma\|f - Df\|^2 \\
&\coloneqq \mathbf{E}[(f(X) - Y)^2] + \gamma\,\mathbf{E}[(f(X) - \mathbf{E}[f(X) \mid U])^2] \\
&= (1-\gamma)\,\mathbf{E}[(f(X) - \frac{1}{1-\gamma}f^*(X) - \frac{\gamma}{1-\gamma}l(X))^2] \\
&- \frac{\gamma^2}{1-\gamma}\,\mathbf{E}[l(X)^2] + \text{ const.}
\end{aligned}$$

where $l(x) \coloneqq \mathbf{E}[\mathbf{E}[f(X) \mid U] \mid X]$. From the optimality of f_w, we have

$$\begin{aligned}
&2(f(X) - f^*(X)) \\
&+ \frac{2(1-w)}{w}\,\mathbf{E}[f(X) - \mathbf{E}[f(X) \mid U] \mid X] \\
&= 2(f(X) - f^*(X)) \\
&+ \frac{2(1-w)}{w}(f(X) - \mathbf{E}[\mathbf{E}[f(X) \mid U] \mid X])
\end{aligned}$$

We are going to bound the two terms in the right hand side of the following:

$$\begin{aligned}
&\mathbf{E}[(\widehat{f}_w(X) - f^*(X))^2] \\
&= 2\,\mathbf{E}[(\widehat{f}_w(X) - f_w(X))^2] + 2\,\mathbf{E}[(f_w(X) - f^*(X))^2].
\end{aligned}$$

Bounding $\mathbf{E}[(\widehat{f}_w(X) - f_w(X))^2]$

$$\begin{aligned}
J_w(\widehat{f}_w) &= \frac{1}{w}\,\mathbf{E}[(\widehat{f}_w(X) - f_w(X) + f_w(X) - Y)^2] \\
&\quad + \frac{1-w}{w}\,\mathbf{E}[(\widehat{f}_w(X) - f_w(X) + f_w(X) - Y)^2]
\end{aligned}$$

Bounding $\mathbf{E}[(f_w(X) - f^*(X))^2]$ Let

$$f_w \coloneqq \operatorname*{arg\,min}_{f \in L^2_X} \mathbf{E}[(f_w(X) - f^*(X))^2] + \frac{1-w}{w}\,\mathbf{E}[(f_w(X) - \mathbf{E}[f^*(X) \mid U])^2].$$

Then, by the optimality of f_w, we have

$$\begin{aligned}
&\mathbf{E}[(f_w(X) - f^*(X))^2] \\
&+ \frac{1-w}{w}\,\mathbf{E}[(f_w(X) - \mathbf{E}[f_w(X) \mid U])^2] \\
&\leq \mathbf{E}[(f^*(X) - f^*(X))^2] \\
&+ \frac{1-w}{w}\,\mathbf{E}[(f^*(X) - \mathbf{E}[f^*(X) \mid U])^2],
\end{aligned}$$

which implies

$$\begin{aligned}&\mathbf{E}[(f_w(X)-f^*(X))^2]\\&\leq \frac{1-w}{w}\,\mathbf{E}[(f^*(X)-\mathbf{E}[f^*(X)\mid U])^2]\\&=V_w(f^*).\end{aligned}$$

□

G. Calculation of the shrinkage term in Joint-RR

We will show the following proposition.

Proposition 7.1. *Under the assumption in Eq. (1), we have*

$$\begin{aligned}&\min_{h\in L_U^2} J_w(f,h)\\&=\mathrm{MSE}(f)+\textit{const.}\\&\quad+\underbrace{\frac{1-w}{w}\,\mathbf{E}[(f(X)-\mathbf{E}[f(X)\mid U])^2]}_{\textit{The shrinkage regularizer.}}\end{aligned}$$

for any $f\in L_X^2$ and any $w\in(0,1)$.

Proof. On one hand, we have

$$\begin{aligned}&J_w(f,h)\\&=\frac{1}{w}\,\mathbf{E}[(f(X)-h(U))^2]+\frac{1}{1-w}\,\mathbf{E}[(h(U)-Y)^2]\\&=\frac{1}{w}\,\mathbf{E}[(f(X)-\mathbf{E}[f(X)\mid U])^2]\\&\quad+\frac{1}{w}\,\mathbf{E}[(\mathbf{E}[f(X)\mid U]-h(U))^2]\\&\quad+\frac{1}{1-w}\,\mathbf{E}[(h(U)-\mathbf{E}[Y\mid U])^2]\\&\quad+\frac{1}{1-w}\,\mathbf{E}[(\mathbf{E}[Y\mid U]-Y)^2]\\&=\frac{1}{w(1-w)}\Big\{w\,\mathbf{E}[(h(U)-\mathbf{E}[Y\mid U])^2]\\&\qquad\qquad+(1-w)\,\mathbf{E}[(\mathbf{E}[f(X)\mid U]-h(U))^2]\Big\}+C_1,\end{aligned}$$

where C_1 is the remaining term that does not depend on h. On the other hand, we have

$$\begin{aligned}&\mathbf{E}[(h(U)-h^\circ(U))^2]\\&=\mathbf{E}\left[\left(h(U)-w\,\mathbf{E}[Y\mid U]-(1-w)\,\mathbf{E}[f(X)\mid U]\right)^2\right]\\&=\mathbf{E}[h(U)^2]+C_2\\&\quad-2w\,\mathbf{E}[h(U)\,\mathbf{E}[Y\mid U]]\\&\quad+2(1-w)\,\mathbf{E}[h(U)\,\mathbf{E}[f(X)\mid U]]^2\\&=w\,\mathbf{E}[(h(U)-\mathbf{E}[Y\mid U])^2]+C_2\\&\quad+(1-w)\,\mathbf{E}[(h(U)-\mathbf{E}[f(X)\mid U])^2],\end{aligned}$$

where C_2 is the remaining term that does not depend on h. Thus, we have

$$J_w(f,h)=\frac{1}{w(1-w)}\,\mathbf{E}[(h(U)-h^\circ(U))^2]+C_3,\quad(27)$$

where C_3 is the remaining term that does not depend on h. This implies

$$h^\circ=\operatorname*{arg\,min}_{h\in L_U^2} J_w(f,h),$$

where $h^\circ(u)\coloneqq w\,\mathbf{E}[Y\mid U=u]+(1-w)\,\mathbf{E}[f(X)\mid U=u]$. Finally, we can calculate the minimizer as

$$\begin{aligned}&\min_{h\in L_U^2} J_w(f,h)\\&=J_w(f,h^\circ)\\&=\frac{1}{w}\,\mathbf{E}[(f(X)-\mathbf{E}[f(X)\mid U])^2]\\&\quad+\frac{1}{w}\,\mathbf{E}[(\mathbf{E}[f(X)\mid U]-h^\circ(U))^2]\\&\quad+\frac{1}{1-w}\,\mathbf{E}[(h^\circ(U)-\mathbf{E}[Y\mid U])^2]\\&\quad+\frac{1}{1-w}\,\mathbf{E}[(\mathbf{E}[Y\mid U]-Y)^2]\\&=\frac{1}{w}\,\mathbf{E}[(f(X)-\mathbf{E}[f(X)\mid U])^2]\\&\quad+\frac{1}{w}w^2\,\mathbf{E}[(\mathbf{E}[f(X)\mid U]-\mathbf{E}[Y\mid U])^2]\\&\quad+\frac{1}{1-w}(1-w)^2\,\mathbf{E}[(\mathbf{E}[f(X)\mid U]-\mathbf{E}[Y\mid U])^2]\\&\quad+\frac{1}{1-w}\,\mathbf{E}[(\mathbf{E}[Y\mid U]-Y)^2]\\&=\frac{1}{w}\,\mathbf{E}[(f(X)-\mathbf{E}[f(X)\mid U])^2]\\&\quad+\mathbf{E}[(\mathbf{E}[f(X)\mid U]-\mathbf{E}[Y\mid U])^2]\\&\quad+\frac{1}{1-w}\,\mathbf{E}[(\mathbf{E}[Y\mid U]-Y)^2]\\&=\left(\frac{1}{w}-1\right)\mathbf{E}[(f(X)-\mathbf{E}[f(X)\mid U])^2]\\&\quad+\mathbf{E}[(\mathbf{E}[f(X)\mid U]-\mathbf{E}[Y\mid U])^2]\\&\quad+\mathbf{E}[(f(X)-\mathbf{E}[f(X)\mid U])^2]\\&\quad+\frac{1}{1-w}\,\mathbf{E}[(\mathbf{E}[Y\mid U]-Y)^2]\\&=\frac{1-w}{w}\,\mathbf{E}[(f(X)-\mathbf{E}[f(X)\mid U])^2]\\&\quad+\mathbf{E}[(f(X)-\mathbf{E}[Y\mid U])^2]\\&\quad+\frac{1}{1-w}\,\mathbf{E}[(\mathbf{E}[Y\mid U]-Y)^2]\\&=\mathrm{MSE}(f)+\frac{1-w}{w}\,\mathbf{E}[(f(X)-\mathbf{E}[f(X)\mid U])^2]+C_4.\end{aligned}$$

□

H. Discussion on the assumption

So far, we have focused on the ideal case in which the conditional mean independence (Eq. (1)) holds. However, it may be difficult to exactly ensure the condition in practice.

Here, we relax Eq. (1) by allowing the gap between the left-hand and right-hand sides to be potentially larger than zero but bounded by c^2 for some constant $c \in (0, \infty)$. We will show that (i) even the best possible method suffers an MSE of at least $c^2/2$ in the worse-case within this scenario while (ii) 2Step-RR suffers an MSE of at most $c^2 + o(1)$.

To see the claim (ii), note that we have already shown that $\widetilde{f}$ converges to $\mathbf{E}[\mathbf{E}[Y \mid U] \mid X = (\cdot)]$. This implies that 2Step-RR method suffers an MSE of

$$\begin{aligned}
&\mathbf{E}[(\mathbf{E}[\mathbf{E}[Y \mid U] \mid X] - \mathbf{E}[Y \mid X])^2] + o(1) \\
&= \mathbf{E}[(\mathbf{E}[\mathbf{E}[Y \mid U] - \mathbf{E}[Y \mid U, X] \mid X])^2] + o(1) \\
&\leq \mathbf{E}[(\mathbf{E}[Y \mid U] - \mathbf{E}[Y \mid U, X])^2] + o(1) \\
&\leq c^2 + o(1)
\end{aligned}$$

under the relaxed assumption.

To show the claim (i), let us first define the class of problem instances satisfying the relaxed assumption. Fix any probability density function (p.d.f.) $p(x)$ defined over $\mathcal{X}$. For $c \in (0, \infty]$, let $\mathcal{P}_c$ denote the set of p.d.f.-s over $\mathcal{X} \times \mathcal{U} \times \mathcal{Y}$ defined as follows. Any p.d.f. $q(x, u, y)$ over $\mathcal{X} \times \mathcal{U} \times \mathcal{Y}$ is a member of $\mathcal{P}_c$ if and only if, for random variables $(X, U, Y) \sim q(x, u, y)$, X follows $p(x)$ and Eq. (1) is violated by c in $L^2(p)$-distance:

$$\mathbf{E}[(\mathbf{E}[Y \mid U] - \mathbf{E}[Y \mid U, X])^2] \leq c^2.$$

In this setup, we will establish an lower bound of

$$E_{\text{minimax}} \coloneqq \inf_{\widehat{f}_{(\cdot)}} \sup_{q \in \mathcal{P}_c} \mathbf{E}[(\widehat{f}_{S_X, S_Y}(X) - \mathbf{E}[Y \mid X])^2],$$

where $\widehat{f}_{(\cdot)}$ represents any estimator that uses mediated uncoupled data (S_X, S_Y) and produces a function from $\mathcal{X}$ to $\mathcal{Y}$, and the expectation is taken over (S_X, S_Y) and $(X, Y) \sim q(x, y)$.

Define $\rho \colon \mathcal{P}_c^2 \to [0, \infty)$ by

$$\begin{aligned}
\rho(q_1, q_2) \coloneqq \mathbf{E}[(&\mathbf{E}[Y_1 \mid X_1 = X_0] \\
&- \mathbf{E}[Y_2 \mid X_2 = X_0])^2]
\end{aligned}$$

for any $q_1, q_2 \in \mathcal{P}_c$, where $X_0 \sim p(x)$, $(X_1, U_1, Y_1) \sim q_1(x, u, y)$, and $(X_2, U_2, Y_2) \sim q_2(x, u, y)$, and they are all independent.

We will use the following lemma.

Lemma 7.3. *Suppose $p(x)$ is a symmetric, centered density function over $\mathcal{X}$, i.e., $p(x) = p(-x)$ and $\int xp(x)\mathrm{d}x = 0$. Then, we have*

$$\inf_{\widehat{p}_{(\cdot)}} \sup_{q \in \mathcal{P}_c} \mathbf{E}\left[\rho\left(\widehat{p}_{S_X, S_Y}, q\right)\right] \geq \frac{1}{2}c^2.$$

Once we establish Lemma 7.3, we immediately obtain the following two propositions.

Proposition 7.2. *Suppose $p(x)$ is a symmetric, centered density function over $\mathcal{X}$, i.e., $p(x) = p(-x)$ and $\int xp(x)\mathrm{d}x = 0$. For any estimator $\widehat{f}_{(\cdot)}$ that takes mediated uncoupled data as input and produces a function from $\mathcal{X}$ to $\mathcal{Y}$, we have*

$$\sup_{p^* \in \mathbb{P}_{p,c}} \mathbf{E}\left[\left(\widehat{f}_{S_X, S_Y}(X) - \mathbf{E}[Y \mid X]\right)^2\right] \geq \frac{1}{2}c^2,$$

where $(X, Y) \sim p^*(x, y)$. This holds no matter how large n and n' are.

Proposition 7.3. *Suppose $p(x)$ is a symmetric, centered density function over $\mathcal{X}$, i.e., $p(x) = p(-x)$ and $\int xp(x)\mathrm{d}x = 0$. For any stochastic estimator $\widehat{Y}_{(\cdot)}$ that takes mediated uncoupled data as input and produces a $\mathcal{Y}$-valued random variable depending on the test sample X, we have*

$$\sup_{q \in \mathcal{P}_c} \mathbf{E}\left[\left(\mathbf{E}[\widehat{Y}_{S_X, S_Y} \mid X] - \mathbf{E}[Y \mid X]\right)^2\right] \geq \frac{1}{2}c^2,$$

where $(X, Y) \sim q(x, y)$. This holds no matter how large n and n' are.

Proof of Proposition 7.3 and Proposition 7.2. If the statements were not to hold, it would contradict Lemma 7.3. □

We present our proof of Lemma 7.3 in details since it provides an intuition about the problem with a concrete example.

Proof of Lemma 7.3. The proof is based on Le Cam's method. We construct two p.d.f.-s within $\mathcal{P}_c$ that are distant enough in terms of ρ, whose corresponding mediated uncoupled data, however, have the identical distribution.

For any $\theta \in \mathbb{R}$, let $p(x, u, y; \theta)$ be the joint p.d.f. of the variables defined by

$$X \sim p(x), \quad U \sim q(u), \quad Y = \theta X$$

with any p.d.f. $q(u)$ over $\mathcal{U}$, where X and U are independent, and $\theta \in \mathbb{R}$.

Take the two density functions $p_1(x, u, y) \coloneqq p(x, u, y; c/\sigma)$ and $p_2(x, u, y) \coloneqq p(x, u, y; -c/\sigma)$, where $\sigma \coloneqq \sqrt{\mathrm{Var}[X]}$. Note that each of their parameters is the negation of the other.

p_1 and p_2 both belong to $\mathcal{P}_c$ because $p_1(x) = p_2(x) = p(x)$, and both for $(X, Y, U) \sim p_1(x, u, y)$ and for $(X, Y, U) \sim p_2(x, u, y)$,

$$\begin{aligned}&\mathbf{E}[(\mathbf{E}[Y \mid U] - \mathbf{E}[Y \mid U, X])^2] \\ &= \mathbf{E}[(0 + (c/\sigma)X)^2] \quad (\text{or } \mathbf{E}[(0 - (c/\sigma)X)^2]) \\ &= c^2.\end{aligned}$$

Obviously, $p_1(x, u) = p(x)q(u) = p_2(x, u)$. Since $p_1(y) = p_2(y)$ from the symmetry $p(x) = p(-x)$, $p_1(u, y) = q(u)p_1(y) = q(u)p_2(y) = p_2(u, y)$. Hence, the distribution of the mediated uncoupled data induced by $p_1(x, u, y)$ and that by $p_2(x, u, y)$ are identical to each other. On the other hand, p_1 and p_2 are $2c$-separated:

$$\begin{aligned}&\rho(p_1(x, u, y), p_2(x, u, y)) \\ &= \mathbf{E}[(\mathbf{E}[Y_1 \mid X] - \mathbf{E}[Y_2 \mid X])^2] \\ &= \mathbf{E}[((c/\sigma)X + (c/\sigma)X)^2] \\ &= (2c)^2.\end{aligned}$$

The argument above intuitively tells that it is impossible to distinguish distinct p.d.f.-s p_1 and p_2 only with the information given by the mediated uncouple data, and even the best possible guess would suffer loss proportionally to c^2 in the worst case. More formally, by applying Le Cam's method, we obtain

$$\inf_{\widehat{p}_{(\cdot)}} \sup_{q \in \mathcal{P}_c} \mathbf{E}\left[\rho\left(\widehat{p}_{S_{n,n'}}(x, u, y), q(x, u, y)\right)\right] \geq \frac{1}{2}c^2.$$

□

I. Proof of Theorem 4.1

See Appendix A. (This subsection is only a stub pointing to Appendix A and will be removed in the next version.)

J. Linear-in-parameter models

We consider the following regularized version of our method with linear-in-parameter models:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{b_\mathcal{F} + b_\mathcal{H}}} \left[\widehat{J}_w(f_{\boldsymbol{\alpha}}, h_{\boldsymbol{\beta}}) + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta}\right],$$

where $f_{\boldsymbol{\alpha}}(\boldsymbol{x}) \coloneqq \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x})$, $h_{\boldsymbol{\beta}}(\boldsymbol{u}) \coloneqq \boldsymbol{\beta}^\top \boldsymbol{\psi}(\boldsymbol{u})$, $\boldsymbol{\theta} \coloneqq (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$, and $\lambda \in (0, \infty)$. Here,

$$\begin{aligned}\widehat{J}_w(f_{\boldsymbol{\alpha}}, h_{\boldsymbol{\beta}}) &= \frac{1}{nw}\sum_{i=1}^{n}(\boldsymbol{\alpha}^\top \boldsymbol{\varphi}(X_i) - \boldsymbol{\beta}^\top \boldsymbol{\psi}(U_i))^2 \\ &\quad + \frac{1}{n'(1-w)}\sum_{i=1}^{n'}(\boldsymbol{\beta}^\top \boldsymbol{\psi}(U'_i) - Y'_i)^2 \\ &= \boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{b}\boldsymbol{\theta} + \text{const.},\end{aligned}$$

where the matrices $\boldsymbol{A}$ and $\boldsymbol{b}$ are defined as

$$\begin{aligned}\boldsymbol{A} &\coloneqq \frac{1}{nw}\sum_{i=1}^{n}\left[\begin{pmatrix}\boldsymbol{\varphi}(X_i) \\ -\boldsymbol{\psi}(U_i)\end{pmatrix}\begin{pmatrix}\boldsymbol{\varphi}(X_i) \\ -\boldsymbol{\psi}(U_i)\end{pmatrix}^\top\right] \\ &\quad + \frac{1}{n'(1-w)}\sum_{i=1}^{n'}\left[\begin{pmatrix}\mathbf{0} \\ \boldsymbol{\psi}(U_i)\end{pmatrix}\begin{pmatrix}\mathbf{0} \\ \boldsymbol{\psi}(U_i)\end{pmatrix}^\top\right], \\ \text{and} \quad \boldsymbol{b} &\coloneqq \frac{1}{n'(1-w)}\sum_{i=1}^{n'}(\mathbf{0}^\top, Y'_i\boldsymbol{\psi}(U'_i)^\top)^\top.\end{aligned}$$

Since A is positive semi-definite, we obtain the minimizer in a closed form by

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &\coloneqq \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^{b_\mathcal{F} + b_\mathcal{H}}} \left[\hat{J}_w(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta}\right] \\ &= \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^{b_\mathcal{F} + b_\mathcal{H}}} \left[\boldsymbol{\theta}^\top(\boldsymbol{A} + \lambda \boldsymbol{I})\boldsymbol{\theta} - \boldsymbol{b}\boldsymbol{\theta}\right] \\ &= (\boldsymbol{A} + \lambda \boldsymbol{I}_{b_\mathcal{F} + b_\mathcal{H}})^{-1}\boldsymbol{b} \qquad (28)\end{aligned}$$

for any positive λ, where $\boldsymbol{I}_k$ denotes the k-by-k identity matrix.

Furthermore, using the block-wise matrix inversion formula, we have

$$\widehat{\boldsymbol{\alpha}} \coloneqq \boldsymbol{M}_1^{-1}\boldsymbol{M}_2\widehat{\boldsymbol{\beta}}, \quad \text{and} \quad \widehat{\boldsymbol{\beta}}(\boldsymbol{M}_3 - \boldsymbol{M}_2^\top \boldsymbol{M}_1^{-1}\boldsymbol{M}_2)^{-1}\boldsymbol{b}_1, \qquad (29)$$

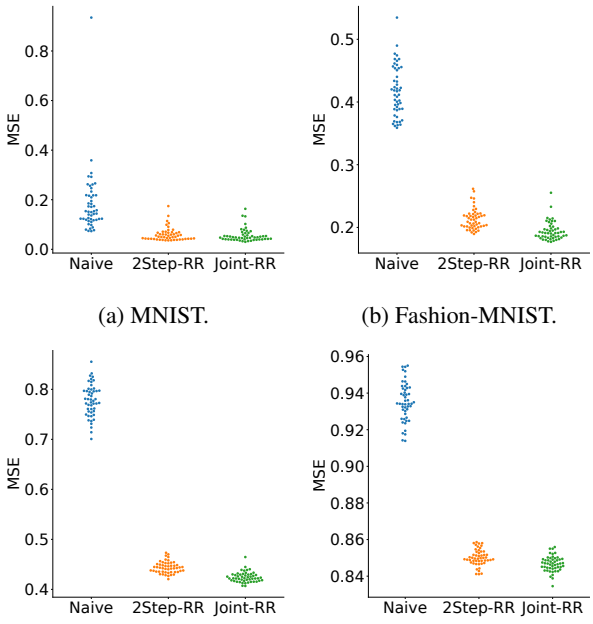
where

$$\begin{aligned}\boldsymbol{M}_1 &\coloneqq \frac{1}{nw}\sum_{i=1}^{n}[\boldsymbol{\varphi}(X_i)\boldsymbol{\varphi}(X_i)^\top] + \lambda \boldsymbol{I}_{b_\mathcal{F}}, \\ \boldsymbol{M}_2 &\coloneqq \frac{1}{nw}\sum_{i=1}^{n}[\boldsymbol{\varphi}(\boldsymbol{x}_i)\boldsymbol{\psi}(U_i)^\top], \\ \boldsymbol{M}_3 &\coloneqq \frac{1}{nw}\sum_{i=1}^{n}[\boldsymbol{\psi}(U_i)\boldsymbol{\psi}(U_i)^\top] \\ &\quad + \frac{1}{n'(1-w)}\sum_{i=1}^{n'}[\boldsymbol{\psi}(U'_i)\boldsymbol{\psi}(U'_i)^\top] + \lambda \boldsymbol{I}_{b_\mathcal{H}}, \\ \boldsymbol{b}_1 &\coloneqq \frac{1}{n'(1-w)}\sum_{i=1}^{n'}[Y'_i\boldsymbol{\psi}(U'_i)].\end{aligned}$$

Eq. (29) involves matrices of size at most $\max(b_\mathcal{F}, b_\mathcal{H})$-by-$\max(b_\mathcal{F}, b_\mathcal{H})$, which requires less computation resources in terms of both space and time compared to Eq. (28) involving inversion of a $(b_\mathcal{F} + b_\mathcal{H})$-by-$(b_\mathcal{F} + b_\mathcal{H})$ matrix.

K. Scatter Plots of MSEs for the Low-quality Image Classification

Figure 5 shows scatter plots of MSEs for the low-quality image classification. We can see that the proposed methods outperform the naive method.



(a) MNIST. (b) Fashion-MNIST.

(c) CIFAR10. (d) CIFAR100.

Figure 5. MSEs for the experiments on low-quality image classification.