Mediated Uncoupled Learning and Validation with Bregman Divergences: Loss Family with Maximal Generality

Ikko Yamane
ENSAI/CREST

Yann Chevaleyre
LAMSADE, CNRS,
Université Paris-Dauphine,
PSL Research University

Takashi Ishida
The University of Tokyo/RIKEN

Florian Yger
LAMSADE, CNRS,
Université Paris-Dauphine,
PSL Research University

Abstract

In *mediated uncoupled learning* (MU-learning), the goal is to predict an output variable Y given an input variable X as in ordinary supervised learning while the training dataset has no joint samples of (X, Y) but only independent samples of (X, U) and (U, Y) each observed with a *mediating* variable U. The existing MU-learning methods can only handle the squared loss, which prohibited the use of other popular loss functions such as the cross-entropy loss. We propose a general MU-learning framework that allows for the problems with Bregman divergences, which cover a wide range of loss functions useful for various types of tasks, in a unified manner. This loss family has *maximal generality* among those whose minimizers characterize the conditional expectation. We prove that the proposed objective function is a tighter approximation to the oracle loss that one would minimize if ordinary supervised samples of (X, Y) were available. We also propose an estimator of an interval containing the expected test loss of predictions of a trained model only using (X, U)- and (U, Y)-data. We provide a theoretical analysis on the excess risk for the proposed method and confirm its practical usefulness with regression experiments with synthetic data and low-quality image classification experiments with benchmark datasets.

1 INTRODUCTION

Supervised learning has found many successful applications and become a standard approach to many real-world pattern recognition and prediction tasks (Mohri et al., 2012; Murphy, 2012; Shalev-Shwartz and Ben-David, 2014). In its standard form, the goal of ordinary supervised learning is to predict an output variable Y given an input variable X from a training dataset consisting of direct input-output correspondences, i.e., joint data samples of (X, Y).

However, collecting such joint data samples can be difficult in some applications (Chapelle et al., 2006; Zhu, 2005; van Engelen and Hoos, 2020). *Mediated uncoupled learning (MU-learning)*, a framework for learning without direct input-output correspondences, can facilitate training data collection in such situations (Yamane et al., 2021). MU-learning does not require joint data of (X, Y) but allows them to be independently observed with another variable U called *mediating variable*. Namely, we only need independent training data of (X, U) and (U', Y'), called *Mediated Uncoupled data (MU-data)*, where (X, U, Y) and (X', U', Y') are i.i.d. with Y and X' being unobserved.

For instance, it may be difficult to collect text translation examples between minor languages (corresponding to X and Y) when we want to train a machine translator between them (Haddow et al., 2022). Instead, we may collect text data written in each of the languages with a major language such as English translations (corresponding to U) and apply an MU-learning method. Other examples include image sentiment analysis (Mittal et al., 2018) from images with text captions (Xu et al., 2015) and text data with sentiment labels (Medhat et al., 2014) and counterfactual prediction (Pearl, 2009; Johansson et al., 2016; Zou et al., 2020).

Yamane et al. (2021) proposed a two-step method for MU-learning that yields a weakly consistent estimator under some conditions on identifiability and model-specification. They also proposed a regularized, one-step variant that jointly performs the two steps, which was empirically shown to improve the two-step method. However, their methods can only handle the squared loss, whose specific properties were necessary for deriving and analyzing their methods. This limitation prohibited the use of, e.g., the cross-entropy loss that is popular for classification.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

In this paper, we propose an MU-learning framework that allows the use of a wide class of loss functions based on *Bregman divergences*. Under mild conditions, this loss family is in fact the *most general* among the ones whose minimizers characterize the conditional expectation (Banerjee et al., 2005a). It includes many different types such as the cross-entropy loss, the Itakura-Saito distance, the generalized I-divergence, and the robust bi-tempered logistic loss (Banerjee et al., 2005b; Amid et al., 2019). The proposed framework enables users to choose an appropriate loss function depending on the task at hand. It is even possible to design a custom loss function tailored to the task by learning the convex function used to define the Bregman divergence (Siahkamari et al., 2020).

Under this setup, we develop a statistically consistent two-step method and a regularized one-step method. Furthermore, we propose a way to validate the prediction performance of models trained with our methods by providing an interval estimate of the test loss only using MU-data, without (X, Y)-data. This is useful for knowing whether the trained model is good or bad before deployment.

We show that the proposed objective function, compared with that of the existing approach, is a tighter approximation to the oracle loss that one would use if (X, Y)-data were available. Our approximation provides the proposed method with another nice property that the asymptotic bias will be zero if either (X, U) or (U', Y') is noise-free in some sense. We also prove finite-sample error bounds for the two proposed methods and confirm the usefulness of the proposed method through experiments in least-squares regression and low-quality image classification setups.

Our contributions: (i) We propose an MU-learning framework that removes the limitation of the existing MU-learning methods on the loss by introducing Bregman divergences which *maximally* generalize the loss family. (ii) We also propose a way to *validate the performance* of a trained model *only using MU-data*. (iii) We prove that the proposed objective function is a *tighter* approximation to the oracle loss compared to the existing one. (iv) We provide *finite-sample* excess risk bounds and a bias analysis for the proposed methods. Our code is available at `https://github.com/i-yamane/mediated_uncoupled_learning`.

2 PROBLEM SETUP

Let X and Y denote an $\mathcal{X}$-valued input variable and a $\mathcal{Y}$-valued output variable with an underlying probability measure P, respectively, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^k$ are measurable spaces. Our goal is to learn the function f best predicting Y from X so that the expected loss $\mathbf{E}[\ell_\phi(Y, f(X))]$ will be minimized, where $\mathbf{E}[\cdot]$ denotes the expectation and $\ell_\phi(\cdot, \cdot)$ is a *Bregman divergence* defined as follows.

Definition 2.1 (Bregman divergences, Bregman risk)**.** Let $\phi\colon \mathbb{R}^k \to \mathbb{R}$ be a differentiable, strictly convex function. We define the Bregman divergence associated with ϕ as

$$\ell_\phi(y_1, y_2) \coloneqq \phi(y_1) - \phi(y_2) - \langle y_1 - y_2, \nabla\phi(y_2)\rangle$$

for any $(y_1, y_2) \in \mathbb{R}^k \times \mathbb{R}^k$, where ∇ is the gradient operator. Furthermore, for any $\mathbb{R}^k$-valued random variables Y_1 and Y_2, we define the Bregman risk associated with ϕ as

$$D_\phi(Y_1, Y_2) \coloneqq \mathbf{E}[\ell_\phi(Y_1, Y_2)].$$

We have $y_1 = y_2$ if and only if $\ell_\phi(y_1, y_2) = 0$. Similarly, $Y_1 = Y_2$ almost surely if and only if $D_\phi(Y_1, Y_2) = 0$. The symmetry $\ell_\phi(y_1, y_2) = \ell_\phi(y_2, y_1)$ or $D(Y_1, Y_2) = D_\phi(Y_2, Y_1)$ does not generally hold.

This paper focuses on the setup called *mediated uncoupled learning (MU-learning)* (Yamane et al., 2021). In the ordinary supervised learning, we are given training samples of any joint data of (X, Y). In MU-learning, on the other hand, we do not observe joint (X, Y)-data as training samples, but we are only given data of (X, U) and (U, Y) independently observed with another $\mathcal{U}$-valued variable U. More formally, the training data that we have are $\{(X_i, U_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{X,U}$ and $\{(U'_i, Y'_i)\}_{i=1}^{n'} \overset{\text{i.i.d.}}{\sim} P_{U,Y}$, where $P_{X,U}$ and $P_{U,Y}$ are probability distributions of (X, U) and (U, Y), respectively, obtained by marginalizing the common joint distribution $P_{X,U,Y}$ of (X, U, Y) defined based on P. Since U mediates between the uncoupled variables X and Y, we refer to data of this form as *mediated uncoupled data (MU-data)*, and we refer to the task of learning from MU-data as *mediated uncoupled learning (MU-learning)*.

Yamane et al. (2021) proposed two methods for MU-learning and theoretically and empirically studied their usefulness. However, their methods are limited to the case in which $\ell_\phi(\cdot, \cdot)$ is the squared loss. In Section 4, we propose a method that can handle any Bregman divergence, which covers a wide class of loss functions including the squared loss and the cross-entropy loss as special cases.

In fact, Bregman divergences are the *most general* loss family among those whose minimizers characterize the conditional expectation. More specifically, any loss ℓ has to be a Bregman divergence if we want $\mathbf{E}[Y \mid X]$ to be a minimizer of $\mathbf{E}[\ell(Y, f(X))]$ over all measurable functions f, under regularity conditions (Banerjee et al., 2005a).

3 EXISTING METHODS

We review existing methods before presenting our methods. To the best of our knowledge, existing methods only focused on the squared loss $(y_1, y_2) \mapsto \frac{1}{2}\|y_1 - y_2\|_2^2$ (Yamane et al., 2021), which is a special instance of the Bregman divergence induced by $\phi(t) = \frac{1}{2}\|t\|_2^2$.

3.1 Naive Approach Based on Separate Estimators

A naive approach to MU-learning is to separately learn two functions and combine them: Let $\widehat{g}_1(x)$ be a function for predicting U from X and $\widehat{g}_2(u)$ be a function for predicting Y from U. Then, we predict U from X as $\widehat{U} \coloneqq \widehat{g}_1(X)$ and then Y based on $\widehat{U}$ as $\widehat{Y} \coloneqq \widehat{g}_2(\widehat{U}) = \widehat{g}_2(\widehat{g}_1(X))$.

Although this approach is simple and intuitive, it introduces avoidable bias unless $\mathbf{E}[Y \mid U]$ is a linear function of U, or U is a deterministic function of X (Yamane et al., 2021).

3.2 Two-step Regressed Regression (2Step-RR)

Two-Step Regressed Regression (2Step-RR) (Yamane et al., 2021) is an MU-learning method that yields a weakly consistent estimator to $\mathbf{E}[Y \mid X]$ as long as U is informative enough in the sense that

$$\mathbf{E}[Y \mid X, U] = \mathbf{E}[Y \mid U], \text{ a.s.}, \tag{1}$$

and models are correctly specified. Eq. 1 is called the *conditional mean independence*. First, 2Step-RR performs regression on $\{(U'_i, Y'_i)\}_{i=1}^{n'}$ to obtain an estimate $\widehat{h}_{\mathrm{u}}^{\text{2RR}}(U)$ of $\mathbf{E}[Y \mid U]$ as

$$\widehat{h}_{\mathrm{u}}^{\text{2RR}} \coloneqq \operatorname*{arg\,min}_{h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \frac{1}{n'} \sum_{i=1}^{n'} \|h_{\mathrm{u}}(U'_i) - Y'_i\|_2^2,$$

where $\mathcal{H}_{\mathrm{u}}$ is a hypothesis class for $\widehat{h}_{\mathrm{u}}^{\text{2RR}}$. Then, it constructs a supervised dataset $\{(X_i, \widehat{Y}_i)\}_{i=1}^{n}$ by replacing each U_i of the dataset $\{(X_i, U_i)\}_{i=1}^{n}$ with $\widehat{Y}_i \coloneqq \widehat{h}_{\mathrm{u}}^{\text{2RR}}(U_i)$. Finally, it performs regression on $\{(X_i, \widehat{Y}_i)\}_{i=1}^{n}$ to obtain an estimate $\widehat{h}_{\mathrm{x}}^{\text{2RR}}(X)$ of $\mathbf{E}[Y \mid X]$:

$$\widehat{h}_{\mathrm{x}}^{\text{2RR}} \coloneqq \operatorname*{arg\,min}_{h_{\mathrm{x}} \in \mathcal{H}_{\mathrm{x}}} \frac{1}{n} \sum_{i=1}^{n} \|h_{\mathrm{x}}(X_i) - \widehat{Y}_i\|_2^2,$$

where $\mathcal{H}_{\mathrm{x}}$ is a hypothesis class of bounded functions for $\widehat{h}_{\mathrm{x}}^{\text{2RR}}$. It only requires performing ordinary regression twice, and one can use any standard machine learning models and optimization methods for each step.

3.3 Joint Regressed Regression (Joint-RR)

In 2Step-RR, the first regression step trains $\widehat{h}_{\mathrm{u}}^{\text{2RR}}$ without any reference to the second regression step for learning $\widehat{h}_{\mathrm{x}}^{\text{2RR}}$. Yamane et al. (2021) also proposed a variant of 2Step-RR called *Joint-RR* which jointly performs the two steps of 2Step-RR. It solves a single optimization problem whose objective function is a convex combination [*1] of those of 2Step-RR:

$$(\widehat{h}_{\mathrm{x}}^{\text{JRR}}, \widehat{h}_{\mathrm{u}}^{\text{JRR}}) \coloneqq \operatorname*{arg\,min}_{(h_{\mathrm{x}}, h_{\mathrm{u}}) \in \mathcal{H}_{\mathrm{x}} \times \mathcal{H}_{\mathrm{u}}} \widehat{J}_{\text{Joint-RR}}(h_{\mathrm{x}}, h_{\mathrm{u}}),$$

where

$$\widehat{J}_{\text{Joint-RR}}(h_{\mathrm{x}}, h_{\mathrm{u}}) \coloneqq \frac{2}{n'} \sum_{i=1}^{n'} \|h_{\mathrm{u}}(U'_i) - Y'_i\|_2^2 + \frac{2}{n} \sum_{i=1}^{n} \|h_{\mathrm{x}}(X_i) - h_{\mathrm{u}}(U_i)\|_2^2. \tag{2}$$

The expectation behind this approach is that h_{u}'s output will be adjusted through the joint optimization so that h_{x} will better fit the generated supervised data $\{(X_i, \widehat{Y}_i)\}_{i=1}^{n}$, where $\widehat{Y}_i \coloneqq h_{\mathrm{u}}(U_i)$.

The expected objective function of Joint-RR is an upper bound of the *oracle* mean squared error (MSE) that we would minimize if (X, Y)-data were available:

$$\begin{aligned} \mathbf{E}[\|h_{\mathrm{x}}(X) - Y\|_2^2] &\leq \mathbf{E}[\widehat{J}_{\text{Joint-RR}}(h_{\mathrm{x}}, h_{\mathrm{u}})] \\ &=: J_{\text{Joint-RR}}(h_{\mathrm{x}}, h_{\mathrm{u}}). \end{aligned} \tag{3}$$

This implies that if we succeed in making the right hand side small, the MSE will not exceed it. It was proven that Joint-RR approximately minimizes the right hand side, and the minimum of the empirical objective approaches to the minimum of its expectation as the sample sizes tend to infinity (Yamane et al., 2021).

4 PROPOSED METHODS

A limitation of 2Step-RR (Section 3.2) and Joint-RR (Section 3.3) is that they only allow us to use the squared loss. In this section, we propose an MU-learning framework that can handle any of the loss functions based on Bregman divergences. It is a wide loss family that includes the squared loss as an instance. The central challenge in this extension is that it is not straightforward to give theoretical guarantees when we replace the squared loss with the general Bregman divergence in 2Step-RR and Joint-RR. For instance, heuristically replacing the loss functions in Joint-RR can fail to keep the basic property analogous to Eq. (3).

4.1 Two-step MU-Learning with Bregman Divergences (2Step-BregMU)

First, we present a new two-step MU-learning method with Bregman divergences. This method generalizes 2Step-RR to the case with any Bregman divergence ℓ_ϕ. We train a function $\widehat{h}_{\mathrm{u}}^{\text{2Breg}} : \mathcal{U} \to \mathcal{Y}$ for predicting Y from U:

$$\widehat{h}_{\mathrm{u}}^{\text{2Breg}} = \operatorname*{arg\,min}_{h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \frac{1}{n'} \sum_{i=1}^{n'} \ell_\phi(Y'_i, h_{\mathrm{u}}(U'_i)), \tag{4}$$

Then, we train another function $\widehat{h}_{\mathrm{x}}^{\text{2Breg}} : \mathcal{X} \to \mathcal{Y}$ so as to predict $\widehat{h}_{\mathrm{u}}^{\text{2Breg}}(U)$ from X:

$$\widehat{h}_{\mathrm{x}}^{\text{2Breg}} = \operatorname*{arg\,min}_{h_{\mathrm{x}} \in \mathcal{H}_{\mathrm{x}}} \frac{1}{n} \sum_{i=1}^{n} \ell_\phi(\widehat{h}_{\mathrm{u}}^{\text{2Breg}}(U_i), h_{\mathrm{x}}(X_i)). \tag{5}$$

[*1] There is a freedom in the choice of the mixing coefficients, but Yamane et al. (2021) suggested setting them to equal weights, which is equivalent to what we present here.

Here, $\mathcal{H}_{\mathrm{u}} \subseteq \{h_{\mathrm{u}} \colon \mathcal{U} \to \mathcal{Y}\}$ and $\mathcal{H}_{\mathrm{x}} \subseteq \{h_{\mathrm{x}} \colon \mathcal{X} \to \mathcal{Y}\}$ are function classes. We call this method *2Step-BregMU*.

Note that swapping the arguments of the loss as $\ell_\phi(h_{\mathrm{u}}(U'_i), Y'_i)$ instead of Eq. (4) or $\ell_\phi(h_{\mathrm{x}}(X_i), \widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U_i))$ instead of Eq. (5) would not generally give the correct result because of the asymmetry of the Bregman divergence.

4.2 Joint MU-Learning with Bregman Divergences (Joint-BregMU)

Next, we present a one-step method that aims to jointly learn the two functions in 2Step-BregMU, which is less straightforward compared to the case of the squared loss. The resulting method is similar to Joint-RR in the sense that it minimizes an upper bound of the *oracle* supervised risk $D_\phi(Y, h_{\mathrm{x}}(X))$. However, it remains to establish a bound in terms of MU-data on the oracle supervised risk only using general properties that all the Bregman divergences admit. We start by introducing the following well-known property of Bregman divergences (Nock et al., 2016; Dhillon and Tropp, 2008).

Lemma 4.1. *For any* $(y_1, y_2, z) \in (\mathbb{R}^k)^3$,

$$\ell_\phi(y_1, y_2) = \ell_\phi(y_1, z) + \ell_\phi(z, y_2) + (y_1 - z)^\top (\nabla\phi(z) - \nabla\phi(y_2)).$$

Lemma 4.1 is useful for developing MU-learning methods because it allows us to express the oracle supervised risk $D_\phi(Y, h_{\mathrm{x}}(X))$, which involves (X, Y)-data, in terms of $D_\phi(Y, h_{\mathrm{u}}(U))$ and $D_\phi(h_{\mathrm{u}}(U), h_{\mathrm{x}}(X))$, which only require (U, Y)- and (X, U)-data, respectively:

$$D_\phi(Y, h_{\mathrm{x}}(X)) = D_\phi(Y, h_{\mathrm{u}}(U)) + D_\phi(h_{\mathrm{u}}(U), h_{\mathrm{x}}(X)) + \mathbf{E}[(Y - h_{\mathrm{u}}(U))^\top (\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X)))].$$

However, we still have the last term $\mathbf{E}[(Y - h_{\mathrm{u}}(U))^\top (\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X)))]$ that we cannot directly approximate using MU-data, $\{(X_i, U_i)\}_{i=1}^n$ and $\{(U'_i, Y'_i)\}_{i=1}^{n'}$. We bound this term using the inequality $-\|u\|_2 \cdot \|v\|_2 \leq u^\top v \leq \|u\|_2 \cdot \|v\|_2$ that holds for any $u, v \in \mathbb{R}^d$ to obtain a tractable expression, where $\|\cdot\|_2$ is the ℓ_2-norm, which leads to the following lemma. To state the lemma, denote $L^2(\mathcal{X}, \mathcal{Y}; P) \coloneqq \{h_{\mathrm{x}} \colon \mathcal{X} \to \mathcal{Y}, \mathbf{E}[h_{\mathrm{x}}(X)^2] < \infty\}$ and $L^2(\mathcal{U}, \mathcal{Y}; P) \coloneqq \{h_{\mathrm{u}} \colon \mathcal{U} \to \mathcal{Y}, \mathbf{E}[h_{\mathrm{u}}(U)^2] < \infty\}$. Also denote $\|f\|_{L^2}^2 \coloneqq \int \|f\|_2^2 \, \mathrm{d}P$ for any square integrable function f. For example, $\|h_{\mathrm{x}}(X)\|_{L^2}^2 = \mathbf{E}[\|h_{\mathrm{x}}(X)\|_2^2]$ for any $h_{\mathrm{x}} \in L^2(\mathcal{X}, \mathcal{U}; P)$ and $\|h_{\mathrm{u}}(U)\|_{L^2}^2 \coloneqq \mathbf{E}[\|h_{\mathrm{u}}(U)\|_2^2]$ for any $h_{\mathrm{u}} \in L^2(\mathcal{X}, \mathcal{U}; P)$.

Lemma 4.2 (Bounds on Bregman divergences)**.** *For any* $h_{\mathrm{x}} \in L^2(\mathcal{X}, \mathcal{U}; P)$ *and any* $h_{\mathrm{u}} \in L^2(\mathcal{X}, \mathcal{U}; P)$, $D_\phi(Y, h_{\mathrm{x}}(X))$ *has the following lower and upper bound:*

$$B_\phi^\times(h_{\mathrm{x}}, h_{\mathrm{u}}, -1) \leq D_\phi(Y, h_{\mathrm{x}}(X)) \leq B_\phi^\times(h_{\mathrm{x}}, h_{\mathrm{u}}, +1), \tag{6}$$

Algorithm 1 Joint-BregMU

$$(\widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}, \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}) \leftarrow \operatorname*{arg\,min}_{(h_{\mathrm{x}}, h_{\mathrm{u}}) \in \mathcal{H}_{\mathrm{x}} \times \mathcal{H}_{\mathrm{u}}} \widehat{B}_\phi^\times(h_{\mathrm{x}}, h_{\mathrm{u}}, +1),$$

where we denote for $s \in \{-1, +1\}$,

$$\widehat{B}_\phi^\times(h_{\mathrm{x}}, h_{\mathrm{u}}, s) \coloneqq \frac{1}{n'} \sum_{i=1}^{n'} \ell_\phi(Y'_i, h_{\mathrm{u}}(U'_i)) + \frac{1}{n} \sum_{i=1}^{n} \ell_\phi(h_{\mathrm{u}}(U_i), h_{\mathrm{x}}(X_i)) + s \sqrt{\frac{1}{n'} \sum_{i=1}^{n'} \|Y'_i - h_{\mathrm{u}}(U'_i)\|_2^2} \times \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|\nabla\phi(h_{\mathrm{u}}(U_i)) - \nabla\phi(h_{\mathrm{x}}(X_i))\|_2^2}. \tag{8}$$

Return $\widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}$.

where we define for $s \in \{-1, +1\}$,

$$B_\phi^\times(h_{\mathrm{x}}, h_{\mathrm{u}}, s) \coloneqq D_\phi(Y, h_{\mathrm{u}}(U)) + D_\phi(h_{\mathrm{u}}(U), h_{\mathrm{x}}(X)) + s \|Y - h_{\mathrm{u}}(U)\|_{L^2} \times \|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|_{L^2}. \tag{7}$$

Notice that each term on the right-hand side of Eq. (7) can be approximated using $\{(X_i, U_i)\}_{i=1}^n$ or $\{(U'_i, Y'_i)\}_{i=1}^{n'}$.

We propose a one-step method, called *Joint-BregMU*, that approximates the upper bound in Lemma 4.2 using MU-data and minimizes the approximated bound as described in Algorithm 1.

4.2.1 Examples with Different Bregman Divergences

We give two examples of the objective function of Joint-BregMU with specific functions ϕ.

Squared Loss: Setting $\phi(t) = \frac{1}{2}\|t\|_2^2$ in the Bregman divergence yields the squared loss: $\ell_\phi(y_1, y_2) = \frac{1}{2}\|y_1 - y_2\|^2$. The bounds used by Joint-BregMU (Eq. (8) in Algorithm 1) will be

$$\widehat{B}_\phi^\times(h_{\mathrm{x}}, h_{\mathrm{u}}, s) = \frac{1}{2} \Bigg(\sqrt{\frac{1}{n'} \sum_{i=1}^{n'} \|Y'_i - h_{\mathrm{u}}(U'_i)\|_2^2} + s \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|h_{\mathrm{u}}(U_i) - h_{\mathrm{x}}(X_i)\|_2^2} \Bigg)^2.$$

Cross-entropy Loss: Define the *cross-entropy loss* as $\ell_{\mathrm{CE}}(y_1, y_2) = -\sum_{j=1}^k [y_1]_j \log [y_2]_j$ for any $y_1 \in [0, 1]^k$ and $y_2 \in (0, 1]^k$ such that $\sum_{j=1}^k [y_1]_j = \sum_{j=1}^k [y_2]_j = 1$, where $[\cdot]_j$ denotes the j-th component of the vector in the argument. Minimizing the expected cross-entropy loss is

equivalent to minimizing the KL-divergence in the following sense:

$$\mathbf{E}[\ell_{\mathrm{CE}}(Y, f(X))] = \mathbf{E}[\underbrace{\ell_\phi(\mathbf{E}[Y \mid X], f(X))}_{\text{KL-divergence}}] - \underbrace{\mathbf{E}\left[\sum_{j=1}^{k} \mathbf{E}[[Y]_j \mid X] \log(\mathbf{E}[[Y]_j \mid X])\right]}_{\text{Constant that does not depend on } f.},$$

where Y is a k-dimensional random variable of a one-hot vector, $f\colon \mathcal{X} \to [0,1]^k$ such that $\sum_{j=1}^k [f(x)]_j = 1$ for all $x \in \mathcal{X}$, and $\phi\colon t \mapsto \sum_{j=1}^k [t]_j \log[t]_j$. Hence, when we have functions $h_{\mathrm{x}}\colon \mathcal{X} \to \mathcal{Y}$ and $h_{\mathrm{u}}\colon \mathcal{U} \to \mathcal{Y}$ whose outputs are positive and normalized, e.g., by a softmax layer, the empirical version of the objective function of Joint-BregMU (Eq. (8) in Algorithm 1) is

$$\begin{aligned}\widehat{B}^{\times}_{\phi}(h_{\mathrm{x}}, h_{\mathrm{u}}, s) = &-\frac{1}{n'}\sum_{i=1}^{n'}\sum_{j=1}^{k}[Y_i']_j \log[h_{\mathrm{u}}(U_i')]_j \\ &+ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}[h_{\mathrm{u}}(U_i)]_j \log\frac{[h_{\mathrm{u}}(U_i)]_j}{[h_{\mathrm{u}}(X_i)]_j} \\ &+ s\sqrt{\frac{1}{n'}\sum_{i=1}^{n'}\sum_{j=1}^{k}([Y_i']_j - [h_{\mathrm{u}}(U_i')]_j)_2^2} \\ &\times \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\log[h_{\mathrm{u}}(U_i)]_j - \log[h_{\mathrm{x}}(X_i)]_j\right)^2} \\ &+ \text{constant}.\end{aligned}$$

4.3 Performance Validation Using Interval Estimates of Test Loss

Predicting whether a trained model is good or bad, is not trivial when we only have MU-data since there is no (X, Y)-data available for comparing the true output Y and the predicted output for X.

Conveniently, Lemma 4.2 provides a way to estimate the test loss in the form of an interval. Let h_{x} and h_{u} be any fixed, already trained models. We can obtain an interval estimate of the expected test loss $D_\phi(Y, h_{\mathrm{x}}(X))$ using the bounds in Eq. (6) approximated only with held-out *MU-data*, without (X, Y)-data:

$$\widetilde{I}^{\times}_{\varepsilon}(h_{\mathrm{x}}, h_{\mathrm{u}}) \coloneqq [\widetilde{B}^{\times}_{\phi}(h_{\mathrm{x}}, h_{\mathrm{u}}, -1) - \varepsilon, \widetilde{B}^{\times}_{\phi}(h_{\mathrm{x}}, h_{\mathrm{u}}, +1) + \varepsilon],$$

where $\widetilde{B}^{\times}_{\phi}(\cdot,\cdot,\cdot)$ is defined similarly to Eq. (8) but calculated using i.i.d. held-out MU-data $\{(\widetilde{X}_i, \widetilde{U}_i)\}_{i=1}^{\widetilde{n}} \sim P_{X,U}$ and $\{\widetilde{U}_i', \widetilde{Y}_i'\}_{i=1}^{\widetilde{n}'} \sim P_{U,Y}$ independent of the training data.

5 THEORETICAL ANALYSIS

In this section, we present upper bounds on the errors of the proposed methods. 2Step-BregMU is weakly consistent but needs stronger assumptions in our analysis. In contrast, Joint-BregMU has a bias but enjoys a bound with a better rate with weaker assumptions. Refer to Section 5.3 for disscussions on the assumptions of our analyses.

5.1 Analysis for 2Step-BregMU

Minimizing the Bregman divergence amounts to estimating the conditional expectation (see Banerjee et al. (2005a, Theorem 1) or Lemma A.2 in the supplementary material). Banerjee et al. (2005a, Theorem 2) further showed that any estimator approximately minimizing the Bregman risk converges in probability to the true conditional expectation.

Based on their results, we can show the weak consistency of our two-step method provided that the estimator of each step is weakly consistent as follows. Suppose that $\widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}$ and $\widehat{h}_{\mathrm{x}}^{\mathrm{2Breg}}$ satisfy, as n tends to infinity,

$$D_\phi(Y, \widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U)) \to D_\phi(Y, \mathbf{E}[Y \mid U]) \tag{9}$$

$$\text{and} \quad D_\phi(\widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U), \widehat{h}_{\mathrm{x}}^{\mathrm{2Breg}}(X)) \to D_\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U), \mathbf{E}[\widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U) \mid X]), \tag{10}$$

which is the case when the models are well-specified, and the function classes are not too complex. Then, $\widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U)$ converges to $\mathbf{E}[Y \mid U]$ and $\widehat{h}_{\mathrm{x}}^{\mathrm{2Breg}}(X)$ to $\mathbf{E}[\widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U) \mid X]$ in probability (Banerjee et al., 2005a, Theorem 2). Thus,

$$\begin{aligned}\widehat{h}_{\mathrm{x}}^{\mathrm{2Breg}}(X) &\xrightarrow{P} \mathbf{E}[\widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U) \mid X] \\ &\xrightarrow{P} \mathbf{E}[\mathbf{E}[Y \mid U, X] \mid X] \\ &= \mathbf{E}[Y \mid X] \quad \text{(Eq. (1))},\end{aligned} \tag{11}$$

where $\xrightarrow{P}$ denotes convergence in probability. Eq. (11) follows from the dominated convergence for conditional expectations (Resnick, 2014, Section 10.3)[*2] and the boundedness of $\mathcal{H}_{\mathrm{x}}$ and $\mathcal{H}_{\mathrm{u}}$.

We present a more refined result on the rate of convergence below. For ease of notation, we denote $\mathcal{H}(v) \coloneqq \{h(v) \mid h \in \mathcal{H}\}$ and $\mathcal{H}(\mathcal{V}) \coloneqq \bigcup_{v\in\mathcal{V}} \mathcal{H}(v)$ for any function class $\mathcal{H}$, any subset $\mathcal{V}$ of the domain of the functions, and any $v \in \mathcal{V}$, whenever they are well-defined.

Theorem 5.1 (Excess risk bound for 2Step-BregMU, in an asymptotic form)**.** *Assume that (i) Eq.* (1) *holds; (ii)* $C_{out} \coloneqq \sup_{y\in\mathcal{Y}\cup\mathcal{H}_{\mathrm{x}}(\mathcal{X})\cup\mathcal{H}_{\mathrm{u}}(\mathcal{U})}\|y\|_2 < \infty$ *and* $C_{loss} \coloneqq \sup_{(y_1,y_2)\in(\mathcal{Y}\cup\mathcal{H}_{\mathrm{x}}(\mathcal{X})\cup\mathcal{H}_{\mathrm{u}}(\mathcal{U}))^2} \ell_\phi(y_1, y_2) < \infty$. *(iii)* $\mathbf{E}[Y \mid U] \in \mathcal{H}_{\mathrm{u}}(U)$ *and* $\mathbf{E}[Y \mid X] \in \mathcal{H}_{\mathrm{x}}(X)$*; (iv)* ϕ *restricted to*

[*2] We take sub-sequences converging almost surely before applying the dominated convergence mentioned in Resnick (2014). See Appendix K of the supplementary material.

$\mathcal{Y} \cup \mathcal{H}_{\mathrm{x}}(\mathcal{X}) \cup \mathcal{H}_{\mathrm{u}}(\mathcal{U})$ is Lipschitz continuous; and (v) for some $C_1, C_2 > 0$, some $p, q \in [1, \infty]$, and some $\alpha \in [1, \infty)$ and $\beta \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} \leq 1$, $\frac{1}{\alpha} + \frac{1}{\beta} \leq 1$, and any $g_1, g_2 \in \mathcal{H}_{\mathrm{u}}(U) \cup \mathcal{H}_{\mathrm{x}}(X)$ satisfy

$$C_1 \|g_1 - g_2\|_{L^p}^{\alpha} \leq D_\phi(g_1, g_2)$$
$$\text{and} \quad \|\nabla_1 \ell_\phi(g_1(\cdot), g_2(\cdot))\|_{L^q}^{\beta} \leq C_2 D_\phi(g_1, g_2),$$

where ∇_1 denotes the gradient operator with respect to the first argument. Then,

$$\begin{aligned} &D_\phi(\mathbf{E}[Y \mid X], \widehat{h}_{\mathrm{x}}^{\mathrm{2Breg}}(X)) \\ &\leq \mathcal{O}_P\Bigg(\left(R_1(n', \phi, \mathcal{H}_{\mathrm{u}}) + \sqrt{\frac{1}{n'}} \right)^{\kappa} \\ &\quad + \left(R_2(n, \phi, \mathcal{H}_{\mathrm{u}}, \mathcal{H}_{\mathrm{x}}) + \sqrt{\frac{1}{n}} \right) \\ &\quad + \left(R_1(n', \phi, \mathcal{H}_{\mathrm{u}}) + \sqrt{\frac{1}{n'}} \right)^{\alpha^{-1}\kappa} \\ &\quad\quad \times \left(R_2(n, \phi, \mathcal{H}_{\mathrm{u}}, \mathcal{H}_{\mathrm{x}}) + \sqrt{\frac{1}{n}} \right)^{\beta^{-1}} \Bigg), \end{aligned}$$

where $\kappa := \alpha^{-1}(1 - \beta^{-1})^{-1} \in (0, 1]$, and $R_l(\dots)$ $(l = 1, 2)$ are model complexity terms that depend on the arguments, and $\mathcal{O}_P(\cdot)$ denotes the stochastic big-O notation.

Theorem 5.1 is an asymptotic summary of our finite-sample result which can be found with a proof in Appendix B of the supplementary material. The precise definitions of $R_l(\cdots)$'s are also in Appendix B.

The upper bound of Theorem 5.1 is at best $\mathcal{O}_P(\sqrt{1/n'} + \sqrt{1/n})$ when $\kappa = 1$, $R_1(n', \phi, \mathcal{H}_{\mathrm{u}}) = \mathcal{O}_P(\sqrt{1/n'})$, and $R_2(n, \phi, \mathcal{H}_{\mathrm{u}}, \mathcal{H}_{\mathrm{x}}) = \mathcal{O}_P(\sqrt{1/n})$. κ depends on α and β in Assumption (v), hence on the loss. For instance, we can show that the squared loss achieves $\kappa = 1$, but we are not aware whether the cross-entropy loss admits $\kappa = 1$. See Section 5.3 for more discussions on the assumption.

A sketch of the proof of Theorem 5.1 goes as follows. First, we upper-bound $D_\phi(\mathbf{E}[Y \mid X], \widehat{h}_{\mathrm{x}}^{\mathrm{2Breg}}(X))$ using some discrepancy measure between $\mathbf{E}[Y \mid U]$ and $\widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U)$ and another between $\widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U)$ and $\widehat{h}_{\mathrm{x}}^{\mathrm{2Breg}}(X)$. The discrepancies involve $\|\widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U) - \mathbf{E}[Y \mid U]\|_{L^p}$ and $\|\nabla \ell_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U) \mid X])\|_{L^q}$. Assumption (v) allows us to bound them in terms of $D_\phi(\mathbf{E}[Y \mid X], \widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U))$ and $D_\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}^{\mathrm{2Breg}}(U) \mid X], \widehat{h}_{\mathrm{x}}^{\mathrm{2Breg}}(X))$. Finally, we can associate those quantities with the minimized training objectives of 2Step-BregMU thorough the emirical process theory.

5.2 Analysis for Joint-BregMU

Next, we present the bound for Joint-BregMU. The bound does not require Assumption (v) in contrast to that of Theorem 5.1.

Theorem 5.2 (Excess risk bound for Joint-BregMU, in an asymptotic form)**.** *Assume the conditions (i)–(iv) of Theorem 5.1 and that $\inf_{h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \mathbf{E}[\|Y - h_{\mathrm{u}}(U)\|_2^2] > 0$ and $\inf_{h_{\mathrm{x}} \in \mathcal{H}_{\mathrm{x}}, h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \mathbf{E}[\|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|_2^2] > 0$. Then,*

$$\begin{aligned} &D_\phi(\mathbf{E}[Y \mid X], \widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}(X)) \\ &\leq \underbrace{e_{n',n}}_{\text{vanishing error}} + 2 \underbrace{\|Y - \mathbf{E}[Y \mid U]\|_{L^2}}_{\text{1st bias factor}} \\ &\qquad \times \underbrace{\|\nabla\phi(\mathbf{E}[Y \mid U]) - \nabla\phi(\mathbf{E}[Y \mid X])\|_{L^2}}_{\text{2nd bias factor}}, \end{aligned}$$

where

$$\begin{aligned} e_{n,n'} \leq \mathcal{O}_P\Bigg(&R_3(n', \phi, \mathcal{H}_{\mathrm{u}}) + R_4(n, \phi, \mathcal{H}_{\mathrm{u}}, \mathcal{H}_{\mathrm{x}}) \\ &+ \sqrt{\frac{1}{n'}} + \sqrt{\frac{1}{n}} \Bigg), \end{aligned}$$

and $R_l(\cdots)$ $(l = 3, 4)$ are model complexity terms that depend on the arguments.

A finite sample version of the bound, its proof, and the precise definitions of $R_l(\cdots)$'s are in Appendix C of the supplementary material.

Notice that the rate of the upper bound in Theorem 5.2 can be faster than that in Theorem 5.1 when $\kappa < 1$. We also analyzed the asymptotic bias of Joint-BregMU as follows.

To illustrate the result, let us consider a one-dimensional, linear example. Let X be a real-valued random variable, $Y = a_{\mathrm{y}} U + b_{\mathrm{y}} + \varepsilon_{\mathrm{y}}$, and $U = a_{\mathrm{u}} X + b_{\mathrm{u}} + \varepsilon_{\mathrm{u}}$, where $a_{\mathrm{y}}, b_{\mathrm{y}}, a_{\mathrm{u}}, b_{\mathrm{u}}$ are constant real numbers, $\varepsilon_{\mathrm{y}}, \varepsilon_{\mathrm{u}}$ are independent centered normal variables with variance σ_{y} and σ_{u}, respectively, and $\phi\colon t \mapsto \frac{1}{2}t^2$. Then, the asymptotic bias of Joint-BregMU is $4 a_{\mathrm{y}} \sigma_{\mathrm{y}} \sigma_{\mathrm{u}}$. This indicates that the bias will be small when $\sigma_{\mathrm{y}} \sigma_{\mathrm{u}}$ is small. In particular, the bias will be zero when either σ_{y} or σ_{u} is zero.

5.3 Discussions on the Assumptions

In this section, we discuss the assumptions used in the paper.

5.3.1 The conditional mean independence (1)

The assumption states about how informative U is, and it can be easy or hard to satisfy depending on the cost of collecting such data. Yamane et al. (2021) assumed the same assumption, and they proved a mini-max lower bound showing that the worst-case L^2 error is at least $\epsilon/\sqrt{2}$ when the assumption is relaxed as $\|\mathbf{E}[Y|U] - \mathbf{E}[Y|U, X]\|_{L^2} \leq \epsilon$ (Yamane et al., 2021, Section 5.5). Intuitively, there is a trade-off between the bias and the violation of the assumption, and if we do not allow bias, $\epsilon = 0$ (i.e., Eq. (1)) is

necessary for any estimator. Note that Eq. (1) only concerns the conditional expectation, which is weaker than the conditional independence.

5.3.2 Assumptions (i-v)

Assumption (iv) is the Lipschitz-continuity (i.e., the boundedness of the gradient and hence the sensitivity) of ϕ. Note that Assumption (iv) is only required on a restricted domain and is relatively easy to satisfy when the domain is bounded as we assume in the paper. The first condition of Assumption (v) is about the strength of the convexity: larger p means stronger convexity. The second condition of (v) is about the strength of the smoothness: larger q means stronger smoothness. Those conditions are orthogonal to each other and do not contradict each other. Note that the result for the one-step method does not require Assumption (v).

For example, the KL-divergence, the squared loss, and $D_{t\mapsto\|t\|^4}(\cdot,\cdot)$ satisfy the Lipschitz-continuity and Assumption (v) with $(p,q,\alpha,\beta)=(1,\infty,2,\infty)$, $(p,q,\alpha,\beta)=(2,2,2,2)$, and $(p,q,\alpha,\beta)=(4,4,4,4)$, respectively. See Appendix J.3 for more details. The boundedness of the function classes is critical for these examples.

5.3.3 Difference between our assumptions and those of Yamane et al. (2021)

Assumptions (i-iv) are commonly assumed in Yamane et al. (2021) and our paper. Our analysis of the two-step method additionally assumes Assumption (v) to bound the target risk using the two objectives minimized in the two steps. For the one-step method, we assume that $Y-h_{\mathrm{u}}(U)$ and $\nabla\phi(h_{\mathrm{u}}(U))-\nabla\phi(h_{\mathrm{x}}(X))$ are not almost surely zero. This is satisfied when the compared terms do not have deterministic relationships. Otherwise, we may add very small random noise to the variables to ensure the conditions.

Our assumptions are weaker than Yamane et al. (2021) in which the loss function must be the squared loss. In fact, all of our results apply to the squared loss.

5.4 Analysis of the Interval Estimates of the Test Loss

As the following theorem states, the proposed test loss interval $\widetilde{I}_{\varepsilon}(h_{\mathrm{x}},h_{\mathrm{u}})$ includes the true test loss with high probability with a slack parameter ε tending to zero.

Theorem 5.3. *Assume that* $C_{out} \coloneqq \sup_{y\in\mathcal{Y}\cup\mathcal{H}_{\mathrm{x}}(\mathcal{X})\cup\mathcal{H}_{\mathrm{u}}(\mathcal{U})}\|y\|_2 < \infty$, $C_{loss} \coloneqq \sup_{(y_1,y_2)\in(\mathcal{Y}\cup\mathcal{H}_{\mathrm{x}}(\mathcal{X})\cup\mathcal{H}_{\mathrm{u}}(\mathcal{U}))^2}\ell_\phi(y_1,y_2) < \infty$, $\inf_{h_{\mathrm{u}}\in\mathcal{H}_{\mathrm{u}}}\mathbf{E}[\|Y-h_{\mathrm{u}}(U)\|_2^2] > 0$, *and* $\inf_{h_{\mathrm{x}}\in\mathcal{H}_{\mathrm{x}},h_{\mathrm{u}}\in\mathcal{H}_{\mathrm{u}}}\mathbf{E}[\|\nabla\phi(h_{\mathrm{u}}(U))-\nabla\phi(h_{\mathrm{x}}(X))\|_2^2] > 0$. *For any measurable functions* $h_{\mathrm{x}}\colon\mathcal{X}\to\mathcal{Y}$, $h_{\mathrm{u}}\colon\mathcal{U}\to\mathcal{Y}$, *any* $\widetilde{n},\widetilde{n}'\in\mathbb{N}$ *sufficiently large, and any* $\delta>0$, *with probability at least* $1-\delta$, *we have*

$$D_\phi(Y,h_{\mathrm{x}}(X))\in\widetilde{I}^{\times}_{\varepsilon}(h_{\mathrm{x}},h_{\mathrm{u}})$$

where $\varepsilon \coloneqq (\sqrt{2}C_{\mathrm{loss}}+2L_\phi^2C_{\mathrm{out}}^{5/2})\sqrt{\frac{1}{\widetilde{n}}\log\frac{16}{\delta}}+(\sqrt{2}C_{\mathrm{loss}}+2L_\phi^{1/2}C_{\mathrm{out}}^{5/2})\sqrt{\frac{1}{\widetilde{n}'}\log\frac{16}{\delta}}$.

The proof is in Appendix L of the supplementary material.

We could obtain a similar interval estimate based on the approach of Yamane et al. (2021):

$$\widetilde{I}^{+}_{\varepsilon}(h_{\mathrm{x}},h_{\mathrm{u}}) \coloneqq [\widetilde{B}^{+}_{\phi}(h_{\mathrm{x}},h_{\mathrm{u}},-1)-\varepsilon,\widetilde{B}^{+}_{\phi}(h_{\mathrm{x}},h_{\mathrm{u}},+1)+\varepsilon],$$

where we define $\widetilde{B}^{+}_{\phi}(h_{\mathrm{x}},h_{\mathrm{u}},s)$ as

$$\begin{aligned}&\frac{1}{\widetilde{n}'}\sum_{i=1}^{\widetilde{n}'}\ell_\phi(\widetilde{Y}'_i,h_{\mathrm{u}}(\widetilde{U}'_i))+\frac{1}{\widetilde{n}}\sum_{i=1}^{\widetilde{n}}\ell_\phi(h_{\mathrm{u}}(\widetilde{U}_i),h_{\mathrm{x}}(\widetilde{X}_i))\\&\quad+\frac{s}{2}\Big(\frac{1}{\widetilde{n}'}\sum_{i=1}^{\widetilde{n}'}\|\widetilde{Y}_i-h_{\mathrm{u}}(\widetilde{U}_i)\|_2^2\\&\qquad+\frac{1}{\widetilde{n}}\sum_{i=1}^{\widetilde{n}}\|\nabla\phi(h_{\mathrm{u}}(\widetilde{U}_i))-\nabla\phi(h_{\mathrm{x}}(\widetilde{X}_i))\|_2^2\Big),\end{aligned}$$

where $\{(\widetilde{X}_i,\widetilde{U}_i)\}_{i=1}^{\widetilde{n}}\sim P_{X,U}$ and $\{\widetilde{U}'_i,\widetilde{Y}'_i\}_{i=1}^{\widetilde{n}'}\sim P_{U,Y}$ are held-out samples independent of the training data. $\widetilde{B}^{+}_{\phi}(h_{\mathrm{x}},h_{\mathrm{u}},+1)$ coincides with the objective function of Joint-RR (Eq. (2)) times $1/2$ when $\phi(t)=\frac{1}{2}\|t\|_2^2$.

However, we show in Proposition 5.1 that the proposed interval $\widetilde{I}^{\times}_{\varepsilon}(\cdot,\cdot)$ is tighter than $\widetilde{I}^{+}_{\varepsilon}(\cdot,\cdot)$.

Proposition 5.1. *For any* $\varepsilon>0$, *any* $h_{\mathrm{x}}\colon\mathcal{X}\to\mathcal{Y}$, *and any* $h_{\mathrm{u}}\colon\mathcal{U}\to\mathcal{Y}$, *we have* $\widetilde{I}^{\times}_{\varepsilon}(h_{\mathrm{x}},h_{\mathrm{u}})\subseteq\widetilde{I}^{+}_{\varepsilon}(h_{\mathrm{x}},h_{\mathrm{u}})$.

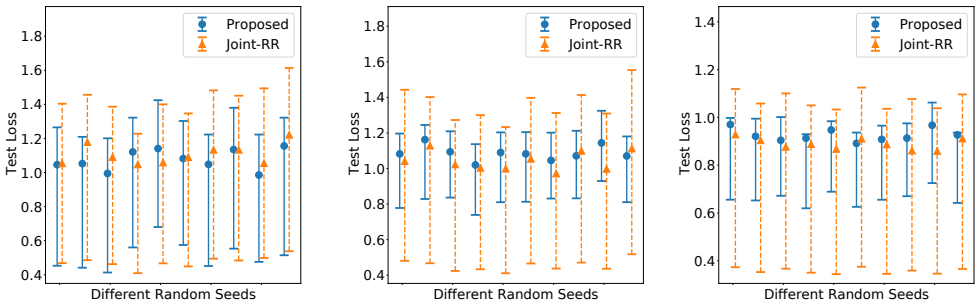
Proposition 5.1 combined with Theorem 5.3 implies that $\widetilde{I}^{\times}_{\varepsilon}(\cdot,\cdot)$ is narrower and more precise than $\widetilde{I}^{+}_{\varepsilon}(\cdot,\cdot)$. This result also implies that the objective function of Joint-BregMU is a tighter approximation to the oracle loss $D_\phi(Y,h_{\mathrm{x}})$ than that of Joint-RR of Yamane et al. (2021).

6 EXPERIMENTS

In this section, we present experimental results.

6.1 Error Interval Estimation in Regression Problems

We conducted experiments of regression problems to compare the proposed interval estimator and that based on the existing approach (Yamane et al., 2021). Similarly to Yamane et al. (2021), we consider the following setup for (X,U,Y): X follows the uniform distribution over $[-1,1]^{10}$, $[U]_j=[X]_j^3+[\varepsilon_{\mathrm{u}}]_j$ for all $j\in\{1,\dots,10\}$, $Y=\|U\|_2^2+\varepsilon_{\mathrm{y}}$, $\varepsilon_{\mathrm{u}}\sim\mathcal{N}(0,0.5I_{10})$, and $\varepsilon_{\mathrm{y}}\sim\mathcal{N}(0,0.5)$. Recall that $[\cdot]_j$ denotes the j-th element of the vector in the argument. We generate independent MU-data $\{(X_i,U_i)\}_{i=1}^{n}$ and $\{(U'_i,Y'_i)\}_{i=1}^{n'}$



(a) $n = n' = 1000$. (b) $n = n' = 5000$. (c) $n = n' = 10000$.

Figure 1: Interval estimates of the test loss for the synthetic data experiments. ε is set to zero. The blue circles and triangles are the test losses of the proposed method and Joint-RR (Yamane et al., 2021), respectively, evaluated using test (X, Y)-data, which are not accessible in the MU-learning setup. The bars indicate the interval estimates of the test losses calculated using validation MU-data. The horizontal axis is for different random seeds used to generate the data.

identically distributed to (X, U) and (U, Y), respectively. The evaluation metric is $D_\phi(\cdot, \cdot)$ with $\phi(t) = \frac{1}{2}\|t\|_2^2$, i.e., the mean squared error (MSE), so that we can directly apply Joint-RR (Yamane et al., 2021) (see Section 3.3). We train models with the proposed Joint-BregMU (see Section 4.2 and 4.2.1) and Joint-RR with different sizes of MU-data, $n = n' \in \{1000, 2000, 3000, 5000, 10000\}$. For the trained models, we compare the interval estimates given by $\widehat{B}_\phi^\times(\cdot, \cdot, \cdot)$ and $\widehat{B}_\phi^+(\cdot, \cdot, \cdot)$, respectively, calculated with validation MU-data. We implemented the methods based on the code of Yamane et al. (2021)[*3]. More details can be found in the supplemental material.

Table 1 shows the average test loss of the two methods. The performances for Joint-RR and the proposed method are almost comparable, but Joint-RR tends to be slightly more advantageous for the largest sample size. This may be because the proposed tighter approximation does not always lead to a better training, and the bias induced by Joint-RR may be preferable to this specific experiment.

Figure 1 shows the results for the interval estimation for $n = n' \in \{1000, 5000, 10000\}$. More results are in the supplemental material. For both methods, the intervals successfully include the test loss in all cases. However, the ones produced by the proposed method tend to be more precise in terms of the sizes of the intervals. The proposed interval estimate becomes drastically narrower as the sample size grows compared with the interval estimate with Joint-RR. This matches the theoretical result in Proposition 5.1.

[*3]Our code and that of Yamane et al. (2021) are both available at `https://github.com/i-yamane/mediated_uncoupled_learning`

6.2 Classification of Low-Resolution Images

Next, we compare the performance of our proposed methods with the existing methods on classification of low-resolution images (Yamane et al., 2021). The task is to learn a function for classifying low-resolution images only using MU-data in training. The MU-data consist of a set of labeled high-resolution images and a set of pairs of high-resolution and low-resolution images.

The motivation behind this setup is that we want to perform predictions with small devices that can only afford low-resolution cameras in the deployment environment, but high-resolution images may be easier for human labelers to classify, and thus we may be able to collect labeled high-resolution images with relatively lower cost. In this case, to obtain information necessary to fill the gap of low- and high-resolution images, we also collect pairs of them. The biggest advantage of this approach is that we only need to collect labels with high-resolution images once even when we need to adjust the deployment resolution to another one or handle different resolutions at the same time.

We use image benchmark datasets prepared for standard classification but modify images to artificially create low-resolution images. More specifically, for each image and

Table 1: Results for the synthetic data experiment. The scores are the average MSEs calculated from 10 repetitions of the experiment (and standard errors). The scores comparable to the best in terms of Wilcoxon's signed rank test with significance level 5% are emphasized in bold fonts.

$n\ (= n')$	1000	2000	3000	5000	10000
Joint-RR	**1.10** (0.01)	**1.23** (0.02)	**1.20** (0.02)	**1.04** (0.01)	**0.89** (0.00)
Joint-BregMU	**1.07** (0.01)	**1.22** (0.02)	**1.19** (0.01)	**1.08** (0.01)	0.92 (0.00)

its class label in each benchmark dataset, we define X as a down-sampled image, U as the original image, and Y as the class label. We take subsamples of (X, U) to define $\{(X_i, U_i)\}_{i=1}^{n}$ and of (U, Y) to define $\{(U'_i, Y'_i)\}_{i=1}^{n'}$.

The task being classification, we use the zero-one loss as the test evaluation metric. For training, we use the cross-entropy (Section 4.2.1) as the surrogate loss for the proposed methods but the squared loss for the previous methods because of its limitation. We implemented the methods based on the code provided by Yamane et al. (2021). More details and our code can be found in the supplemental material.

Table 2 shows the results for the low-quality image classification experiment. We can see that the performance of Joint-BregMU is consistently among the best for all datasets. The previous methods 2Step-RR and Joint-RR perform especially well for MNIST, the dataset with the least uncertainty. However, with more uncertainty, Joint-BregMU become more advantageous and performs significantly better than the other methods. 2Step-BregMU did not show as good performance as that of Joint-BregMU in these experiments. This might be because of the advantage of Joint-BregMU in convergence as Theorems 5.1 and 5.2 suggest.

7 CONCLUSION

We proposed two MU-learning methods that can handle a wide range of loss functions defined with Bregman divergences. One is a two-step method, each step of which is reduced to simple, ordinary supervised learning. The other one is a one-step method that has an appealing convergence property and empirical performance. The one-step method also yields a provably tighter interval estimate of test loss compared with the existing approach (Yamane et al., 2021). The current analyses rely on the assumptions which may not be satisfied in some cases. Also, when the loss is strongly convex, the analysis may not provide the optimal rates. Future work includes relaxing those assumptions and improving the rates. MU-learning could be seen as a way to recover couplings of attributes reavealed independently. This might seem to raise privacy issues when applied to combinations of sensitive attributes such as face images and regions of birth. However, MU-learning can only recover statistical associations but cannot recover individual-level couplings. Further investigating this topic using the theory of differential privacy may be an important direction of future research.

Acknowledgements

We thank the anonymous reviewers for their helpful comments that we received for the current and the past version of this paper. 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). TI was supported by KAKENHI Grant Number 22K17946.

References

Amid, E., Warmuth, M. K., Anil, R., and Koren, T. (2019). Robust Bi-Tempered Logistic Loss Based on Bregman Divergences. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Banerjee, A., Guo, X., and Wang, H. (2005a). On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005b). Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6(58):1705–1749.

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

Dhillon, I. S. and Tropp, J. A. (2008). Matrix nearness problems with Bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146.

Haddow, B., Bawden, R., Barone, A. V. M., Helcl, J., and

Table 2: The results for the experiment on classification of low-quality images. Each cell shows the average (and the standard error) of classification accuracy rates calculated over 10 repetitions of the experiment for the corresponding dataset and the method. The scores comparable to the best in terms of Wilcoxon's signed rank test with significance level 5% are emphasized in bold fonts. The bottom of the table shows the result for training the same model h_{x} as that of Joint-BregMU with the cross-entropy loss using (X, Y)-data (OracleXY) for a reference. OracleXY is excluded from the comparison since it does not perform MU-learning.

	MNIST	FashionMNIST	CIFAR10	CIFAR100
Naive	0.878 (0.016)	0.705 (0.020)	0.453 (0.007)	0.076 (0.002)
2Step-RR	**0.977** (0.001)	0.877 (0.003)	0.704 (0.002)	0.275 (0.001)
Joint-RR	**0.976** (0.002)	**0.885** (0.001)	**0.687** (0.031)	0.279 (0.002)
2Step-BregMU	0.972 (0.002)	0.871 (0.002)	0.694 (0.003)	0.317 (0.002)
Joint-BregMU	**0.972** (0.003)	**0.884** (0.002)	**0.718** (0.002)	**0.324** (0.002)
OracleXY	0.964 (0.005)	0.893 (0.002)	0.777 (0.001)	0.373 (0.002)

Birch, A. (2022). Survey of Low-Resource Machine Translation. *arXiv:2109.00486 [cs]*. arXiv: 2109.00486.

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

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

Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 3020–3029, New York, New York, USA. PMLR.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

LeCun, Y., Cortes, C., and Burges, C. (2010). MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

Linder-Norén, E. (2018). Pytorch-gan. `https://github.com/eriklindernoren/PyTorch-GAN`. Accessed: 2022-5-20.

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

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

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

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

Nock, R., Menon, A., and Ong, C. S. (2016). A scaled bregman theorem with applications. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Resnick, S. I. (2014). *A Probability Path*. Birkhäuser Boston, Boston, MA.

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

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

Siahkamari, A., XIA, X., Saligrama, V., Castañón, D., and Kulis, B. (2020). Learning to Approximate a Bregman Divergence. In *Advances in Neural Information Processing Systems*, volume 33, pages 3603–3612. Curran Associates, Inc.

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

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). 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*, pages 2048–2057. PMLR.

Yamane, I., Honda, J., Yger, F., and Sugiyama, M. (2021). Mediated uncoupled learning: Learning functions without direct input-output correspondences. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11637–11647. PMLR.

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

Zou, H., Cui, P., Li, B., Shen, Z., Ma, J., Yang, H., and He, Y. (2020). Counterfactual prediction for bundle treatment. In *Advances in Neural Information Processing Systems*, volume 33, pages 19705–19715.

A Properties of Bregman Divergences

Below is a well-known fact about Bregman divergences.

Lemma A.1 (Decomposition with the conditional expectation)**.** *For any $y \in \mathcal{Y}$, it holds that*

$$D_\phi(Y, y) = D_\phi(Y, \mathbf{E}[Y \mid X]) + D_\phi(\mathbf{E}[Y \mid X], y). \tag{12}$$

Proof.

$$\begin{aligned}
&D_\phi(Y, y) - D_\phi(Y, \mathbf{E}[Y \mid X]) \\
&= \mathbf{E}[\phi(\mathbf{E}[Y \mid X]) - \phi(y) - (Y - y)^\top \nabla\phi(y) + (Y - \mathbf{E}[Y \mid X])^\top \nabla\phi(\mathbf{E}[Y \mid X])] \\
&= \mathbf{E}[\phi(\mathbf{E}[Y \mid X]) - \phi(y) - (\mathbf{E}[Y \mid X] - y)^\top \nabla\phi(y) + (\mathbf{E}[Y \mid X] - \mathbf{E}[Y \mid X])^\top \nabla\phi(\mathbf{E}[Y \mid X])] \\
&= \mathbf{E}[\phi(\mathbf{E}[Y \mid X]) - \phi(y) - (\mathbf{E}[Y \mid X] - y)^\top \nabla\phi(y)] \\
&= D_\phi(\mathbf{E}[Y \mid X], y).
\end{aligned}$$

□

Banerjee et al. (2005a, Theorem 1) proved that the conditional expectation is the minimizer of any Bregman divergence. We restate their result using our notation below.

Lemma A.2 (Theorem 1 of Banerjee et al. (2005a))**.**

$$\min_{h_x \in \mathcal{H}_\mathrm{x}} D_\phi(Y, h_x(X)) = D_\phi(Y, \mathbf{E}[Y \mid X])$$

when $\mathbf{E}[Y \mid X] \in \mathcal{H}_\mathrm{x}(X)$.

Proof. This is a corollary of Lemma A.1. □

B Excess Risk Bound for 2Step-BregMU

Theorem B.1 (Excess risk bound for 2Step-BregMU, in an asymptotic form)**.** *Assume the following conditions.*

(i) *Eq.* (1) *holds.*

(ii) $C_{out} \coloneqq \sup_{y \in \mathcal{Y} \cup \mathcal{H}_\mathrm{x}(\mathcal{X}) \cup \mathcal{H}_\mathrm{u}(\mathcal{U})} \|y\|_2 < \infty$*, and* $C_{loss} \coloneqq \sup_{(y_1, y_2) \in (\mathcal{Y} \cup \mathcal{H}_\mathrm{x}(\mathcal{X}) \cup \mathcal{H}_\mathrm{u}(\mathcal{U}))^2} \ell_\phi(y_1, y_2) < \infty$.

(iii) $\mathbf{E}[Y \mid U] \in \mathcal{H}_\mathrm{u}(U)$ *and* $\mathbf{E}[Y \mid X] \in \mathcal{H}_\mathrm{x}(X)$.

(iv) ϕ *restricted to* $\mathcal{Y} \cup \mathcal{H}_\mathrm{x}(\mathcal{X}) \cup \mathcal{H}_\mathrm{u}(\mathcal{U})$ *is* L_ϕ*-Lipschitz continuous for some* $L_\phi > 0$.

(v) *For some* $C_1, C_2 > 0$, $p, q \in [1, \infty]$, $\alpha \in [1, \infty)$, *and* $\beta \in [1, \infty]$ *such that* $\frac{1}{p} + \frac{1}{q} \leq 1$, $\frac{1}{\alpha} + \frac{1}{\beta} \leq 1$, *and any* $g_1, g_2 \in \mathcal{H}_\mathrm{u} \circ U \cup \mathcal{H}_\mathrm{x} \circ X$ *satisfy*

$$\int [\phi(g_1) - \phi(g_2)] \mathrm{d}P \geq \int (g_1 - g_2)^\top \nabla\phi(g_2) \mathrm{d}P + C_1 \|g_1 - g_2\|_{L^p}^\alpha, \tag{13}$$

$$\|\nabla\phi(g_1) - \nabla\phi(g_2)\|_{L^p}^\beta \leq C_2 D_\phi(g_1, g_2). \tag{14}$$

These are equivalent to saying

$$C_1 \|g_1 - g_2\|_{L^p}^\alpha \leq D_\phi(g_1, g_2), \text{ and} \tag{15}$$

$$\|\nabla_1 \ell_\phi(g_1(\cdot), g_2(\cdot))\|_{L^q}^\beta \leq C_2 D_\phi(g_1, g_2), \tag{16}$$

where ∇_1 *denotes the gradient operator with respect to the first argument. When* $\alpha = \infty$*, we mean* $C_1 \|g_1 - g_2\|_{L^p} \leq 1$ *by Eq.* (15)*. When* $\beta = \infty$*, we mean* $\|\nabla_1 \ell_\phi(g_1(\cdot), g_2(\cdot))\|_{L^q} \leq C_2$ *by Eq.* (16)*.*

Then,

$$
\begin{aligned}
&D_\phi(\mathbf{E}[Y \mid X], \widehat{h}_x(X)) \\
&\leq C_1^{-\eta} C_2^{\frac{1}{\beta-1}} \left(R_1(n', \phi, \mathcal{H}_\mathrm{u}) + 2C_{loss}\sqrt{\frac{2}{n'}\log\frac{1}{\delta}} \right)^{\eta} + \left(R_2(n, \phi, \mathcal{H}_\mathrm{x}, \mathcal{H}_\mathrm{u}) + 2C_{loss}\sqrt{\frac{2}{n}\log\frac{1}{\delta}} \right) \qquad (17) \\
&\quad + C_1^{\frac{(1-\eta)}{\alpha}} C_2^{\frac{1+\eta}{\beta}} \left(R_1(n', \phi, \mathcal{H}_\mathrm{u}) + 2C_{loss}\sqrt{\frac{1}{n}\log\frac{1}{\delta}} \right)^{\frac{\eta}{\alpha}} \left(R_2(n', \phi, \mathcal{H}_\mathrm{x}, \mathcal{H}_\mathrm{u}) + 2C_{loss}\sqrt{\frac{2}{n}\log\frac{1}{\delta}} \right)^{\frac{1}{\beta}}, \qquad (18)
\end{aligned}
$$

and in any asymptotic form,

$$
\begin{aligned}
D_\phi(\mathbf{E}[Y \mid X], \widehat{h}_x(X)) \leq \mathcal{O}_P\Bigg(&\left(R_1(n', \phi, \mathcal{H}_\mathrm{u}) + \sqrt{\frac{1}{n'}} \right)^{\eta} + \left(R_2(n, \phi, \mathcal{H}_\mathrm{u}, \mathcal{H}_\mathrm{x}) + \sqrt{\frac{1}{n}} \right) \\
&+ \left(R_1(n', \phi, \mathcal{H}_\mathrm{u}) + \sqrt{\frac{1}{n'}} \right)^{\alpha^{-1}\eta} \times \left(R_2(n, \phi, \mathcal{H}_\mathrm{u}, \mathcal{H}_\mathrm{x}) + \sqrt{\frac{1}{n}} \right)^{\beta^{-1}} \Bigg),
\end{aligned}
$$

where $\eta \coloneqq \alpha^{-1}(1-\beta^{-1})^{-1} \in (0, 1]$, $R_l(\dots)$ $(l = 1, 2)$ are model complexity terms that depend on the arguments, and we denote $f(n) \leq \mathcal{O}_P(g(n))$ for non-negative functions f, g of $n \in \mathbb{N}$ if and only if for all $\epsilon > 0$, there exist $C > 0$ and $N_0 \in \mathbb{N}$ such that for all $n \geq N$, $1 - \varepsilon \leq P\left[f(n) \leq Cg(n)\right]$.

Proof. Let

$$
\begin{aligned}
\widehat{h}_\mathrm{u} &\coloneqq \operatorname*{arg\,min}_{h_\mathrm{u} \in \mathcal{H}_\mathrm{u}} \frac{1}{n'} \sum_{i=1}^{n'} \ell_\phi(Y'_i, h_\mathrm{u}(U'_i)), \\
\widehat{h}_\mathrm{x} &\coloneqq \operatorname*{arg\,min}_{h_\mathrm{x} \in \mathcal{H}_\mathrm{x}} \frac{1}{n} \sum_{i=1}^{n} \ell_\phi(h_\mathrm{u}(U_i), h_\mathrm{x}(X_i)).
\end{aligned}
$$

Reduction to Two Regression Analyses. From Lemma A.1, the excess risk can be expressed as

$$
(\text{Excess risk of } \widehat{h}_\mathrm{x}) \equiv D_\phi(Y, \widehat{h}_\mathrm{x}(X)) - D_\phi(Y, \mathbf{E}[Y \mid X]) = D_\phi(\mathbf{E}[Y \mid X], \widehat{h}_\mathrm{x}(X)). \qquad (19)
$$

Furthermore,

$$
\begin{aligned}
&D_\phi(\mathbf{E}[Y \mid X], \widehat{h}_\mathrm{x}(X)) \\
&= D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X]) + D_\phi(\mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X], \widehat{h}_\mathrm{x}(X)) \\
&\quad + \mathbf{E}[(\mathbf{E}[Y \mid X] - \mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X])^\top (\nabla\phi(\mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X]) - \nabla\phi(\widehat{h}_\mathrm{x}(X)))] \\
&\leq D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X]) + D_\phi(\mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X], \widehat{h}_\mathrm{x}(X)) \\
&\quad + \|\mathbf{E}[Y \mid X] - \mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X]\|_{L^p} \cdot \|\nabla\phi(\mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X]) - \nabla\phi(\widehat{h}_\mathrm{x}(X))\|_{L^q}.
\end{aligned}
$$

From Eqs. (15) and (16),

$$
\begin{aligned}
&D_\phi(\mathbf{E}[Y \mid X], \widehat{h}_\mathrm{x}(X)) \\
&\leq D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X]) + D_\phi(\mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X], \widehat{h}_\mathrm{x}(X)) \\
&\quad + C_1^{-1/\alpha} C_2^{1/\beta} D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X])^{1/\alpha} \cdot D_\phi(\mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X], \widehat{h}_\mathrm{x}(X))^{1/\beta}. \qquad (20)
\end{aligned}
$$

It suffices to bound each of $D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X])$ (the error of the first regression step) and $D_\phi(\mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X], \widehat{h}_\mathrm{x}(X))$ (the error of the second regression step).

Bound on $D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X])$: Since $\mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X = x]$ is the minimizer of the functional $h_\mathrm{x} \mapsto D_\phi(\widehat{h}_\mathrm{u}(U), h_\mathrm{x}(X))$, we have

$$
D_\phi(\widehat{h}_\mathrm{u}(U), \mathbf{E}[\widehat{h}_\mathrm{u}(U) \mid X]) \leq D_\phi(\widehat{h}_\mathrm{u}(U), \mathbf{E}[Y \mid X]),
$$

and thus

$$
\begin{aligned}
&D_\phi(\widehat{h}_{\mathrm{u}}(U), \mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X]) \\
&= D_\phi(\widehat{h}_{\mathrm{u}}(U), \mathbf{E}[Y \mid X]) + D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X]) \\
&\quad + \mathbf{E}[(\widehat{h}_{\mathrm{u}}(U) - \mathbf{E}[Y \mid X])^\top (\nabla\phi(\mathbf{E}[Y \mid X]) - \nabla\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X]))] \leq D_\phi(\widehat{h}_{\mathrm{u}}(U), \mathbf{E}[Y \mid X]).
\end{aligned}
$$

Rearranging the both sides, we get

$$
\begin{aligned}
&D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X]) \\
&\leq -\mathbf{E}[(\widehat{h}_{\mathrm{u}}(U) - \mathbf{E}[Y \mid X])^\top (\nabla\phi(\mathbf{E}[Y \mid X]) - \nabla\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X]))] \\
&= -\mathbf{E}[\mathbf{E}[\widehat{h}_{\mathrm{u}}(U) - \mathbf{E}[Y \mid U, X] \mid X]^\top (\nabla\phi(\mathbf{E}[Y \mid X]) - \nabla\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X]))] \\
&= -\mathbf{E}[(\widehat{h}_{\mathrm{u}}(U) - \mathbf{E}[Y \mid U])^\top (\nabla\phi(\mathbf{E}[Y \mid X]) - \nabla\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X]))] \\
&= -\mathbf{E}[(\widehat{h}_{\mathrm{u}}(U) - \mathbf{E}[Y \mid U])^\top \nabla\ell_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X])] \\
&\leq \|\widehat{h}_{\mathrm{u}}(U) - \mathbf{E}[Y \mid U]\|_{L^p} \times \|\nabla\ell_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X])\|_{L^q}
\end{aligned}
$$

by Hölder's inequality. From Eqs. (15) and (16), we can further bound the right hand side as

$$
\begin{aligned}
&\|\widehat{h}_{\mathrm{u}}(U) - \mathbf{E}[Y \mid U]\|_{L^p} \times \|\nabla\ell_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X])\|_{L^q} \\
&\leq C_1^{-1/\alpha} C_2^{1/\beta} D_\phi(\widehat{h}_{\mathrm{u}}(U), \mathbf{E}[Y \mid U])^{1/\alpha} \times D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X])^{1/\beta}.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X]) &\leq C_1^{-1/\alpha} C_2^{1/\beta} D_\phi(\mathbf{E}[Y \mid U], \widehat{h}_{\mathrm{u}}(U))^{1/\alpha} \times D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X])^{1/\beta}, \\
D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X])^{1-1/\beta} &\leq C_1^{-1/\alpha} C_2^{1/\beta} D_\phi(\mathbf{E}[Y \mid U], \widehat{h}_{\mathrm{u}}(U))^{1/\alpha}, \\
D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X]) &\leq (C_1^{-1/\alpha} C_2^{1/\beta} D_\phi(\mathbf{E}[Y \mid U], \widehat{h}_{\mathrm{u}}(U))^{1/\alpha})^{(1-1/\beta)^{-1}} \\
&= C_1^{-\alpha^{-1}(1-1/\beta)^{-1}} C_2^{\beta^{-1}(1-1/\beta)^{-1}} D_\phi(\mathbf{E}[Y \mid U], \widehat{h}_{\mathrm{u}}(U))^{\alpha^{-1}(1-1/\beta)^{-1}} \\
&= C_1^{-\eta} C_2^{\beta^{-1}(1-1/\beta)^{-1}} D_\phi(\mathbf{E}[Y \mid U], \widehat{h}_{\mathrm{u}}(U))^{\eta}, \qquad (21)
\end{aligned}
$$

where $\eta \coloneqq \alpha^{-1}(1 - 1/\beta)^{-1}$. Hence, it suffices to bound $D_\phi(\mathbf{E}[Y \mid U], \widehat{h}_{\mathrm{u}}(U))$ to bound $D_\phi(\mathbf{E}[Y \mid X], \mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X])$.

Note that in the bound above, we could have introduced $D_\phi(\widehat{h}_{\mathrm{u}}(U), \mathbf{E}[Y \mid U])$ instead of $D_\phi(\mathbf{E}[Y \mid U], \widehat{h}_{\mathrm{u}}(U))$. We found that our choice will be more convenient in the very last step of this subsection in which Lemma A.1 works nicely together with $D_\phi(\mathbf{E}[Y \mid U], \widehat{h}_{\mathrm{u}}(U))$.

Upper bound on $D_\phi(\mathbf{E}[Y \mid U], \widehat{h}_{\mathrm{u}}(U))$:

with probability at least $1 - \delta$, we have

$$
D_\phi(\mathbf{E}[Y \mid U], \widehat{h}_{\mathrm{u}}(U)) \leq R_1(n', \phi, \mathcal{H}_{\mathrm{u}}) + 2C_{\mathrm{loss}}\sqrt{\frac{2}{n'}\log\frac{1}{\delta}}, \qquad (22)
$$

where

$$
R_1(n', \phi, \mathcal{H}_{\mathrm{u}}) \coloneqq 2\mathfrak{R}(\phi \circ \mathcal{H}_u(S')) + 8(L_\phi + C_{\mathrm{out}}) \sum_{j=1}^{k} \left(\mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_u(S')) + \mathfrak{R}(\mathcal{H}_{\mathrm{u}}(S')) \right). \qquad (23)
$$

Upper bound on $D_\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X], \widehat{h}_{\mathrm{x}}(X))$: From Lemma C.4 and Lemma C.5, with probability at least $1 - \delta$, we have

$$
D_\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}(U) \mid X], \widehat{h}_{\mathrm{x}}(X)) \leq R_2(n, \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) + 2C_{\mathrm{loss}}\sqrt{\frac{2}{n}\log\frac{1}{\delta}}, \qquad (24)
$$

where

$$R_2(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}}) \coloneqq 2\mathfrak{R}(\phi\circ\mathcal{H}_{\mathrm{x}}(S)) + 2\mathfrak{R}(\phi\circ\mathcal{H}_{\mathrm{u}}(S)) \tag{25}$$

$$+ 8(L_\phi + C_{\mathrm{out}})\sum_{j=1}^{k}(\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_{\mathrm{x}}(S)) + \mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_{\mathrm{u}}(S))) \tag{26}$$

$$+ 8(L_\phi + C_{\mathrm{out}})\sum_{j=1}^{k}(\mathfrak{R}(\mathcal{H}_{\mathrm{x}}^{(j)}(S)) + \mathfrak{R}(\mathcal{H}_{\mathrm{u}}^{(j)}(S))). \tag{27}$$

Combining Eqs. (20), (21), (22), and (24), we obtain

$$\begin{aligned}
&D_\phi(\mathbf{E}[Y\mid X],\widehat{h}_{\mathrm{x}}(X))\\
&\le D_\phi(\mathbf{E}[Y\mid X],\mathbf{E}[\widehat{h}_{\mathrm{u}}(U)\mid X]) + D_\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}(U)\mid X],\widehat{h}_{\mathrm{x}}(X))\\
&\quad + C_1^{1/p}C_2^{1/\beta}D_\phi(\mathbf{E}[Y\mid X],\mathbf{E}[\widehat{h}_{\mathrm{u}}(U)\mid X])^{1/p}\cdot D_\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}(U)\mid X],\widehat{h}_{\mathrm{x}}(X))^{1/\beta} && (28)\\
&\le C_1^{-\eta}C_2^{(\beta-1)^{-1}}D_\phi(\mathbf{E}[Y\mid U],\widehat{h}_{\mathrm{u}}(U))^{\eta} + D_\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}(U)\mid X],\widehat{h}_{\mathrm{x}}(X)) && (29)\\
&\quad + C_1^{p^{-1}}C_2^{\beta^{-1}}(C_1^{-\eta}C_2^{\beta^{-1}(1-1/\beta)^{-1}}D_\phi(\mathbf{E}[Y\mid U],\widehat{h}_{\mathrm{u}}(U))^{\eta})^{p^{-1}} && (30)\\
&\qquad \times D_\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}(U)\mid X],\widehat{h}_{\mathrm{x}}(X))^{1/\beta} && (31)\\
&\le C_1^{-\eta}C_2^{(\beta-1)^{-1}}D_\phi(\mathbf{E}[Y\mid U],\widehat{h}_{\mathrm{u}}(U))^{\eta} && (32)\\
&\quad + D_\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}(U)\mid X],\widehat{h}_{\mathrm{x}}(X)) && (33)\\
&\quad + C_1^{\alpha^{-1}(1-\eta)}C_2^{\beta^{-1}(1+\eta)}D_\phi(\mathbf{E}[Y\mid U],\widehat{h}_{\mathrm{u}}(U))^{\alpha^{-1}\eta}\times D_\phi(\mathbf{E}[\widehat{h}_{\mathrm{u}}(U)\mid X],\widehat{h}_{\mathrm{x}}(X))^{\beta^{-1}} && (34)\\
&\le C_1^{-\eta}C_2^{\frac{1}{\beta-1}}\left(R_1(n',\phi,\mathcal{H}_{\mathrm{u}}) + 2C_{\mathrm{loss}}\sqrt{\frac{2}{n'}\log\frac{1}{\delta}}\right)^{\eta} + \left(R_2(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}}) + 2C_{\mathrm{loss}}\sqrt{\frac{2}{n}\log\frac{1}{\delta}}\right) && (35)\\
&\quad + C_1^{\frac{(1-\eta)}{\alpha}}C_2^{\frac{1+\eta}{\beta}}\left(R_1(n',\phi,\mathcal{H}_{\mathrm{u}}) + 2C_{\mathrm{loss}}\sqrt{\frac{1}{n}\log\frac{1}{\delta}}\right)^{\frac{\eta}{\alpha}}\left(R_2(n',\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}}) + 2C_{\mathrm{loss}}\sqrt{\frac{2}{n}\log\frac{1}{\delta}}\right)^{\frac{1}{\beta}}, && (36)
\end{aligned}$$

By summarizing the result in an asymptotic form, we obtain

$$\begin{aligned}
&D_\phi(\mathbf{E}[Y\mid X],\widehat{h}_{\mathrm{x}}(X)) && (37)\\
&\le \mathcal{O}_P\Bigg(\left(R_1(n',\phi,\mathcal{H}_{\mathrm{u}}) + \sqrt{\frac{1}{n'}}\right)^{\eta} + \left(R_2(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}}) + \sqrt{\frac{1}{n}}\right) && (38)\\
&\quad + \left(R_1(n',\phi,\mathcal{H}_{\mathrm{u}}) + \sqrt{\frac{1}{n'}}\right)^{\alpha^{-1}\eta}\times\left(R_2(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}}) + \sqrt{\frac{1}{n}}\right)^{\beta^{-1}}\Bigg), && (39)
\end{aligned}$$

where

$$R_1(n',\phi,\mathcal{H}_{\mathrm{u}}) \coloneqq 2\mathfrak{R}(\phi\circ\mathcal{H}_u(S')) + 8(L_\phi + C_{\mathrm{out}})\sum_{j=1}^{k}\left(\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_u(S')) + \mathfrak{R}(\mathcal{H}_{\mathrm{u}}(S'))\right), \tag{40}$$

$$R_2(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}}) \coloneqq \mathfrak{R}(\phi\circ\mathcal{H}_{\mathrm{x}}(S)) + \mathfrak{R}(\phi\circ\mathcal{H}_{\mathrm{u}}(S)) \tag{41}$$

$$+ 8(L_\phi + C_{\mathrm{out}})\sum_{j=1}^{k}\left(\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_{\mathrm{x}}(S)) + \mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_{\mathrm{u}}(S)) + \mathfrak{R}(\mathcal{H}_{\mathrm{x}}^{(j)}(S)) + \mathfrak{R}(\mathcal{H}_{\mathrm{u}}^{(j)}(S))\right). \tag{42}$$

□

C Excess Risk Bound for Joint-BregMU

Let

$$(\widehat{h}_{\mathrm{x}}^{\text{JBreg}}, \widehat{h}_{\mathrm{u}}^{\text{JBreg}}) \coloneqq \operatorname*{arg\,min}_{(h_{\mathrm{x}},h_{\mathrm{u}})\in\mathcal{H}_{\mathrm{x}}\times\mathcal{H}_{\mathrm{u}}} \widehat{B}_{\phi}^{\times}(h_{\mathrm{x}}, h_{\mathrm{u}}, +1), \tag{43}$$

$$(h_{\mathrm{x}}^{\text{JBreg}}, h_{\mathrm{u}}^{\text{JBreg}}) \coloneqq \operatorname*{arg\,min}_{(h_{\mathrm{x}},h_{\mathrm{u}})\in\mathcal{H}_{\mathrm{x}}\times\mathcal{H}_{\mathrm{u}}} B_{\phi}^{\times}(h_{\mathrm{x}}, h_{\mathrm{u}}, +1), \tag{44}$$

where

$$\begin{aligned}\widehat{B}_{\phi}^{\times}(h_{\mathrm{x}}, h_{\mathrm{u}}, +1) \coloneqq{}& \frac{1}{n'}\sum_{i=1}^{n'} \ell_{\phi}(Y_i', h_{\mathrm{u}}(U_i')) + \frac{1}{n}\sum_{i=1}^{n} \ell_{\phi}(h_{\mathrm{u}}(U_i), h_{\mathrm{x}}(X_i)) \\ &+ \sqrt{\frac{1}{n'}\sum_{i=1}^{n'} \|Y_i' - h_{\mathrm{u}}(U_i')\|_2^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\nabla\phi(h_{\mathrm{u}}(U_i)) - \nabla\phi(h_{\mathrm{x}}(X_i))\|_2^2}, \end{aligned} \tag{45}$$

$$\begin{aligned}B_{\phi}^{\times}(h_{\mathrm{x}}, h_{\mathrm{u}}, +1) \coloneqq{}& \mathbf{E}[\ell_{\phi}(Y, h_{\mathrm{u}}(U))] + \mathbf{E}[\ell_{\phi}(h_{\mathrm{u}}(U), h_{\mathrm{x}}(X))] \\ &+ \sqrt{\mathbf{E}[\|Y - h_{\mathrm{u}}(U)\|_2^2]} \times \sqrt{\mathbf{E}[\|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|_2^2]}. \end{aligned} \tag{46}$$

Theorem C.1 (Excess risk bound for Joint-BregMU)**.** *Assume that the conditions (i)–(iv) of Theorem 5.1,* $\inf_{h_{\mathrm{u}}\in\mathcal{H}_{\mathrm{u}}} \mathbf{E}[\|Y - h_{\mathrm{u}}(U)\|_2^2] > 0$*, and* $\inf_{h_{\mathrm{x}}\in\mathcal{H}_{\mathrm{x}}, h_{\mathrm{u}}\in\mathcal{H}_{\mathrm{u}}} \mathbf{E}[\|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|_2^2] > 0$ *hold. Let*

$$(\widehat{h}_{\mathrm{x}}^{\text{JBreg}}, \widehat{h}_{\mathrm{u}}^{\text{JBreg}}) \coloneqq \operatorname*{arg\,min}_{(h_{\mathrm{x}},,h_{\mathrm{u}})\in\mathcal{H}_{\mathrm{x}}\times\mathcal{H}_{\mathrm{u}}} \widehat{B}_{\phi}^{\times}(h_{\mathrm{x}}, h_{\mathrm{u}}, +1), \quad (h_{\mathrm{x}}^{\text{JBreg}}, h_{\mathrm{u}}^{\text{JBreg}}) \coloneqq \operatorname*{arg\,min}_{(h_{\mathrm{x}},,h_{\mathrm{u}})\in\mathcal{H}_{\mathrm{x}}\times\mathcal{H}_{\mathrm{u}}} B_{\phi}^{\times}(h_{\mathrm{x}}, h_{\mathrm{u}}, +1).$$

Then, for any $\delta > 0$*, with probability at least* $1 - \delta$*, it holds that*

$$B_{\phi}^{\times}(\widehat{h}_{\mathrm{x}}^{\text{JBreg}}, \widehat{h}_{\mathrm{u}}^{\text{JBreg}}, +1) - B_{\phi}^{\times}(h_{\mathrm{x}}^{\text{JBreg}}, h_{\mathrm{u}}^{\text{JBreg}}, +1) \tag{47}$$

$$\leq \widetilde{R}_5(n', \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) + \widetilde{R}_6(n, \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) + C_{\text{loss}}\sqrt{2\log\frac{1}{\delta}}\left(\sqrt{\frac{1}{n'}} + \sqrt{\frac{1}{n}}\right) \tag{48}$$

$$+ \begin{cases} \sqrt{\frac{2L_\phi C_{\text{out}}}{c_1}}\left(\widetilde{R}_7(n', \mathcal{H}_{\mathrm{u}}) + 2C_{\text{out}}\sqrt{\frac{1}{2n'}\log\frac{2}{\delta}}\right) + \sqrt{\frac{2C_{\text{out}}}{c_2}}\left(\widetilde{R}_8(n, \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) + 2L_\phi\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}\right) \\ \quad \textit{if } e_{1,n'} \leq 3c_1/4 \textit{ and } e_{2,n} \leq 3c_2/4, \\ \sqrt{2L_\phi C_{\text{out}}}\sqrt{\widetilde{R}_7(n', \mathcal{H}_{\mathrm{u}}) + 2C_{\text{out}}\sqrt{\frac{1}{2n'}\log\frac{2}{\delta}}} + \sqrt{2C_{\text{out}}}\sqrt{\widetilde{R}_8(n, \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) + 2L_\phi\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}} \\ \quad \textit{otherwise}, \end{cases} \tag{49}$$

where

$$\widetilde{R}_5(n', \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) \coloneqq \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{u}}(S')) + 4(L_\phi + C_{\text{out}})\sum_{j=1}^{k}(\mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{u}}(S')) + \mathfrak{R}(\mathcal{H}_{\mathrm{x}}^{(j)}(S'))), \tag{50}$$

$$\widetilde{R}_6(n, \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) \coloneqq \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{x}}(S)) + \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{u}}(S)) \tag{51}$$

$$+ 4(L_\phi + C_{\text{out}})\sum_{j=1}^{k}(\mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{x}}(S)) + \mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{u}}(S))) \tag{52}$$

$$+ 4(L_\phi + C_{\text{out}})\sum_{j=1}^{k}(\mathfrak{R}(\mathcal{H}_{\mathrm{x}}^{(j)}(S)) + \mathfrak{R}(\mathcal{H}_{\mathrm{u}}^{(j)}(S))), \tag{53}$$

$$\widetilde{R}_7(n', \phi, \mathcal{H}_{\mathrm{u}}) \coloneqq 2C_{\text{out}}\mathfrak{R}(\mathcal{H}_{\mathrm{u}}^{(j)}(S')), \tag{54}$$

$$\widetilde{R}_8(n, \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) \coloneqq 2L_\phi\sum_{j=1}^{k}(\mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{x}}(S)) + \mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_u(S'))). \tag{55}$$

In an asymptotic form, we have

$$B_\phi^\times(\widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}, \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}, +1) - B_\phi^\times(h_{\mathrm{x}}^{\mathrm{JBreg}}, h_{\mathrm{u}}^{\mathrm{JBreg}}, +1) \tag{56}$$

$$\leq \mathcal{O}_P\left(R_5(n', \mathcal{H}_{\mathrm{u}}) + R_6(n', \phi, \mathcal{H}_{\mathrm{u}}) + R_7(n, \phi, \mathcal{H}_{\mathrm{u}}, \mathcal{H}_{\mathrm{x}}) + \sqrt{\frac{1}{n'}} + \sqrt{\frac{1}{n}}\right), \tag{57}$$

where

$$R_5(n', \mathcal{H}_{\mathrm{u}}) \coloneqq \mathfrak{R}_{n'}(\mathcal{H}_{\mathrm{u}}^{(j)}), \tag{58}$$

$$R_6(n', \phi, \mathcal{H}_{\mathrm{u}}) \coloneqq \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{u}}) + \sum_{j=1}^{k}(\mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{u}}) + \mathfrak{R}(\mathcal{H}_{\mathrm{x}}^{(j)})), \tag{59}$$

$$R_7(n, \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) \coloneqq \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{x}}) + \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{u}}) \tag{60}$$

$$+ \sum_{j=1}^{k}(\mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{x}}) + \mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{u}})) \tag{61}$$

$$+ \sum_{j=1}^{k}(\mathfrak{R}(\mathcal{H}_{\mathrm{x}}^{(j)}) + \mathfrak{R}(\mathcal{H}_{\mathrm{u}}^{(j)})) \tag{62}$$

$$+ \sum_{j=1}^{k}(\mathfrak{R}_n(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{x}}) + \mathfrak{R}_{n'}(\nabla^{(j)}\phi \circ \mathcal{H}_u)). \tag{63}$$

Proof of Theorem C.1. Decompose the excess risk as

$$B_\phi^\times(\widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}, \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}, +1) - B_\phi^\times(h_{\mathrm{x}}^{\mathrm{JBreg}}, h_{\mathrm{u}}^{\mathrm{JBreg}}, +1) \tag{64}$$

$$= \mathbf{E}[\ell_\phi(Y, \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U))] + \mathbf{E}[\ell_\phi(\widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U), \widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}(X))] \tag{65}$$

$$+ \sqrt{\mathbf{E}[\|Y - \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U)\|_2^2]} \times \sqrt{\mathbf{E}[\|\nabla\phi(\widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U)) - \nabla\phi(\widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}(X))\|_2^2]} \tag{66}$$

$$- \mathbf{E}[\ell_\phi(Y, h_{\mathrm{u}}^{\mathrm{JBreg}}(U))] - \mathbf{E}[\ell_\phi(h_{\mathrm{u}}^{\mathrm{JBreg}}(U), h_{\mathrm{x}}^{\mathrm{JBreg}}(X))] \tag{67}$$

$$- \sqrt{\mathbf{E}[\|Y - h_{\mathrm{u}}^{\mathrm{JBreg}}(U)\|_2^2]} \times \sqrt{\mathbf{E}[\|\nabla\phi(h_{\mathrm{u}}^{\mathrm{JBreg}}(U)) - \nabla\phi(h_{\mathrm{x}}^{\mathrm{JBreg}}(X))\|_2^2]} \tag{68}$$

$$= A_1 + A_2 + A_3, \tag{69}$$

where

$$A_1 \coloneqq \mathbf{E}[\ell_\phi(Y, \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U))] + \mathbf{E}[\ell_\phi(\widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U), \widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}(X))] \tag{70}$$

$$+ \sqrt{\mathbf{E}[\|Y - \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U)\|_2^2]} \times \sqrt{\mathbf{E}[\|\nabla\phi(\widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U)) - \nabla\phi(\widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}(X))\|_2^2]} \tag{71}$$

$$- \frac{1}{n'}\sum_{i=1}^{n'}\ell_\phi(Y_i', \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U_i')) + \frac{1}{n}\sum_{i=1}^{n}\ell_\phi(\widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U_i), \widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}(X_i)) \tag{72}$$

$$- \sqrt{\frac{1}{n'}\sum_{i=1}^{n'}\|Y_i' - \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U_i')\|_2^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\nabla\phi(\widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U_i)) - \nabla\phi(\widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}(X_i'))\|_2^2}, \tag{73}$$

$$A_2 \coloneqq \frac{1}{n'}\sum_{i=1}^{n'} \ell_\phi(Y'_i, \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U'_i)) + \frac{1}{n}\sum_{i=1}^{n} \ell_\phi(\widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U_i), \widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}(X_i)) \tag{74}$$

$$+ \sqrt{\frac{1}{n'}\sum_{i=1}^{n'} \|Y'_i - \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U'_i)\|_2^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\nabla\phi(\widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}(U_i)) - \nabla\phi(\widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}(X'_i))\|_2^2} \tag{75}$$

$$- \frac{1}{n'}\sum_{i=1}^{n'} \ell_\phi(Y'_i, h_{\mathrm{u}}^{\times}(U'_i)) + \frac{1}{n}\sum_{i=1}^{n} \ell_\phi(h_{\mathrm{u}}^{\mathrm{JBreg}}(U_i), h_{\mathrm{x}}^{\mathrm{JBreg}}(X_i)) \tag{76}$$

$$- \sqrt{\frac{1}{n'}\sum_{i=1}^{n'} \|Y'_i - h_{\mathrm{u}}^{\mathrm{JBreg}}(U'_i)\|_2^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\nabla\phi(h_{\mathrm{u}}^{\mathrm{JBreg}}(U_i)) - \nabla\phi(h_{\mathrm{x}}^{\mathrm{JBreg}}(X'_i))\|_2^2}, \tag{77}$$

and

$$A_3 \coloneqq \frac{1}{n'}\sum_{i=1}^{n'} \ell_\phi(Y'_i, h_{\mathrm{u}}^{\times}(U'_i)) + \frac{1}{n}\sum_{i=1}^{n} \ell_\phi(h_{\mathrm{u}}^{\times}(U_i), h_{\mathrm{x}}^{\times}(X_i)) \tag{78}$$

$$+ \sqrt{\frac{1}{n'}\sum_{i=1}^{n'} \|Y'_i - h_{\mathrm{u}}^{\times}(U'_i)\|_2^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\nabla\phi(h_{\mathrm{u}}^{\times}(U_i)) - \nabla\phi(h_{\mathrm{x}}^{\times}(X'_i))\|_2^2} \tag{79}$$

$$- \mathbf{E}[\ell_\phi(Y, h_{\mathrm{u}}^{\mathrm{JBreg}}(U))] - \mathbf{E}[\ell_\phi(h_{\mathrm{u}}^{\mathrm{JBreg}}(U), h_{\mathrm{x}}^{\mathrm{JBreg}}(X))] \tag{80}$$

$$- \sqrt{\mathbf{E}[\|Y - h_{\mathrm{u}}^{\mathrm{JBreg}}(U)\|_2^2]} \times \sqrt{\mathbf{E}[\|\nabla\phi(h_{\mathrm{u}}^{\mathrm{JBreg}}(U)) - \nabla\phi(h_{\mathrm{x}}^{\mathrm{JBreg}}(X))\|_2^2]}. \tag{81}$$

$A_2 \leq 0$ from the optimality of $(\widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}, \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}})$. To bound A_1 and A_3, it suffices to bound

$$A_{4,1} \coloneqq \sup_{h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \left| \mathbf{E}[\ell_\phi(Y, h_{\mathrm{u}}(U))] - \frac{1}{n'}\sum_{i=1}^{n'} \ell_\phi(Y'_i, h_{\mathrm{u}}(U'_i)) \right|, \tag{82}$$

$$A_{4,2} \coloneqq \sup_{h_{\mathrm{x}} \in \mathcal{H}_{\mathrm{x}}, h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \left| \mathbf{E}[\ell_\phi(h_{\mathrm{u}}(U), h_{\mathrm{x}}(X))] - \frac{1}{n}\sum_{i=1}^{n} \ell_\phi(h_{\mathrm{u}}(U_i), h_{\mathrm{x}}(X_i)) \right|, \tag{83}$$

$$A_{4,3} \coloneqq \sup_{h_{\mathrm{x}} \in \mathcal{H}_{\mathrm{x}}, h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \left| \sqrt{\mathbf{E}[\|Y - h_{\mathrm{u}}(U)\|_2^2]} \times \sqrt{\mathbf{E}[\|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|_2^2]} \right. \tag{84}$$

$$\left. - \sqrt{\frac{1}{n'}\sum_{i=1}^{n'} \|Y'_i - h_{\mathrm{u}}(U'_i)\|_2^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\nabla\phi(h_{\mathrm{u}}(U_i)) - \nabla\phi(h_{\mathrm{x}}(X'_i))\|_2^2} \right|. \tag{85}$$

From Lemma C.4,

$$A_{4,1} = \sup_{h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \left| \mathbf{E}[\ell_\phi(Y, h_{\mathrm{u}}(U))] - \frac{1}{n'}\sum_{i=1}^{n'} \ell_\phi(Y'_i, h_{\mathrm{u}}(U'_i)) \right| \tag{86}$$

$$\leq \widetilde{R}_5(n', \phi, \mathcal{H}_{\mathrm{u}}) + C_{\mathrm{loss}}\sqrt{\frac{1}{2n'}\log\frac{1}{\delta}}, \tag{87}$$

and

$$A_{4,2} = \sup_{h_{\mathrm{x}} \in \mathcal{H}_{\mathrm{x}}, h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \left| \mathbf{E}[\ell_\phi(h_{\mathrm{u}}(U), h_{\mathrm{x}}(X))] - \frac{1}{n}\sum_{i=1}^{n} \ell_\phi(h_{\mathrm{u}}(U_i), h_{\mathrm{x}}(X_i)) \right| \tag{88}$$

$$\leq \widetilde{R}_6(n, \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) + C_{\mathrm{loss}}\sqrt{\frac{1}{2n}\log\frac{1}{\delta}}, \tag{89}$$

where

$$\widetilde{R}_5(n', \phi, \mathcal{H}_{\mathrm{u}}) \coloneqq \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{u}}) + 4(L_\phi + C_{\mathrm{out}}) \sum_{j=1}^{k} (\mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{u}}) + \mathfrak{R}(\mathcal{H}_{\mathrm{x}}^{(j)})), \tag{90}$$

$$\widetilde{R}_6(n, \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) \coloneqq \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{x}}) + \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{u}}) \tag{91}$$

$$+ 4(L_\phi + C_{\mathrm{out}}) \sum_{j=1}^{k} (\mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{x}}) + \mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{u}})) \tag{92}$$

$$+ 4(L_\phi + C_{\mathrm{out}}) \sum_{j=1}^{k} (\mathfrak{R}(\mathcal{H}_{\mathrm{x}}^{(j)}) + \mathfrak{R}(\mathcal{H}_{\mathrm{u}}^{(j)})). \tag{93}$$

We will bound $A_{4,3}$ next. Denote $c_1 \coloneqq \inf_{h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \mathbf{E}[\|Y - h_{\mathrm{u}}(U)\|_2^2] > 0$ and $c_2 \coloneqq \inf_{h_{\mathrm{x}} \in \mathcal{H}_{\mathrm{x}}, h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \mathbf{E}[\|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|_2^2] > 0$. Let $B_1 \coloneqq \|Y - h_{\mathrm{u}}(U)\|_2^2$ and $B_2 \coloneqq \|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|_2^2$. Note that $B_1 \leq 2C_{\mathrm{out}}$ and $B_2 \leq 2L_\phi C_{\mathrm{out}}$. Let $\widehat{\mathbf{E}}[\cdot]$ denote the empirical average using the empirical measure defined by the training data. Then, we want to bound

$$\left| \sqrt{\mathbf{E}[B_1]\,\mathbf{E}[B_2]} - \sqrt{\widehat{\mathbf{E}}[B_1]\widehat{\mathbf{E}}[B_2]} \right|$$

uniformly over the choice of $(h_{\mathrm{x}}, h_{\mathrm{u}}) \in \mathcal{H}_{\mathrm{x}} \times \mathcal{H}_{\mathrm{u}}$. We are going to use the bounds obtained from Lemmas C.6 and C.7,

$$\left| \mathbf{E}[B_1] - \widehat{\mathbf{E}}[B_1] \right| \leq e_{1,n'} \text{ and } \left| \mathbf{E}[B_2] - \widehat{\mathbf{E}}[B_2] \right| \leq e_{2,n} \tag{94}$$

that hold uniformly over the choice of $(h_{\mathrm{x}}, h_{\mathrm{u}}) \in \mathcal{H}_{\mathrm{x}} \times \mathcal{H}_{\mathrm{u}}$ with probability at least $1 - \delta$, where

$$e_{1,n'} \coloneqq \widetilde{R}_7(n', \mathcal{H}_{\mathrm{u}}) + 2C_{\mathrm{out}} \sqrt{\frac{1}{2n'} \log \frac{2}{\delta}}, \tag{95}$$

$$e_{2,n} \coloneqq \widetilde{R}_8(n, \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) + 2L_\phi \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}, \tag{96}$$

$$\widetilde{R}_7(n', \phi, \mathcal{H}_{\mathrm{u}}) \coloneqq 2C_{\mathrm{out}} \mathfrak{R}(\mathcal{H}_{\mathrm{u}}^{(j)}(S')), \tag{97}$$

$$\widetilde{R}_8(n, \phi, \mathcal{H}_{\mathrm{x}}, \mathcal{H}_{\mathrm{u}}) \coloneqq 2L_\phi \sum_{j=1}^{k} (\mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{x}}(S)) + \mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_u(S))). \tag{98}$$

The case in which $e_{1,n'} \leq 3c_1/4$ and $e_{2,n} \leq 3c_2/4$: In this case, denoting $\delta_1 \coloneqq \widehat{\mathbf{E}}[B_1] - \mathbf{E}[B_1]$ and $\delta_2 \coloneqq \widehat{\mathbf{E}}[B_2] - \mathbf{E}[B_2]$, we have

$$\sqrt{\widehat{\mathbf{E}}[B_1]} - \sqrt{\mathbf{E}[B_1]} = \sqrt{\mathbf{E}[B_1] + \delta_1} - \sqrt{\mathbf{E}[B_1]} \leq \frac{\delta_1}{2\sqrt{\mathbf{E}[B_1]}} \leq \frac{e_{1,n'}}{2\sqrt{c_1}}, \tag{99}$$

where we used the inequality

$$\sqrt{a + \Delta a} - \sqrt{a} \leq \frac{\Delta a}{2\sqrt{a}} \tag{100}$$

that holds for any $a > 0$ and $\Delta a \in \mathbb{R}$ such that $a + \Delta a \geq 0$. The second inequality follows because of the definition of c_1. Similarly,

$$\sqrt{\widehat{\mathbf{E}}[B_2]} - \sqrt{\mathbf{E}[B_2]} \leq \frac{e_{2,n}}{2\sqrt{c_2}}. \tag{101}$$

On the other hand,

$$\sqrt{\mathbf{E}[B_1]} - \sqrt{\widehat{\mathbf{E}}[B_1]} = \sqrt{\widehat{\mathbf{E}}[B_1] - \delta_1} - \sqrt{\widehat{\mathbf{E}}[B_1]} \tag{102}$$
$$\leq \frac{-\delta_1}{2\sqrt{\widehat{\mathbf{E}}[B_1]}} \tag{103}$$
$$\leq \frac{e_{1,n'}}{2\sqrt{\widehat{\mathbf{E}}[B_1]}} \tag{104}$$
$$\leq \frac{e_{1,n'}}{\sqrt{c_1}}. \tag{105}$$

since

$$\widehat{\mathbf{E}}[B_1] = \mathbf{E}[B_1] + \widehat{\mathbf{E}}[B_1] - \mathbf{E}[B_1] \tag{106}$$
$$= |\mathbf{E}[B_1]| - \left|\widehat{\mathbf{E}}[B_1] - \mathbf{E}[B_1]\right| \tag{107}$$
$$\geq c_1 - e_{1,n'} \tag{108}$$
$$\geq c_1 - \frac{3}{4}c_1 = \frac{c_1}{4}. \tag{109}$$

Similarly,

$$\sqrt{\widehat{\mathbf{E}}[B_2]} - \sqrt{\mathbf{E}[B_2]} \leq \frac{e_{2,n}}{\sqrt{c_2}}. \tag{110}$$

Thus,

$$\left|\sqrt{\mathbf{E}[B_1]} - \sqrt{\widehat{\mathbf{E}}[B_1]}\right| \leq \frac{e_{1,n'}}{\sqrt{c_1}}, \tag{111}$$
$$\left|\sqrt{\mathbf{E}[B_2]} - \sqrt{\widehat{\mathbf{E}}[B_2]}\right| \leq \frac{e_{2,n}}{\sqrt{c_2}}. \tag{112}$$

Using these inequalities, we obtain

$$\left|\sqrt{\mathbf{E}[B_1]\,\mathbf{E}[B_2]} - \sqrt{\widehat{\mathbf{E}}[B_1]\widehat{\mathbf{E}}[B_2]}\right| \tag{113}$$
$$\leq \left|\left(\sqrt{\mathbf{E}[B_1]} - \sqrt{\widehat{\mathbf{E}}[B_1]} + \sqrt{\widehat{\mathbf{E}}[B_1]}\right)\sqrt{\mathbf{E}[B_2]} - \sqrt{\widehat{\mathbf{E}}[B_1]\widehat{\mathbf{E}}[B_2]}\right| \tag{114}$$
$$\leq \left|\sqrt{\mathbf{E}[B_1]} - \sqrt{\widehat{\mathbf{E}}[B_1]}\right| \times \sqrt{\mathbf{E}[B_2]} + \sqrt{\widehat{\mathbf{E}}[B_1]} \times \left|\sqrt{\mathbf{E}[B_2]} - \sqrt{\widehat{\mathbf{E}}[B_2]}\right| \tag{115}$$
$$\leq \sqrt{\mathbf{E}[B_2]}\frac{e_{1,n'}}{\sqrt{c_1}} + \sqrt{\widehat{\mathbf{E}}[B_1]}\frac{e_{2,n}}{\sqrt{c_2}} \tag{116}$$
$$\leq \sqrt{\frac{2L_\phi C_{\text{out}}}{c_1}}e_{1,n'} + \sqrt{\frac{2C_{\text{out}}}{c_2}}e_{2,n}. \tag{117}$$

Other cases:

$$\left|\sqrt{\mathbf{E}[B_1]\,\mathbf{E}[B_2]}-\sqrt{\widehat{\mathbf{E}}[B_1]\widehat{\mathbf{E}}[B_2]}\right| \tag{118}$$

$$=\sqrt{\left(\mathbf{E}[B_1]-\widehat{\mathbf{E}}[B_1]+\widehat{\mathbf{E}}[B_1]\right)\mathbf{E}[B_2]}-\sqrt{\widehat{\mathbf{E}}[B_1]\widehat{\mathbf{E}}[B_2]} \tag{119}$$

$$\leq\sqrt{\left(\mathbf{E}[B_1]-\widehat{\mathbf{E}}[B_1]\right)\mathbf{E}[B_2]}+\sqrt{\widehat{\mathbf{E}}[B_1]\,\mathbf{E}[B_2]}-\sqrt{\widehat{\mathbf{E}}[B_1]\widehat{\mathbf{E}}[B_2]} \tag{120}$$

$$\leq\sqrt{\left(\mathbf{E}[B_1]-\widehat{\mathbf{E}}[B_1]\right)\mathbf{E}[B_2]}+\sqrt{\widehat{\mathbf{E}}[B_1]}\left(\sqrt{\mathbf{E}[B_2]}-\sqrt{\widehat{\mathbf{E}}[B_2]}\right) \tag{121}$$

$$\leq\sqrt{\left(\mathbf{E}[B_1]-\widehat{\mathbf{E}}[B_1]\right)\mathbf{E}[B_2]}+\sqrt{\widehat{\mathbf{E}}[B_1]}\left(\sqrt{\mathbf{E}[B_2]}-\sqrt{\widehat{\mathbf{E}}[B_2]}\right) \tag{122}$$

$$\leq\sqrt{\left(\mathbf{E}[B_1]-\widehat{\mathbf{E}}[B_1]\right)\mathbf{E}[B_2]}+\sqrt{\widehat{\mathbf{E}}[B_1]}\left(\sqrt{\mathbf{E}[B_2]-\widehat{\mathbf{E}}[B_2]+\widehat{\mathbf{E}}[B_2]}-\sqrt{\widehat{\mathbf{E}}[B_2]}\right) \tag{123}$$

$$\leq\sqrt{\left(\mathbf{E}[B_1]-\widehat{\mathbf{E}}[B_1]\right)\mathbf{E}[B_2]}+\sqrt{\widehat{\mathbf{E}}[B_1]}\sqrt{\mathbf{E}[B_2]-\widehat{\mathbf{E}}[B_2]} \tag{124}$$

$$\leq\sqrt{\mathbf{E}[B_2]}\sqrt{e_{1,n'}}+\sqrt{\widehat{\mathbf{E}}[B_1]}\sqrt{e_{2,n}} \tag{125}$$

$$\leq\sqrt{2L_\phi C_{\text{out}}e_{1,n'}}+\sqrt{2C_{\text{out}}e_{2,n}}. \tag{126}$$

These bounds hold uniformly for all $h_{\mathrm{x}}\in\mathcal{H}_{\mathrm{x}}$ and $h_{\mathrm{u}}\in\mathcal{H}_{\mathrm{u}}$. Hence, summarizing the results above gives

$$A_{4,3}\leq\begin{cases}\sqrt{\frac{2L_\phi C_{\text{out}}}{c_1}}e_{1,n'}+\sqrt{\frac{2C_{\text{out}}}{c_1}}e_{2,n} & \text{if } e_{1,n'}\leq 3c_1/4 \text{ and } e_{2,n}\leq 3c_2/4,\\ \sqrt{2L_\phi C_{\text{out}}e_{1,n'}}+\sqrt{2C_{\text{out}}e_{2,n}} & \text{otherwise.}\end{cases} \tag{127}$$

Note that when $\min\{n,n'\}$ is sufficiently large, the first case holds with high probability. Hence, in an asymptotic form, we obtain

$$\left|\sqrt{\mathbf{E}[B_1]\,\mathbf{E}[B_2]}-\sqrt{\widehat{\mathbf{E}}[B_1]\widehat{\mathbf{E}}[B_2]}\right|\leq\mathcal{O}\left(e_{1,n'}+e_{1,n'}\right). \tag{128}$$

Collecting the results we have obtained, we conclude that with probability at least $1-\delta$, it holds that

$$B_\phi^{\times}(\widehat{h}_{\mathrm{x}}^{\text{JBreg}},\widehat{h}_{\mathrm{u}}^{\text{JBreg}},+1)-B_\phi^{\times}(h_{\mathrm{x}}^{\text{JBreg}},h_{\mathrm{u}}^{\text{JBreg}},+1) \tag{129}$$

$$\leq A_1+A_2+A_3\leq A_{4,1}+A_{4,2}+A_{4,3} \tag{130}$$

$$\leq\widetilde{R}_5(n',\phi,\mathcal{H}_{\mathrm{u}})+\widetilde{R}_6(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}})+C_{\text{loss}}\sqrt{\frac{1}{2}\log\frac{1}{\delta}}\left(\sqrt{\frac{1}{n'}}+\sqrt{\frac{1}{n}}\right) \tag{131}$$

$$+\begin{cases}\sqrt{\frac{2L_\phi C_{\text{out}}}{c_1}}\left(\widetilde{R}_7(n',\mathcal{H}_{\mathrm{u}})+2C_{\text{out}}\sqrt{\frac{1}{2n'}\log\frac{2}{\delta}}\right)+\sqrt{\frac{2C_{\text{out}}}{c_2}}\left(\widetilde{R}_8(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}})+2L_\phi\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}\right)\\ \quad\text{if } e_{1,n'}\leq 3c_1/4 \text{ and } e_{2,n}\leq 3c_2/4,\\ \sqrt{2L_\phi C_{\text{out}}}\sqrt{\widetilde{R}_7(n',\mathcal{H}_{\mathrm{u}})+2C_{\text{out}}\sqrt{\frac{1}{2n'}\log\frac{2}{\delta}}}+\sqrt{2C_{\text{out}}}\sqrt{\widetilde{R}_8(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}})+2L_\phi\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}}\\ \quad\text{otherwise.}\end{cases} \tag{132}$$

Summarizing this in an asymptotic form, we get

$$B_\phi^{\times}(\widehat{h}_{\mathrm{x}}^{\text{JBreg}},\widehat{h}_{\mathrm{u}}^{\text{JBreg}},+1)-B_\phi^{\times}(h_{\mathrm{x}}^{\text{JBreg}},h_{\mathrm{u}}^{\text{JBreg}},+1) \tag{133}$$

$$\leq\mathcal{O}_P\left(R_5(n',\mathcal{H}_{\mathrm{u}})+R_6(n',\phi,\mathcal{H}_{\mathrm{u}})+R_7(n,\phi,\mathcal{H}_{\mathrm{u}},\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}})+\sqrt{\frac{1}{n'}}+\sqrt{\frac{1}{n}}\right), \tag{134}$$

where

$$R_5(n',\mathcal{H}_{\mathrm{u}}):=\widetilde{R}_7(n',\mathcal{H}_{\mathrm{u}}), \tag{135}$$

$$R_6(n',\phi,\mathcal{H}_{\mathrm{u}}):=\widetilde{R}_5(n',\phi,\mathcal{H}_{\mathrm{u}}), \tag{136}$$

$$R_7(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}}):=\widetilde{R}_6(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}})+\widetilde{R}_8(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}}). \tag{137}$$

□

Theorem 5.2 (Excess risk bound for Joint-BregMU, in an asymptotic form)**.** *Assume the conditions (i)–(iv) of Theorem 5.1 and that* $\inf_{h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \mathbf{E}[\|Y - h_{\mathrm{u}}(U)\|_2^2] > 0$ *and* $\inf_{h_{\mathrm{x}} \in \mathcal{H}_{\mathrm{x}}, h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \mathbf{E}[\|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|_2^2] > 0$*. Then,*

$$
\begin{aligned}
&D_\phi(\mathbf{E}[Y \mid X], \widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}(X)) \\
&\leq \underbrace{e_{n',n}}_{\textit{vanishing error}} + 2 \underbrace{\|Y - \mathbf{E}[Y \mid U]\|_{L^2}}_{\textit{1st bias factor}} \\
&\qquad \times \underbrace{\|\nabla\phi(\mathbf{E}[Y \mid U]) - \nabla\phi(\mathbf{E}[Y \mid X])\|_{L^2}}_{\textit{2nd bias factor}},
\end{aligned}
$$

where

$$
\begin{aligned}
e_{n,n'} \leq \mathcal{O}_P \Bigg(& R_3(n', \phi, \mathcal{H}_{\mathrm{u}}) + R_4(n, \phi, \mathcal{H}_{\mathrm{u}}, \mathcal{H}_{\mathrm{x}}) \\
& + \sqrt{\frac{1}{n'}} + \sqrt{\frac{1}{n}} \Bigg),
\end{aligned}
$$

and $R_l(\cdots)$ *(*$l = 3, 4$*) are model complexity terms that depend on the arguments.*

Proof of Theorem 5.2.

$$
\begin{aligned}
&\mathbf{E}[\ell_\phi(Y, \widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}(X))] && (138)\\
&\leq B_\phi^\times(\widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}, \widehat{h}_{\mathrm{u}}^{\mathrm{JBreg}}, +1) && (139)\\
&\leq B_\phi^\times(h_{\mathrm{x}}^{\mathrm{JBreg}}, h_{\mathrm{u}}^{\mathrm{JBreg}}, +1) + e_{n,n'} \quad \text{(from Theorem C.1)} && (140)\\
&\leq B_\phi^\times(\mathbf{E}[Y \mid X = (\cdot)], \mathbf{E}[Y \mid U = (\cdot)], +1) + e_{n,n'} && (141)\\
&\quad \text{(from the optimality of } (h_{\mathrm{x}}^{\mathrm{JBreg}}, h_{\mathrm{u}}^{\mathrm{JBreg}})\text{).} && (142)
\end{aligned}
$$

The first term of the last expression can be further bounded as

$$
\begin{aligned}
&B_\phi^\times(\mathbf{E}[Y \mid X = (\cdot)], \mathbf{E}[Y \mid U = (\cdot)], +1) && (143)\\
&= \mathbf{E}[\ell_\phi(Y, \mathbf{E}[Y \mid U])] + \mathbf{E}[\ell_\phi(\mathbf{E}[Y \mid U], \mathbf{E}[Y \mid X])] && (144)\\
&\quad + \|Y - \mathbf{E}[Y \mid U]\|_{L^2} \times \|\nabla\phi(\mathbf{E}[Y \mid U]) - \nabla\phi(\mathbf{E}[Y \mid X])\|_{L^2} && (145)\\
&= \mathbf{E}[\ell_\phi(Y, \mathbf{E}[Y \mid U])] + \mathbf{E}[\ell_\phi(\mathbf{E}[Y \mid U], \mathbf{E}[Y \mid X])] && (146)\\
&\quad + \|Y - \mathbf{E}[Y \mid U]\|_{L^2} \times \|\nabla\phi(\mathbf{E}[Y \mid U]) - \nabla\phi(\mathbf{E}[Y \mid X])\|_{L^2} && (147)\\
&\quad + \mathbf{E}[(Y - \mathbf{E}[Y \mid U])^\top (\nabla\phi(\mathbf{E}[Y \mid U]) - \nabla\phi(\mathbf{E}[Y \mid X]))] && (148)\\
&\quad - \mathbf{E}[(Y - \mathbf{E}[Y \mid U])^\top (\nabla\phi(\mathbf{E}[Y \mid U]) - \nabla\phi(\mathbf{E}[Y \mid X]))] && (149)\\
&= \mathbf{E}[\ell_\phi(Y, \mathbf{E}[Y \mid X])] && (150)\\
&\quad + \|Y - \mathbf{E}[Y \mid U]\|_{L^2} \times \|\nabla\phi(\mathbf{E}[Y \mid U]) - \nabla\phi(\mathbf{E}[Y \mid X])\|_{L^2} && (151)\\
&\quad - \mathbf{E}[(Y - \mathbf{E}[Y \mid U])^\top (\nabla\phi(\mathbf{E}[Y \mid U]) - \nabla\phi(\mathbf{E}[Y \mid X]))] && (152)\\
&\leq \mathbf{E}[\ell_\phi(Y, \mathbf{E}[Y \mid X])] + 2\|Y - \mathbf{E}[Y \mid U]\|_{L^2} \times \|\nabla\phi(\mathbf{E}[Y \mid U]) - \nabla\phi(\mathbf{E}[Y \mid X])\|_{L^2}. && (153)
\end{aligned}
$$

From Lemma A.1, we get

$$
\mathbf{E}[\ell_\phi(\mathbf{E}[Y \mid X], \widehat{h}_{\mathrm{x}}^{\mathrm{JBreg}}(X))] \leq e_{n',n} + 2\|Y - \mathbf{E}[Y \mid U]\|_{L^2} \times \|\nabla\phi(\mathbf{E}[Y \mid U]) - \nabla\phi(\mathbf{E}[Y \mid X])\|_{L^2}. \tag{154}
$$

□

C.1 Lemmas for Excess Risk Bounds

Definition C.1 (Rademacher Complexity)**.** Let $A = \{\{a_i\}_{i=1}^N \subseteq \mathbb{R}\}$ be a set of real sequences. Define the Rademacher complexity of A as

$$
\mathfrak{R}(A) \coloneqq \mathop{\mathbf{E}}_{\varepsilon_1, \ldots, \varepsilon_N} \left[\sup_{\{a_i\}_{i=1}^N \in A} \frac{1}{N} \sum_{i=1}^n \varepsilon_i a_i \right], \tag{155}
$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are Rademacher variables, namely, independent, {-1, 1}-valued, uniform random variables. Furthermore, for any function class $\mathcal{H}$, define the Rademacher complexity of $\mathcal{H}$ over $S \coloneqq \{x_i\}_{i=1}^N \subseteq \mathbb{R}^d$ as

$$\mathfrak{R}(\mathcal{H}(S)) = \mathop{\mathbf{E}}_{\varepsilon_1,\ldots,\varepsilon_N}\left[\sup_{\{y_i\}_{i=1}^N \in \mathcal{H}(S)} \frac{1}{N}\sum_{i=1}^n \varepsilon_i y_i\right] \tag{156}$$

$$= \mathop{\mathbf{E}}_{\varepsilon_1,\ldots,\varepsilon_N}\left[\sup_{h\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^n \varepsilon_i h(x_i)\right], \tag{157}$$

where we used the notation

$$\mathcal{H}(S) \coloneqq \{\{h(x_i)\}_{i=1}^N \mid h \in \mathcal{H}\}. \tag{158}$$

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

Lemma C.1 (McDiarmid's inequality)**.** *Let $\varphi : \mathcal{D}^N \to \mathbb{R}$ be a measurable function. Assume that there exists $B \in (0, \infty)$ such that*

$$|\varphi(v_1, \ldots, v_N) - \varphi(v'_1, \ldots, v'_N)| \leq B, \tag{159}$$

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

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

Lemma C.2. *Let $\mathcal{Z}$ and $\mathcal{W}$ be measurable spaces. Let $\mathcal{H}_{\mathrm{z}} \subseteq \{h\colon \mathcal{Z} \to \mathcal{Y}\}$ and $\mathcal{H}_{\mathrm{w}} \subseteq \{h\colon \mathcal{W} \to \mathcal{Y}\}$ be function classes such that there exists a constant $C \in \mathbb{R}$ satisfying*

$$\ell_\phi(h_1(z), h_1(w)) \leq C \tag{160}$$

for all $h_1 \in \mathcal{H}_z$, $h_2 \in \mathcal{H}_w$, $z \in \mathcal{Z}$, and $w \in \mathcal{W}$. Define $\psi(\cdot)$ by

$$\psi\left(\{(z_i, w_i)\}_{i=1}^N; \mathcal{H}_{\mathrm{z}}, \mathcal{H}_{\mathrm{w}}\right) \coloneqq \sup_{h_1\in\mathcal{H}_z, h_2\in\mathcal{H}_{\mathrm{w}}} \left|\frac{1}{N}\sum_{i=1}^N \ell_\phi(h_1(z_i), h_2(w_i)) - \mathbf{E}[\ell_\phi(h_1(Z_i), h_2(W_i))]\right|, \tag{161}$$

where $\{(z_i, w_i)\}_{i=1}^N \subseteq \mathcal{Z} \times \mathcal{W}$, and Z_i and W_i are $\mathcal{Z}$-valued and $\mathcal{W}$-valued independent random variables. Then, with probability at least $1 - \delta$, it holds that

$$\psi\left(\{(Z_i, W_i)\}_{i=1}^N; \mathcal{H}_{\mathrm{z}}, \mathcal{H}_{\mathrm{w}}\right) \leq \mathbf{E}\left[\psi\left(\{(Z_i, W_i)\}_{i=1}^N; \mathcal{H}_{\mathrm{z}}, \mathcal{H}_{\mathrm{w}}\right)\right] + C\sqrt{\frac{1}{2N}\log\frac{1}{\delta}}. \tag{162}$$

Proof. Let $\{(z_i, w_i)\}_{i=1}^N \subseteq \mathcal{Z} \times \mathcal{W}$ and $\{(z'_i, w'_i)\}_{i=1}^N \subseteq \mathcal{Z} \times \mathcal{W}$ such that $(z_i, w_i) = (z'_i, w'_i)$ for all $i \in [N]$ but some $j \in [N]$.

$$\left|\psi\left(\{(z_i, w_i)\}_{i=1}^N; \mathcal{H}_{\mathrm{z}}, \mathcal{H}_{\mathrm{w}}\right) - \psi\left(\{(z'_i, w'_i)\}_{i=1}^N; \mathcal{H}_{\mathrm{z}}, \mathcal{H}_{\mathrm{w}}\right)\right| \tag{163}$$

$$\leq \frac{1}{N}\sup_{h_1\in\mathcal{H}_z, h_2\in\mathcal{H}_{\mathrm{w}}} \left|\ell_\phi(h_1(z_j), h_2(w_j)) - \ell_\phi(h_1(z'_j), h_2(w'_j))\right| \tag{164}$$

$$\leq \frac{1}{N}\sup_{h_1\in\mathcal{H}_z, h_2\in\mathcal{H}_{\mathrm{w}}} \max\{\ell_\phi(h_1(z_j), h_2(w_j)), \ell_\phi(h_1(z'_j), h_2(w'_j))\} \tag{165}$$

$$\leq \frac{C}{N}, \tag{166}$$

where we used $|a - b| \leq \max\{a - b, b - a\} \leq \max\{a, b\}$ for any $(a, b) \in [0, \infty)^2$. We obtain the result of the lemma by applying McDiarmid's inequality. □

Lemma C.3. *Let $S_{\mathrm{z,w}} \coloneqq \{(Z_i, W_i)\}_{i=1}^N$ with $\{(Z_i, W_i)\}_{i=1}^N$ defined as in Lemma C.2. Let $S_{\mathrm{z}} \coloneqq \{Z_i \mid (Z_i, W_i) \in S_{\mathrm{z,w}}, i \in [N]\}$ and $S_{\mathrm{w}} \coloneqq \{W_i \mid (Z_i, W_i) \in S_{\mathrm{z,w}}, i \in [N]\}$. Define $\mathcal{H}_{\mathrm{z}}$, $\mathcal{H}_{\mathrm{w}}$, and $\psi(\cdot)$ as in Lemma C.2. Then, we have*

$$\mathbf{E}[\psi\left(S_{\mathrm{z,w}}; \mathcal{H}_{\mathrm{z}}, \mathcal{H}_{\mathrm{w}}\right)] \leq \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{z}}(S_{\mathrm{z}})) + \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{w}}(S_{\mathrm{w}})) \tag{167}$$

$$+ 2(2L_\phi + C_{\text{out}}) \sum_{j=1}^{K} (\mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{z}}(S_{\mathrm{z}})) + \mathfrak{R}(\nabla^{(j)}\phi \circ \mathcal{H}_{\mathrm{w}}(S_{\mathrm{w}}))) \tag{168}$$

$$+ 2(L_\phi + 2C_{\text{out}}) \sum_{j=1}^{K} (\mathfrak{R}(\mathcal{H}_{\mathrm{z}}^{(j)}(S_{\mathrm{z}})) + \mathfrak{R}(\mathcal{H}_{\mathrm{w}}^{(j)}(S_{\mathrm{w}}))), \tag{169}$$

where $\mathcal{H}_z^{(j)} \coloneqq \{z \mapsto [h(z)]_j \mid h \in \mathcal{H}_z\}$ and $\mathcal{H}_w^{(j)} \coloneqq \{w \mapsto [h(w)]_j \mid h \in \mathcal{H}_w\}$.

Proof.

$$\mathbf{E}\left[\psi\left(\{(Z_i, W_i)\}_{i=1}^N; \mathcal{H}_{\mathrm{z}}, \mathcal{H}_{\mathrm{w}}\right)\right] \tag{170}$$

$$\leq \mathbf{E}\left[\frac{1}{N} \sup_{h_1 \in \mathcal{H}_{\mathrm{z}}, h_2 \in \mathcal{H}_{\mathrm{w}}} \left| \sum_{i=1}^N \ell_\phi(h_1(Z_i), h_2(W_i)) - \sum_{i=1}^N \ell_\phi(h_1(Z_i'), h_2(W_i')) \right| \right]$$

(from Jensen's inequality, where (Z_i', W_i') is an independent copy of (Z_i, W_i)) (171)

$$\leq \mathbf{E}\left[\frac{1}{N} \sup_{h_1 \in \mathcal{H}_{\mathrm{z}}, h_2 \in \mathcal{H}_{\mathrm{w}}} \sum_{i=1}^N \varepsilon_i \ell_\phi(h_1(Z_i), h_2(W_i)) \right] \tag{172}$$

$$\leq \mathfrak{R}(\{(z, w) \mapsto \ell_\phi(h_1(z), h_2(w)) \mid h_1 \in \mathcal{H}_{\mathrm{z}}, h_2 \in \mathcal{H}_{\mathrm{w}}\}(S_{\mathrm{z,w}})) \tag{173}$$

$$\leq \mathfrak{R}(\{(z, w) \mapsto \phi(h_1(z)) - \nabla\phi(h_2(w))^\top h_1(z) \tag{174}$$

$$- \phi(h_2(w)) + \nabla\phi(h_2(w))^\top h_2(w) \mid h_1 \in \mathcal{H}_{\mathrm{z}}, h_2 \in \mathcal{H}_{\mathrm{w}}\}(S_{\mathrm{z,w}})) \tag{175}$$

$$\leq \mathfrak{R}(\{z \mapsto \phi(h_1(z)) \mid h_1 \in \mathcal{H}_{\mathrm{z}}\}(S_{\mathrm{z}})) + \mathfrak{R}(\{(z, w) \mapsto \nabla\phi(h_2(w))^\top h_1(z) \mid h_1 \in \mathcal{H}_{\mathrm{z}}, h_2 \in \mathcal{H}_{\mathrm{w}}\}(S_{\mathrm{z,w}})) \tag{176}$$

$$+ \mathfrak{R}(\{w \mapsto \phi(h_2(w)) \mid h_2 \in \mathcal{H}_{\mathrm{w}}\}(S_{\mathrm{w}})) + \mathfrak{R}(\{(z, w) \mapsto \nabla\phi(h_1(z))^\top h_2(w) \mid h_1 \in \mathcal{H}_{\mathrm{z}}, h_2 \in \mathcal{H}_{\mathrm{w}}\}(S_{\mathrm{z,w}})). \tag{177}$$

Denote

$$\phi \circ \mathcal{H}_{\mathrm{z}} \coloneqq \{z \mapsto \phi(h_1(z)) \mid h_1 \in \mathcal{H}_{\mathrm{z}}\}, \tag{178}$$

$$\phi \circ \mathcal{H}_{\mathrm{w}} \coloneqq \{w \mapsto \phi(h_2(w)) \mid h_2 \in \mathcal{H}_{\mathrm{w}}\}. \tag{179}$$

Then,

$$\mathfrak{R}(\{z \mapsto \phi(h_1(z)) \mid h_1 \in \mathcal{H}_{\mathrm{z}}\}) = \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{z}}), \tag{180}$$

$$\mathfrak{R}(\{w \mapsto \phi(h_2(w)) \mid h_2 \in \mathcal{H}_{\mathrm{w}}\}) = \mathfrak{R}(\phi \circ \mathcal{H}_{\mathrm{w}}). \tag{181}$$

For a function class $\mathcal{H}$ bounded as $\sup_{h \in \mathcal{H}, z \in \mathcal{Z}} |h(z)| < B \in (0, \infty)$, we have $\mathfrak{R}(\{h(z)^2 \mid h \in \mathcal{H}\}) = 2B\mathfrak{R}(\mathcal{H})$ from the

Ledoux-Talagrand contraction lemma. Thus,

$$\mathfrak{R}(\{(z,w)\mapsto \nabla\phi(h_2(w))^\top h_1(z) \mid h_1\in\mathcal{H}_{\mathrm{z}}, h_2\in\mathcal{H}_{\mathrm{w}}\}(S_{\mathrm{z,w}})) \tag{182}$$
$$\leq \sum_{j=1}^{K}\mathfrak{R}(\{(z,w)\mapsto \nabla^{(j)}\phi(h_2(w))[h_1(z)]_j \mid h_1\in\mathcal{H}_{\mathrm{z}}, h_2\in\mathcal{H}_{\mathrm{w}}\}(S_{\mathrm{z,w}})) \tag{183}$$
$$\leq \sum_{j=1}^{K}\mathfrak{R}(\{(z,w)\mapsto (\nabla^{(j)}\phi(h_2(w))-[h_1(z)]_j)^2-(\nabla^{(j)}\phi(h_2(w)))^2-[h_1(z)]_j^2 \tag{184}$$
$$\mid h_1\in\mathcal{H}_{\mathrm{z}}, h_2\in\mathcal{H}_{\mathrm{w}}\}(S_{\mathrm{z,w}})) \tag{185}$$
$$\leq \sum_{j=1}^{K}\mathfrak{R}(\{(z,w)\mapsto (\nabla^{(j)}\phi(h_2(w))-[h_1(z)]_j)^2 \mid h_1\in\mathcal{H}_{\mathrm{z}}, h_2\in\mathcal{H}_{\mathrm{w}}\}(S_{\mathrm{z,w}})) \tag{186}$$
$$+\sum_{j=1}^{K}\mathfrak{R}(\{w\mapsto \nabla^{(j)}\phi(h_2(w))^2 \mid h_2\in\mathcal{H}_{\mathrm{w}}\}(S_{\mathrm{w}})) \tag{187}$$
$$+\sum_{j=1}^{K}\mathfrak{R}(\{z\mapsto [h_1(z)]_j^2 \mid h_1\in\mathcal{H}_{\mathrm{z}}\}(S_{\mathrm{z}})) \tag{188}$$
$$\leq 2\sum_{j=1}^{K}(L_\phi+C_{\text{out}})(\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_{\mathrm{w}}(S_{\mathrm{w}}))+\mathfrak{R}(\mathcal{H}_{\mathrm{z}}^{(j)}(S_{\mathrm{z}}))) \tag{189}$$
$$+2\sum_{j=1}^{K}L_\phi\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_{\mathrm{w}}(S_{\mathrm{w}}))+2\sum_{j=1}^{K}C_{\text{out}}\mathfrak{R}(\mathcal{H}_{\mathrm{z}}^{(j)}(S_{\mathrm{z}})) \tag{190}$$
$$\leq 2(2L_\phi+C_{\text{out}})\sum_{j=1}^{K}\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_{\mathrm{w}}(S_{\mathrm{w}}))+2(L_\phi+2C_{\text{out}})\sum_{j=1}^{K}\mathfrak{R}(\mathcal{H}_{\mathrm{z}}^{(j)}(S_{\mathrm{z}})). \tag{191}$$

Similarly,

$$\mathfrak{R}(\{(z,w)\mapsto \nabla\phi(h_1(z))^\top h_2(w) \mid h_1\in\mathcal{H}_{\mathrm{z}}, h_2\in\mathcal{H}_{\mathrm{w}}\}(S_{\mathrm{z,w}})) \tag{192}$$
$$\leq 2(2L_\phi+C_{\text{out}})\sum_{j=1}^{K}\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_{\mathrm{z}}(S_{\mathrm{z}}))+2(L_\phi+2C_{\text{out}})\sum_{j=1}^{K}\mathfrak{R}(\mathcal{H}_{\mathrm{w}}^{(j)}(S_{\mathrm{w}})). \tag{193}$$

We obtain the result of the lemma by combining the inequalities. □

Lemma C.4.

$$\psi\left(S_{\mathrm{z,w}};\mathcal{H}_{\mathrm{z}},\mathcal{H}_{\mathrm{w}}\right) \tag{194}$$
$$\leq \mathfrak{R}(\phi\circ\mathcal{H}_{\mathrm{z}}(S_{\mathrm{z}}))+\mathfrak{R}(\phi\circ\mathcal{H}_{\mathrm{w}}(S_{\mathrm{w}})) \tag{195}$$
$$+2(2L_\phi+C_{\text{out}})\sum_{j=1}^{K}(\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_{\mathrm{z}}(S_{\mathrm{z}}))+\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_{\mathrm{w}}(S_{\mathrm{w}}))) \tag{196}$$
$$+2(L_\phi+2C_{\text{out}})\sum_{j=1}^{K}(\mathfrak{R}(\mathcal{H}_{\mathrm{z}}^{(j)}(S_{\mathrm{z}}))+\mathfrak{R}(\mathcal{H}_{\mathrm{w}}^{(j)}(S_{\mathrm{w}})))+C_{\text{loss}}\sqrt{\frac{1}{2N}\log\frac{1}{\delta}}. \tag{197}$$

Proof. Combine Lemma C.2 and Lemma C.3. □

Lemma C.5.

$$D_\phi(\mathbf{E}[h_1(Z)\mid W],\widehat{h}_2(W))\leq 2\psi\left(\{(Z_i,W_i)\}_{i=1}^N;\mathcal{H}_{\mathrm{z}},\mathcal{H}_{\mathrm{w}}\right). \tag{198}$$

Proof. From Lemma A.1,

$$
\begin{aligned}
& D(\mathbf{E}[h_1(Z) \mid W], \widehat{h}_2(W)) && (199)\\
& = D_\phi(h_1(Z), \widehat{h}_2(W)) - D_\phi(h_1(Z), \mathbf{E}[h_2(Z) \mid Z]) && (200)\\
& \leq D_\phi(h_1(Z), \widehat{h}_2(W)) - \frac{1}{N}\sum_{i=1}^{N} \ell_\phi(h_1(Z_i), \widehat{h}_2(W_i)) && (201)\\
& \quad + \underbrace{\frac{1}{N}\sum_{i=1}^{N} \ell_\phi(h_1(Z_i), \widehat{h}_2(W_i)) - \frac{1}{N}\sum_{i=1}^{N} \ell_\phi(h_1(Z_i), \mathbf{E}[h_1(Z) \mid W = W_i])}_{= 0 \text{ (from the optimality of } \widehat{h}_2)} && (202)\\
& \quad + \frac{1}{N}\sum_{i=1}^{N} \ell_\phi(h_1(Z_i), \mathbf{E}[h_1(Z) \mid W = W_i]) - D_\phi(h_1(Z), \mathbf{E}[h_1(Z) \mid W]) && (203)\\
& \leq 2 \sup_{g_1 \in \mathcal{H}_z,\ g_2 \in \mathcal{H}_w} \left| \frac{1}{N}\sum_{i=1}^{N} \ell_\phi(g_1(Z_i), g_2(W_i)) - \mathbf{E}[\ell_\phi(g_1(Z), g_2(W))] \right| && (204)\\
& \leq 2\psi(\{(Z_i, W_i)\}_{i=1}^N; \mathcal{H}_z, \mathcal{H}_w). && (205)
\end{aligned}
$$

□

Lemma C.6. *Assume* $c_1 \coloneqq \inf_{h_u \in \mathcal{H}_u} \mathbf{E}[\|Y - h_u(U)\|_2^2] > 0$. *Let* $B_1 \coloneqq \|Y - h_u(U)\|_2^2$. *Let* $\widehat{\mathbf{E}}[\cdot]$ *denote the empirical average using the empirical measure defined by the training data. Then, it holds that*

$$
\left| \mathbf{E}[B_1] - \widehat{\mathbf{E}}[B_1] \right| \leq e_{1,n'} \tag{206}
$$

uniformly over the choice of $h_u \in \mathcal{H}_u$ *with probability at least* $1 - \delta$, *where*

$$
e_{1,n'} \coloneqq \widetilde{R}_7(n', \mathcal{H}_u) + 2C_{out}\sqrt{\frac{1}{2n'}\log\frac{2}{\delta}}, \tag{207}
$$

$$
\widetilde{R}_7(n', \phi, \mathcal{H}_u) \coloneqq 2C_{out}\sum_{j=1}^{k} \mathfrak{R}(\mathcal{H}_u^{(j)}(S')). \tag{208}
$$

Proof. Note that $B_1 \leq 2C_{\text{out}}$. Define $\psi(\cdot)$ by

$$
\psi\left(\{(u_i, y_i)\}_{i=1}^{n'}; \mathcal{H}_u\right) \coloneqq \sup_{h_1 \in \mathcal{H}_u} \left| \frac{1}{n'}\sum_{i=1}^{n'} \|h_1(u_i) - y_i\|_2^2 - \mathbf{E}[\|h_1(U), Y\|_2^2] \right|, \tag{209}
$$

where $\{(u_i, y_i)\}_{i=1}^N \subseteq \mathcal{U} \times \mathcal{Y}$, and U_i and Y_i are $\mathcal{U}$-valued and $\mathcal{Y}$-valued independent random variables, respectively. Let $\{(u_i, y_i)\}_{i=1}^{n'} \subseteq \mathcal{U} \times \mathcal{Y}$ and $\{(u_i', y_i')\}_{i=1}^{n'} \subseteq \mathcal{U} \times \mathcal{Y}$ such that $(u_i, y_i) = (u_i', y_i')$ for all $i \in [n']$ but some $j \in [n']$.

$$
\begin{aligned}
& \left| \psi\left(\{(u_i, y_i)\}_{i=1}^{n'}; \mathcal{H}_u, \mathcal{H}_y\right) - \psi\left(\{(u_i', y_i')\}_{i=1}^{n'}; \mathcal{H}_u, \mathcal{H}_y\right) \right| && (210)\\
& \leq \frac{1}{n'} \sup_{h_1 \in \mathcal{H}_u, h_2 \in \mathcal{H}_y} \left| \|h_1(u_j) - y_j\|_2^2 - \|h_1(u_j'), h_2(y_j')\|_2^2 \right| && (211)\\
& \leq \frac{1}{n'} \sup_{h_1 \in \mathcal{H}_u, h_2 \in \mathcal{H}_y} \max\{\|h_1(u_j), (y_j)\|_2^2, \|h_1(u_j'), y_j'\|_2^2\} && (212)\\
& \leq \frac{2C_{\text{out}}}{n'}, && (213)
\end{aligned}
$$

where we used $|a - b| \leq \max\{a - b, b - a\} \leq \max\{a, b\}$ for any $(a, b) \in [0, \infty)^2$. From McDiarmid's inequality, for any $\delta > 0$ the following holds with probability at least $1 - \delta$:

$$
\psi\left(\{(u_i, y_i)\}_{i=1}^{n'}; \mathcal{H}_u, \mathcal{H}_y\right) \leq \mathbf{E}\left[\psi\left(\{(u_i, y_i)\}_{i=1}^{n'}; \mathcal{H}_u, \mathcal{H}_y\right)\right] + \sqrt{\frac{2C_{\text{out}}^2}{n'}\log\frac{1}{\delta}}. \tag{214}
$$

The first term of the right hand side can be bounded as

$$\mathbf{E}\left[\psi\left(\{(U_i,Y_i)\}_{i=1}^N;\mathcal{H}_{\mathrm{u}}\right)\right] \tag{215}$$

$$\leq \mathbf{E}\left[\frac{1}{N}\sup_{h_1\in\mathcal{H}_{\mathrm{u}},h_2\in\mathcal{H}_{\mathrm{y}}}\left|\sum_{i=1}^N\|h_1(U_i')-Y_i'\|_2^2-\sum_{i=1}^N\|h_1(\widetilde{U}_i')-\widetilde{Y}_i'\|_2^2\right|\right]$$

(from Jensen's inequality, where $(\widetilde{U}_i',\widetilde{Y}_i)$ is an independent copy of (U_i',Y_i')) (216)

$$\leq \mathbf{E}\left[\frac{1}{N}\sup_{h_1\in\mathcal{H}_{\mathrm{u}},h_2\in\mathcal{H}_{\mathrm{y}}}\sum_{i=1}^N\varepsilon_i\|h_1(U_1')-Y_i'\|_2^2\right] \tag{217}$$

$$\leq \sum_{j=1}^k\mathfrak{R}(\{(u,y)\mapsto([h_1(u)]_j-[y]_j)^2\mid h_1\in\mathcal{H}_{\mathrm{u}}\}(S')) \tag{218}$$

$$\leq 2C_{\mathrm{out}}\sum_{j=1}^k\mathfrak{R}(\{u\mapsto[h_1(u)]_j\mid h_1\in\mathcal{H}_{\mathrm{u}}\}(S')) \tag{219}$$

$$= 2C_{\mathrm{out}}\sum_{j=1}^k\mathfrak{R}(\mathcal{H}_{\mathrm{u}}^{(j)}(S')), \tag{220}$$

where we used the Ledoux-Talagrand contraction lemma. Combining the above with 214, we conclude the claim of Lemma C.6. □

Lemma C.7. *Assume* $c_2 \coloneqq \inf_{h_{\mathrm{x}}\in\mathcal{H}_{\mathrm{x}},h_{\mathrm{u}}\in\mathcal{H}_{\mathrm{u}}}\mathbf{E}[\|\nabla\phi\circ h_{\mathrm{x}}(X)-\nabla\phi\circ h_{\mathrm{u}}(U)\|_2^2]>0$. *Let* $B_2\coloneqq\|\nabla\phi\circ h_{\mathrm{x}}(X)-\nabla\phi\circ h_{\mathrm{u}}(U)\|_2^2$. *Let* $\widehat{\mathbf{E}}[\cdot]$ *denote the empirical average using the empirical measure defined by the training data. Then, it holds that with probability at least* $1-\delta$*,*

$$\left|\mathbf{E}[B_2]-\widehat{\mathbf{E}}[B_2]\right|\leq e_{2,n} \tag{221}$$

uniformly over the choice of $h_{\mathrm{x}}\in\mathcal{H}_{\mathrm{x}}$ *and* $h_{\mathrm{u}}\in\mathcal{H}_{\mathrm{u}}$*, where*

$$e_{2,n}\coloneqq\widetilde{R}_8(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}})+2L_\phi C_{out}\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}, \tag{222}$$

$$\widetilde{R}_8(n,\phi,\mathcal{H}_{\mathrm{x}},\mathcal{H}_{\mathrm{u}})\coloneqq 2L_\phi\sum_{j=1}^k(\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_{\mathrm{x}}(S))+\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_u(S'))). \tag{223}$$

Proof. Since $B_2\leq 2L_\phi C_{\mathrm{out}}$, similarly to the proof of Lemma C.6, it holds that with probability at least $1-\delta$,

$$\sup_{h_{\mathrm{x}}\in\mathcal{H}_{\mathrm{x}},h_{\mathrm{u}}\in\mathcal{H}_{\mathrm{u}}}\left|\frac{1}{n}\sum_{i=1}^n\|\nabla\phi\circ h_{\mathrm{x}}(X_i)-\nabla\phi\circ h_{\mathrm{u}}(U_i)\|_2^2-\mathbf{E}[\|h_{\mathrm{x}}(X)-h_{\mathrm{u}}(U)\|_2^2]\right| \tag{224}$$

$$\leq\mathfrak{R}(\{(x,u)\mapsto\|\nabla\phi\circ h_{\mathrm{x}}(x)-\nabla\phi\circ h_{\mathrm{u}}(u)\|_2^2\mid h_{\mathrm{x}}\in\mathcal{H}_{\mathrm{x}},h_{\mathrm{u}}\in\mathcal{H}_{\mathrm{u}}\}(S))+\sqrt{\frac{2L_\phi^2C_{\mathrm{out}}^2}{n}\log\frac{1}{\delta}} \tag{225}$$

$$\leq 2L_\phi\sum_{j=1}^k(\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_{\mathrm{x}}(S))+\mathfrak{R}(\nabla^{(j)}\phi\circ\mathcal{H}_u(S')))+\sqrt{\frac{2L_\phi^2C_{\mathrm{out}}^2}{n}\log\frac{1}{\delta}}. \tag{226}$$

□

D One-dimensional example

We present a one-dimensional example for illustrating Theorem B.1.

To illustrate our objective function, let us consider a one-dimensional linear model. Let X be a real-valued random variable, $Y=a_{\mathrm{y}}U+b_{\mathrm{y}}+\varepsilon_{\mathrm{y}}$, and $U=a_{\mathrm{u}}X+b_{\mathrm{u}}+\varepsilon_{\mathrm{u}}$, where $a_{\mathrm{y}},b_{\mathrm{y}},a_{\mathrm{u}},b_{\mathrm{u}}$ are constant real numbers, and $\varepsilon_{\mathrm{y}},\varepsilon_{\mathrm{u}}$ are independent

normal variables. $\phi\colon \mathbb{R} \ni t \mapsto \frac{1}{2}t^2 \in \mathbb{R}$. Then,

$$\begin{aligned}
&2\|Y - \mathbf{E}[Y \mid U]\|_{L^2} \\
&= 2\|a_{\mathrm{y}}U + b_{\mathrm{y}} + \varepsilon_{\mathrm{y}} - \mathbf{E}[a_{\mathrm{y}}U + b_{\mathrm{y}} + \varepsilon_{\mathrm{y}} \mid U]\|_{L^2} \\
&= 2\|\varepsilon_{\mathrm{y}}\|_{L^2} = 2\sigma_{\mathrm{y}},
\end{aligned}$$

and

$$\begin{aligned}
&2\|\nabla\phi(\mathbf{E}[Y \mid U]) - \nabla\phi(\mathbf{E}[Y \mid X])\|_{L^2} \\
&= 2\|\mathbf{E}[Y \mid U] - \mathbf{E}[Y \mid X]\|_{L^2} \\
&= 2\|\mathbf{E}[a_{\mathrm{y}}U + b_{\mathrm{y}} + \varepsilon_{\mathrm{y}} \mid U] - \mathbf{E}[a_{\mathrm{y}}U + b_{\mathrm{y}} + \varepsilon_{\mathrm{y}} \mid X]\|_{L^2} \\
&= 2a_{\mathrm{y}}\|a_{\mathrm{u}}X + b_{\mathrm{u}} + \varepsilon_{\mathrm{u}} - \mathbf{E}[a_{\mathrm{u}}X + b_{\mathrm{u}} + \varepsilon_{\mathrm{u}} \mid X]\|_{L^2} \\
&= 2a_{\mathrm{y}}\|\varepsilon_{\mathrm{u}} - \mathbf{E}[\varepsilon_{\mathrm{u}} \mid X]\|_{L^2} = 2\|\varepsilon_{\mathrm{u}}\|_{L^2} = 2a_{\mathrm{y}}\sigma_{\mathrm{u}}.
\end{aligned}$$

In total, the irreducible error is $2\sigma_{\mathrm{y}} + 2a_{\mathrm{y}}\sigma_{\mathrm{u}}$.

E $B_\phi^\times(\dots)$ is Tighter than $B_\phi^+(\dots)$

We can show that $B_\phi^\times(\dots)$ is tighter than $B_\phi^+(\dots)$ as follows.

$$\begin{aligned}
&D_\phi(Y, h_{\mathrm{u}}(U)) + D_\phi(h_{\mathrm{u}}(U), h_{\mathrm{x}}(X)) \\
&\quad + \|Y - h_{\mathrm{u}}(U)\|_{L^2} \times \|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|_{L^2} \\
&\le D_\phi(Y, h_{\mathrm{u}}(U)) + D_\phi(h_{\mathrm{u}}(U), h_{\mathrm{x}}(X)) \\
&\quad + \frac{1}{2}\|Y - h_{\mathrm{u}}(U)\|_{L^2}^2 + \frac{1}{2}\|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|_{L^2}^2,
\end{aligned}$$

where we used the inequality $ab \le (a^2 + b^2)/2$.

F Examples with Different Loss Functions

We show two examples by instantiating the Bregman loss with specific functions ϕ.

F.1 Squared loss

Setting $\phi(t) = \frac{1}{2}\|t\|_2^2$ in the Bregman loss yields the squared loss:

$$\ell_\phi(y_1, y_2) = \frac{1}{2}\|y_1 - y_2\|^2.$$

The gradient of ϕ are

$$[\nabla\phi(\boldsymbol{v})]_j = 2[\boldsymbol{v}]_j.$$

The objective function of BregMU-ProdUB will be

$$\begin{aligned}
\widehat{B}_\phi^\times(h_{\mathrm{x}}, h_{\mathrm{u}}, s) &= \frac{1}{2n'}\sum_{i=1}^{n'}\|Y_i' - h_{\mathrm{x}}(U_i')\|_2^2 + \frac{1}{2n}\sum_{i=1}^{n}\|h_{\mathrm{u}}(U_i) - h_{\mathrm{x}}(X_i)\|_2^2 \\
&\quad + s\sqrt{\frac{1}{n'}\sum_{i=1}^{n'}\|Y_i' - h_{\mathrm{u}}(U_i')\|_2^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|h_{\mathrm{u}}(U_i) - h_{\mathrm{x}}(X_i)\|_2^2} \\
&= \frac{1}{2}\left(\sqrt{\frac{1}{n'}\sum_{i=1}^{n'}\|Y_i' - h_{\mathrm{x}}(U_i')\|_2^2} + s\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|h(U_i) - h_{\mathrm{x}}(X_i)\|_2^2}\right)^2.
\end{aligned}$$

Minimization of this objective is equivalent to minimizing

$$\sqrt{\frac{1}{n'}\sum_{i=1}^{n'}\|Y_i' - h_{\mathrm{x}}(U_i')\|_2^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|h(U_i) - h_{\mathrm{x}}(X_i)\|_2^2}.$$

F.2 Generalized I-divergence

Another example is the generalized I-divergence (Banerjee et al., 2005b). For $\phi(v) = \sum_{j=1}^{k} [v]_j \log [v]_j$ ($v \in (0,\infty)^k$), whose gradient is

$$[\nabla \phi(v)]_j = \log [v]_j + 1, \text{ for } j = 1, \dots, k,$$

the Bregman loss ℓ_ϕ will be instantiated as the *generalized I-divergence*:

$$\ell_\phi(y_1, y_2) = \sum_{j=1}^{k} [y_1]_j \log \left(\frac{[y_1]_j}{[y_2]_j} \right) - \sum_{j=1}^{k} ([y_1]_j - [y_2]_j), \tag{227}$$

where $y_1, y_2 \in (0,\infty)^k$.

F.3 KL-divergence

When y_1 and y_2 are normalized vectors in Eq. (227), i.e., $\sum_{j=1}^{k} [y_1]_j = \sum_{j=1}^{k} [y_2]_j = 1$, the generalized I-divergence recovers the KL-divergence as a special case:

$$\ell_\phi(y_1, y_2) \coloneqq \sum_{j=1}^{k} [y_1]_j \log \left(\frac{[y_1]_j}{[y_2]_j} \right), \tag{228}$$

which is a popular divergence for comparing probability density functions.

F.4 Cross-entropy Loss

Define the *cross-entropy loss* as

$$\ell_{\mathrm{CE}}(y_1, y_2) = -\sum_{j=1}^{k} [y_1]_j \log([y_2]_j)$$

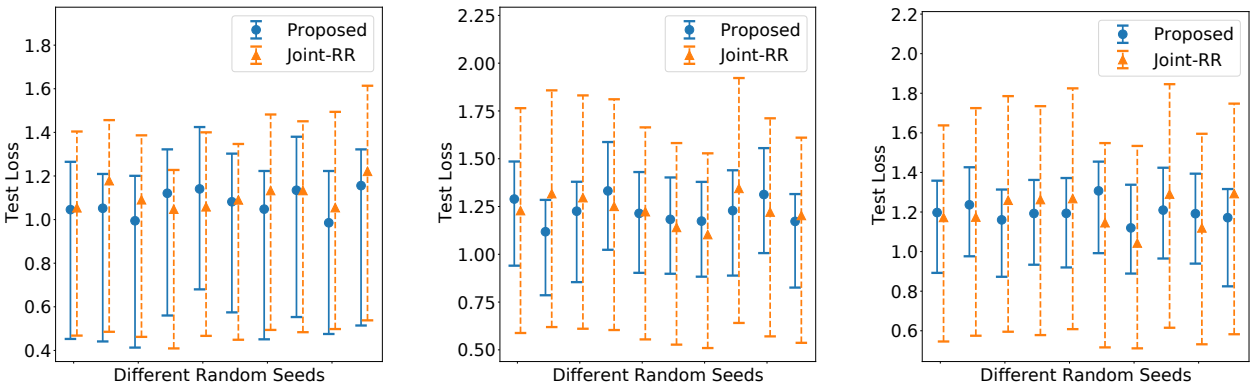
for any $y_1 \in [0,1]^k$ and $y_2 \in (0,1]^k$ such that $\sum_{j=1}^{k} [y_1]_j = \sum_{j=1}^{k} [y_2]_j = 1$, where $[\cdot]_j$ denotes the j-th component of the vector in the argument. Minimizing the expected cross-entropy loss is equivalent to minimizing the KL-divergence in the following sense:

$$\mathbf{E}[\ell_{\mathrm{CE}}(Y, f(X))] = \mathbf{E}[\underbrace{\ell_\phi(\mathbf{E}[Y \mid X], f(X))}_{\text{KL-divergence}}] - \underbrace{\mathbf{E}\left[\sum_{j=1}^{k} \mathbf{E}[[Y]_j \mid X] \log(\mathbf{E}[[Y]_j \mid X]) \right]}_{\text{Constant that does not depend on } f.},$$

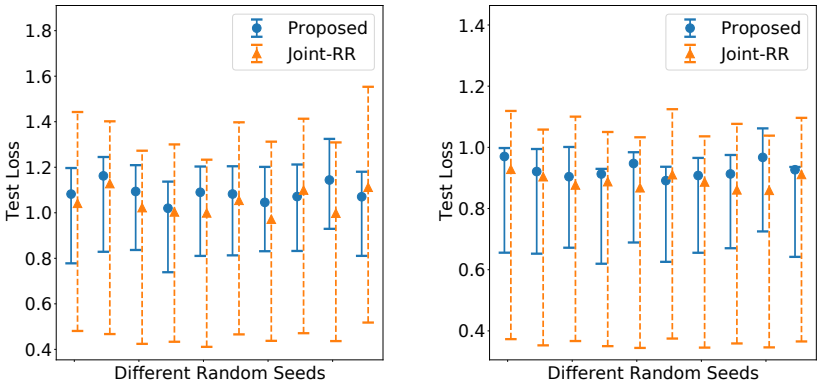
where Y is a k-dimensional random variable of a one-hot vector, $f\colon \mathcal{X} \to [0,1]^k$ such that $\sum_{j=1}^{k} [f(x)]_j = 1$ for all $x \in \mathcal{X}$, and $\phi\colon t \mapsto \sum_{j=1}^{k} [t]_j \log [t]_j$. Hence, when we have function models $h_{\mathrm{x}}\colon \mathcal{X} \to \mathcal{Y}$ and $h_{\mathrm{u}}\colon \mathcal{U} \to \mathcal{Y}$ whose outputs are positive and normalized, e.g., by a softmax layer, the empirical version of the objective function of Joint-BregMU (Eq. (8) in Algorithm 1) is

$$\widehat{B}^{\times}_{\phi}(h_{\mathrm{x}}, h_{\mathrm{u}}, s) = -\frac{1}{n'} \sum_{i=1}^{n'} \sum_{j=1}^{k} [Y'_i]_j \log [h_{\mathrm{u}}(U'_i)]_j + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} [h_{\mathrm{u}}(U_i)]_j \log \frac{[h_{\mathrm{u}}(U_i)]_j}{[h_{\mathrm{u}}(X_i)]_j}$$
$$+ s \sqrt{\frac{1}{n'} \sum_{i=1}^{n'} \sum_{j=1}^{k} ([Y'_i]_j - [h_{\mathrm{u}}(U'_i)]_j)_2^2} \times \sqrt{\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(\log [h_{\mathrm{u}}(U_i)]_j - \log [h_{\mathrm{x}}(X_i)]_j \right)^2} + \text{constant}.$$

G Figures for Interval Estimates



(a) $n = n' = 1000$. (b) $n = n' = 2000$. (c) $n = n' = 3000$.



(d) $n = n' = 5000$. (e) $n = n' = 10000$.

Figure 2: Results for the interval estimate experiment with the synthetic data.

H Details of Experiments

We conducted experiments using a computer with

- CPU: AMD EPYC (with IBPB) (16) @ 3.800GHz,
- GPU: NVIDIA GeForce GTX 1080 Ti,
- Memory: 64058MiB.

Regression with synthetic data: Similarly to Yamane et al. (2021), we consider the following setup for (X, U, Y): X follows the uniform distribution over $[-1, 1]^{10}$, $[U]_j = [X]_j^3 + [\varepsilon_{\mathrm{u}}]_j$ for all $j \in \{1, \ldots, 10\}$, $Y = \|U\|_2^2 + \varepsilon_{\mathrm{y}}$, $\varepsilon_{\mathrm{u}} \sim \mathcal{N}(0, 0.5I_{10})$, and $\varepsilon_{\mathrm{y}} \sim \mathcal{N}(0, 0.001)$. Recall that $[\cdot]_j$ denotes the j-th element of the vector in the argument. We generate independent MU-data $\{(X_i, U_i)\}_{i=1}^n$ and $\{(U_i', Y_i')\}_{i=1}^{n'}$ identically distributed to (X, U) and (U, Y), respectively. The evaluation metric is $D_\phi(\cdot, \cdot)$ with $\phi(t) = \frac{1}{2}\|t\|_2^2$, i.e., the mean squared error (MSE), so that we can directly apply Joint-RR (Yamane et al., 2021) (see Section 3.3). We train models with the proposed Joint-BregMU (see Section 4.2 and 4.2.1) and Joint-RR with different sizes of MU-data, $n = n' \in \{1000, 2000, 3000, 5000, 10000\}$. For the trained models, we compare the interval estimates given by $\widehat{B^{\times}_{\phi\ \phi}}(\cdot, \cdot, \cdot)$ and $\widehat{B^{+}_{\phi\ \phi}}(\cdot, \cdot, \cdot)$, respectively, calculated with validation MU-data.

For all models, we used 4-layer multi-layer perceptrons with the ReLU activation and 50 hidden units in each hidden layer. We trained the models using Adam with no weight decay, learning rate 0.001, batch size 512 for 1000 epochs. We set the other parameters of Adam as in the default provided by PyTorch.

For Joint-BregMU, we perform warm-starting initialized by 2Step-BregMU. We found that this significantly accelerates the training.

We used MU-data of size $n = n' \in \{1000, 2000, 3000, 5000, 10000\}$ for training. We used 10000 validation MU-data for estimating the interval for predicting test losses. For calculating test losses for evaluation, we used 10000 (X, Y)-data.

We implemented the methods based on the code of Yamane et al. (2021) licensed under the GNU General Public License v3.0. Our code can be found in the supplemental material and will be published under the same license.

Classification of Low-Resolution Images: We used four standard image classification benchmark datasets, MNIST (LeCun et al., 2010), FashionMNIST (Xiao et al., 2017), CIFAR10 (Krizhevsky, 2009), and CIFAR100 (Krizhevsky, 2009), but we modified images to artificially create low-resolution images. More specifically, for each image and its class label in each benchmark dataset, we define X as a down-sampled image, U as the original image, and Y as the class label. For training data, we take 10000 subsamples of (X, U) as $\{(X_i, U_i)\}_{i=1}^n$ and 10000 subsamples of (U, Y) as $\{(U_i', Y_i')\}_{i=1}^{n'}$. For calculating test losses for evaluation, we used 10000 (X, Y)-data.

The task being classification, we use the zero-one loss as the test evaluation metric. For training, we use the cross-entropy (Section 4.2.1) as the surrogate loss for the proposed methods but the squared loss for the previous methods because of its limitation.

For the naive method, we used a U-Net (Ronneberger et al., 2015) implemented by Linder-Norén (2018) for predicting U from X since the task is essentially image-to-image translation. For the model predicting Y from U, we used a ResNet (He et al., 2016) with 20 layers implemented by Idelbayev (2020).

In order to adapt the multi-class classification to 2Step-RR and Joint-RR, we use the "squared-softmax" layer to the output of the models as proposed by Yamane et al. (2021). For 2Step-BregMU and Joint-BregMU, we apply the ordinary softmax layer to the output.

We trained the models using Adam with no weight decay and batch size 512 for 200 epochs. We set the other parameters of Adam as in the default provided by PyTorch. We set the learning rate to 0.01 for MNIST, FashionMNIST, CIFAR10 and 0.001 for CIFAR100.

For Joint-BregMU and Joint-RR, we perform warm-starting initialized by 2Step-BregMU and 2Step-RR, respectively. We found that this significantly accelerates the training.

We implemented the methods based on the code of Yamane et al. (2021) licensed under the GNU General Public License v3.0. Our code can be found in the supplemental material and will be published under the same license.

I Maximal generality of Bregman divergences

Banerjee et al. (2005a, Theorems 3–4) proved that "under mild regularity conditions, that if $F\colon \mathbb{R}^d \times \mathbb{R}^d\colon \to \mathbb{R}$ is a non-negative loss function such that $\arg\min_{Y\in G} \mathbf{E}[F(X,Y)] = \mathbf{E}[X \mid G]$, for all random variable X, then F has to be a Bregman Loss Function (BLF)."

Theorem I.1 (Theorem 3 of Banerjee et al. (2005a))**.** *Let $F\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ be a nonnegative function such that $F(x,x)=0$, for all $x\in\mathbb{R}$. Assume that F and F_x are both continuous, where F_x denotes F's partial derivative with respect to the first argument. If for all random variables X taking values in $\mathbb{R}$, $\mathbf{E}[X]$ is the unique minimizer of $\mathbf{E}[F(X,y)]$ over all constants $y\in\mathbb{R}$, i.e.,*

$$\arg\min_{y\in\mathbb{R}} \mathbf{E}[F(X,y)] = \mathbf{E}[X],$$

then $F(x,y)=\ell_\phi(x,y)$ for some strictly convex differentiable function $\phi\colon \mathbb{R}\to\mathbb{R}$.

The following is the multi-dimensional version of their theorem, which requires slightly stronger assumptions.

Theorem I.2 (Theorem 4 of Banerjee et al. (2005a))**.** *Let $F\colon \mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ be a nonnegative function such that $F(x,x)=0$, for all $x\in\mathbb{R}^d$. Assume that $F(x_1,x_2)$ and $\frac{\partial^2 F(x_1,x_2)}{\partial[x_1]_i\partial[x_2]_j}$, $i,j\in[d]$, are all continuous. For all random variables X taking values in $\mathbb{R}^d$, if $\mathbf{E}[X]$ is the unique minimizer of $\mathbf{E}[F(X,y)]$ over all constants $y\in\mathbb{R}^d$, i.e.,*

$$\arg\min_{y\in\mathbb{R}^d} \mathbf{E}[F(X,y)] = \mathbf{E}[X],$$

then $F(x,y)=\ell_\phi(x,y)$ for some strictly convex differentiable function $\phi\colon \mathbb{R}^d\to\mathbb{R}$.

J Discussions on the Assumptions

In this section, we discuss the assumptions used in the paper.

J.1 Difference between our assumptions and those of Yamane et al. (2021)

Assumptions (i-iv) are commonly assumed in Yamane et al. (2021) and our paper. Additional assumptions of ours compared with Yamane et al. (2021) are as follows.

- The analysis of the two-step method assumes (v) to bound the target risk using the two objectives minimized in the two steps.
- The analysis of the one-step method assumes that $Y - h_{\mathrm{u}}(U)$ and $\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))$ are not almost surely zero. This is satisfied when the compared terms do not have deterministic relationships. Otherwise, we may add very small random noise to the variables to ensure the conditions.

J.2 Discussions on Eq. (1)

The assumption states about how informative U is, and it can be easy or hard to satisfy depending on the cost of collecting such data. Yamane et al. (2021) assumed the same assumption, and they proved a mini-max lower bound showing that the worst-case L^2 error is at least $\epsilon/\sqrt{2}$ when the assumption is relaxed as $\|\mathbf{E}[Y|U] - \mathbf{E}[Y|U,X]\|_{L^2} \le \epsilon$ (Yamane et al., 2021, Section 5.5). Intuitively, there is a trade-off between the bias and the violation of the assumption, and if we do not allow bias, $\epsilon = 0$ (i.e., Eq. (1)) is necessary for any estimator. Note that Eq. (1) only concerns the conditional expectation, which is weaker than the conditional independence.

On the other hand, the methods of Yamane et al. (2021) and our methods would suffer asymptotic bias at most ϵ in L^2 since what they do is essentially estimating $\mathbf{E}[\mathbf{E}[Y|U]|X]$ and the gap from $\mathbf{E}[Y|X]$ is at most $\|\mathbf{E}[\mathbf{E}[Y|U] - \mathbf{E}[Y|U,X]|X]\|_{L^2} \le \mathbf{E}[\|\mathbf{E}[Y|U] - \mathbf{E}[Y|U,X]\|_{L^2}|X] \le \epsilon$. It would be interesting to extend this result to the general Bregman divergence, which is future work.

J.3 Discussions on the assumptions of 2Step-BregMU

Assumption (iv) is Lipschitz-continuity (i.e., the boundedness of the gradient and hence the sensitivity) of ϕ. Note that Assumption (iv) is only required on a restricted domain and is relatively easy to satisfy when the domain is bounded as we assume in the paper. The first condition of Assumption (v) is about the strength of the convexity: larger p means stronger convexity. The second condition of (v) is about the strength of the smoothness: larger q means stronger smoothness. Those conditions are orthogonal to each other and do not contradict each other.

We give three examples that satisfy the Lipschitz-continuity and Assumption (v) here.

The KL-divergence: The KL-divergence satisfies the strong convexity in L^1 (it is known as Pinsker's inequality):

$$\frac{1}{2}\|g_1 - g_2\|_{L^1}^2 \leq D_\phi(g_1, g_2), \tag{229}$$

where $\phi\colon \mathbb{R}^d \ni t \mapsto \sum_{j=1}^d [t]_j \log[t]_j \in \mathbb{R}$. When $p_{\text{inf}} \coloneqq \inf_{g\in\mathcal{H}, x\in\mathcal{X}, j=1,\dots,d}[g]_j > 0$, we also have

$$\begin{aligned}\|\nabla\phi(g_1) - \nabla\phi(g_2)\|_\infty &= \sup_{x\in\mathcal{X},\ j\in[d]} |\log[g_1]_j - \log[g_2]_j| \\ &\leq 2 \sup_{x\in\mathcal{X},\ j\in[d],\ h\in\mathcal{H}} |\log[g]_j| \\ &\leq 2\log p_{\text{inf}}.\end{aligned}$$

Note that $g \in (0, 1]$. Therefore, Assumption (v) holds with $p = 1$, $q = \infty$, $\alpha = 2$, and $\beta = \infty$.

The squared loss.

$$|\phi(g_1) - \phi(g_2)| = |(g_1 + g_2)^\top (g_1 - g_2)| \leq 2C_{\text{out}}\|g_1 - g_2\|_{L^2},$$

where $C_{\text{out}} := \sup \mathcal{Y} \cup \mathcal{H}_{\text{x}}(\mathcal{X}) \cup \mathcal{H}_{\text{u}}(\mathcal{U})$, i.e., the Lipschitz continuity holds. Also,

$$D_\phi(g_1, g_2) = \|g_1 - g_2\|_{L^2}^2 = \frac{1}{4}\|\nabla\phi(g_1) - \nabla\phi(g_2)\|_{L^2}^2,$$

and thus Assumption (v) holds with $p = q = \alpha = \beta = 2$.

The loss corresponding to $\phi(t) = \|t\|^4$. $|\phi(g_1) - \phi(g_2)| = |\|g_1\|_{L^4}^4 - \|g_2\|_{L^4}^4| \leq 3C_{\text{out}}^3\|g_1 - g_2\|_{L^2}$, i.e., the Lipschitz continuity holds. Also, $D_\phi(g_1, g_2) = \mathbf{E}\left[\frac{2}{3}\sum_{j=1}^d [g_1 + 2g_2]_j^2 \times [g_1 - g_2]_j^2 + \frac{1}{3}\sum_{j=1}^d [g_1 - g_2]^4\right]$. Thus, $\frac{1}{3}\|g_1 - g_2\|_{L^4}^4 \leq D_\phi(g_1, g_2)$ and $\|\nabla\phi(g_1) - \nabla\phi(g_2)\|_{L^4}^4 = 4^4\mathbf{E}[([g_1]_j^3 - [g_2]_j^3)^4] \leq 4^4 \times 3^4 C_{\text{out}}^4\|g_1 - g_2\|_{L^4}^4 \leq 4^4 \times 3^5 C_{\text{out}}^4 D_\phi(g_1, g_2)$, i.e., Assumption (v) holds with $p = q = \alpha = \beta = 4$.

As we can see, the boundedness of the function classes is critical for these examples. We conjecture that Assumption (v) holds for $\phi(t) = \|t\|^{2k}$ with any positive integer k, but we do not have a proof yet.

Note that the theorem for the one-step method does not require Assumption (v) and does not have this weakness.

The proposed method has many assumptions in the theory, but this does not mean the proposed method needs stronger assumptions. In fact, it relaxes the strong condition of the previous approach in which the loss function must be the squared loss. All of our results apply to the squared loss.

On the other hand, our proposed one-step method minimizes

$$\begin{aligned}\widehat{B}_\phi^\times(h_{\text{x}}, h_{\text{u}}, +1) = \widehat{D}_\phi(Y, h_{\text{u}}(U)) + \widehat{D}_\phi(h_{\text{u}}(U), h_{\text{x}}(X)) \\ + \sqrt{\frac{1}{n'}\sum_i \|Y_i' - h_{\text{u}}(U_i')\|_2^2} \times \sqrt{\frac{1}{n}\sum_i \|\nabla\phi(h_{\text{u}}(U_i)) - \nabla\phi(h_{\text{x}}(X_i))\|_2^2}.\end{aligned} \tag{230}$$

Using the inequality $(a + b)/2 \leq \sqrt{a}\sqrt{b}$ that holds for any $a, b \geq 0$, one can show that the latter objective function is a lower bound of the former, and thus the latter is a tighter approximation to $D_\phi(Y, h_{\text{x}}(X))$ because of Proposition 6.1 and Lemma 4.2.

K Dominated Convergence in Probability for Conditional Expectations

Here, we will show the following fact using the dominated convergence theorem in terms of almost sure convergence.

Proposition K.1. $\mathbf{E}[Y_n \mid X] \xrightarrow{P} \mathbf{E}[Y \mid X]$ *for any sequence of random variables* $\{Y_n\}_{n=1}^{\infty}$ *when* $Y_n \xrightarrow{P} Y$ *and* $|Y_n| \leq U$ *for some* $U \in L^1$*, where* $\sigma(X)$ *is the sigma-algebra generated by* X*.*

Proof. Since any sequence of random variables converging in probability has a sub-sequence converging almost surely, we can take a sub-sequence $\{Y_{n_k}\}_{k=1}^{\infty}$ of $\{Y_n\}_{n=1}^{\infty}$ that converges almost surely. Resnick (2014, Section 10.3) states the dominated almost sure convergence of the conditional expectations:

$$\mathbf{E}[Y_{n_k} \mid X] \to \mathbf{E}[Y \mid X] \text{ a.s.}$$

The fact that the sequence $\{\mathbf{E}[Y_n \mid X]\}_{n=1}^{\infty}$ has a sub-sequence $\{\mathbf{E}[Y_{n_k} \mid X]\}_{k=1}^{\infty}$ converging almost surely implies that $\{\mathbf{E}[Y_n \mid X]\}_{n=1}^{\infty}$ converges in probability to the same limit. □

L Proof of Theorem 5.3

Fix $s \in \{-1, 1\}$. We have

$$\left|\widetilde{B}_{\phi}^{\times}(h_{\mathrm{u}}, h_{\mathrm{x}}, s) - B_{\phi}^{\times}(h_{\mathrm{u}}, h_{\mathrm{x}}, s)\right| \leq A_{5,1} + A_{5,2} + A_{5,3}, \tag{231}$$

where

$$\begin{aligned}
A_{5,1} &\coloneqq \left|\widetilde{\mathbf{E}}[\ell_{\phi}(Y, h_{\mathrm{u}}(U))] - \mathbf{E}[\ell_{\phi}(Y, h_{\mathrm{u}}(U))]\right|, \\
A_{5,2} &\coloneqq \left|\widetilde{\mathbf{E}}[\ell_{\phi}(h_{\mathrm{u}}(U), h_{\mathrm{x}}(X))] - \mathbf{E}[\ell_{\phi}(h_{\mathrm{u}}(U), h_{\mathrm{x}}(X))]\right|, \\
A_{5,3} &\coloneqq \Bigg|\sqrt{\widetilde{\mathbf{E}}[\|Y - h_{\mathrm{u}}(U)\|_2^2]} \times \sqrt{\widetilde{\mathbf{E}}[\|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|]} \\
&\qquad - \sqrt{\mathbf{E}[\|Y - h_{\mathrm{u}}(U)\|_2^2]} \times \sqrt{\mathbf{E}[\|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|]}\Bigg|,
\end{aligned}$$

where $\widetilde{\mathbf{E}}$ denotes the empirical average using i.i.d. samples $\{(\widetilde{X}_i, \widetilde{U}_i)\}_{i=1}^{\widetilde{n}} \sim P_{X,U}$ and $\{(\widetilde{U}'_i, \widetilde{Y}'_i)\}_{i=1}^{\widetilde{n}'} \sim P_{U,Y}$. From Hoeffding's inequality and the union bound, for any $\delta > 0$, with probability at least $1 - \delta/4$,

$$A_{5,1} \leq \sqrt{\frac{C_{\text{loss}}^2}{2\widetilde{n}'} \log \frac{16}{\delta}}, \tag{232}$$

$$A_{5,2} \leq \sqrt{\frac{C_{\text{loss}}^2}{2\widetilde{n}} \log \frac{16}{\delta}}. \tag{233}$$

Let $B_{5,1} \coloneqq \|Y - h_{\mathrm{u}}(U)\|_2^2$ and $B_{5,2} \coloneqq \|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|_2^2$. Note that $B_{5,1} \leq 2C_{\text{out}}^2$ and $B_{5,2} \leq 2L_{\phi}C_{\text{out}}^2$. Denote $\delta_{5,1} \coloneqq \widetilde{\mathbf{E}}[B_{5,1}] - \mathbf{E}[B_{5,1}]$ and $\delta_{5,2} \coloneqq \widetilde{\mathbf{E}}[B_{5,2}] - \mathbf{E}[B_{5,2}]$. From Hoeffding's inequality and the union bound, with probability at least $1 - \delta/4$,

$$\begin{aligned}
|\delta_{5,1}| &\leq \sqrt{\frac{2C_{\text{out}}^4}{\widetilde{n}'} \log \frac{16}{\delta}} =: e_{5,1}, \\
|\delta_{5,2}| &\leq \sqrt{\frac{2L_{\phi}^4 C_{\text{out}}^4}{\widetilde{n}} \log \frac{16}{\delta}} =: e_{5,2}.
\end{aligned}$$

Denote $c_1 \coloneqq \inf_{h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \mathbf{E}[\|Y - h_{\mathrm{u}}(U)\|_2^2] > 0$ and $c_2 \coloneqq \inf_{h_{\mathrm{x}} \in \mathcal{H}_{\mathrm{x}}, h_{\mathrm{u}} \in \mathcal{H}_{\mathrm{u}}} \mathbf{E}[\|\nabla\phi(h_{\mathrm{u}}(U)) - \nabla\phi(h_{\mathrm{x}}(X))\|_2^2] > 0$. Suppose that $\widetilde{n}$ and $\widetilde{n}'$ are sufficiently large so that $e_{5,1} \leq \frac{3}{4}c_1$ and $e_{5,2} \leq \frac{3}{4}c_2$. Then,

$$\begin{aligned}
\widetilde{\mathbf{E}}[B_{5,1}] &\geq |\mathbf{E}[B_{5,1}]| - |\widetilde{\mathbf{E}}[B_{5,1}] - \mathbf{E}[B_{5,1}]| \\
&\geq c_1 - e_{5,1} \\
&\geq c_1 - \frac{3}{4}c_1 = \frac{c_1}{4}.
\end{aligned}$$

Similarly, $\widetilde{\mathbf{E}}[B_{5,2}] \geq c_2/4$. Thus,

$$\sqrt{\widetilde{\mathbf{E}}[B_{5,1}]} - \sqrt{\mathbf{E}[B_{5,1}]} = \sqrt{\mathbf{E}[B_{5,1}] + \delta_{5,1}} - \sqrt{\mathbf{E}[B_{5,1}]} \leq \frac{\delta_{5,1}}{2\sqrt{\mathbf{E}[B_{5,1}]}} \leq \frac{e_{5,1}}{2\sqrt{c_1}},$$
$$\sqrt{\widetilde{\mathbf{E}}[B_{5,1}]} - \sqrt{\mathbf{E}[B_{5,1}]} = \sqrt{\mathbf{E}[B_{5,2}] + \delta_{5,2}} - \sqrt{\mathbf{E}[B_{5,2}]} \leq \frac{\delta_{5,2}}{2\sqrt{\mathbf{E}[B_{5,2}]}} \leq \frac{e_{5,2}}{2\sqrt{c_2}},$$
$$\sqrt{\mathbf{E}[B_{5,1}]} - \sqrt{\widetilde{\mathbf{E}}[B_{5,1}]} = \sqrt{\widetilde{\mathbf{E}}[B_{5,1}] + \delta_{5,1}} - \sqrt{\widetilde{\mathbf{E}}[B_{5,1}]} \leq \frac{\delta_{5,1}}{2\sqrt{\widetilde{\mathbf{E}}[B_{5,1}]}} \leq \frac{e_{5,1}}{\sqrt{c_1}},$$
$$\sqrt{\mathbf{E}[B_{5,2}]} - \sqrt{\widetilde{\mathbf{E}}[B_{5,2}]} = \sqrt{\widetilde{\mathbf{E}}[B_{5,2}] + \delta_{5,2}} - \sqrt{\widetilde{\mathbf{E}}[B_{5,2}]} \leq \frac{\delta_{5,2}}{2\sqrt{\widetilde{\mathbf{E}}[B_{5,2}]}} \leq \frac{e_{5,2}}{\sqrt{c_2}}.$$

Summarizing the inequalities above, we get

$$\left|\sqrt{\widetilde{\mathbf{E}}[B_{5,1}]} - \sqrt{\mathbf{E}[B_{5,1}]}\right| \leq \frac{e_{5,1}}{\sqrt{c_1}},$$
$$\left|\sqrt{\widetilde{\mathbf{E}}[B_{5,2}]} - \sqrt{\mathbf{E}[B_{5,2}]}\right| \leq \frac{e_{5,2}}{\sqrt{c_2}}.$$

Using these inequalities, we obtain

$$\begin{aligned}
A_{5,3} &= \left|\sqrt{\mathbf{E}[B_{5,1}]\,\mathbf{E}[B_{5,2}]} - \sqrt{\widetilde{\mathbf{E}}[B_{5,1}]\widetilde{\mathbf{E}}[B_{5,2}]}\right| \\
&\leq \left|\left(\sqrt{\mathbf{E}[B_{5,1}]} - \sqrt{\widetilde{\mathbf{E}}[B_{5,1}]} + \sqrt{\widetilde{\mathbf{E}}[B_{5,1}]}\right)\sqrt{\mathbf{E}[B_{5,2}]} - \sqrt{\widetilde{\mathbf{E}}[B_{5,1}]\widetilde{\mathbf{E}}[B_{5,2}]}\right| \\
&\leq \left|\sqrt{\mathbf{E}[B_{5,1}]} - \sqrt{\widetilde{\mathbf{E}}[B_{5,1}]}\right| \times \sqrt{\mathbf{E}[B_{5,2}]} + \sqrt{\widetilde{\mathbf{E}}[B_{5,1}]} \times \left|\sqrt{\mathbf{E}[B_{5,2}]} - \sqrt{\widetilde{\mathbf{E}}[B_{5,2}]}\right| \\
&\leq \sqrt{\mathbf{E}[B_{5,2}]}\frac{e_{5,1}}{\sqrt{c_1}} + \sqrt{\widetilde{\mathbf{E}}[B_{5,1}]}\frac{e_{5,2}}{\sqrt{c_2}} \\
&\leq \sqrt{\frac{2L_\phi C_{\text{out}}}{c_1}}e_{5,1} + \sqrt{\frac{2C_{\text{out}}}{c_2}}e_{5,2}.
\end{aligned} \tag{234}$$

From Eqs. (231-234) and the union bound, with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\widetilde{B}_\phi^\times(h_{\text{u}}, h_{\text{x}}, s) - B_\phi^\times(h_{\text{u}}, h_{\text{x}}, s)\right| \\
&\leq A_{5,1} + A_{5,2} + A_{5,3} \\
&\leq \sqrt{\frac{2C_{\text{loss}}^2}{\widetilde{n}}\log\frac{16}{\delta}} + \sqrt{\frac{2C_{\text{loss}}^2}{\widetilde{n}'}\log\frac{16}{\delta}} + \sqrt{\frac{4L_\phi C_{\text{out}}^5}{\widetilde{n}'}\log\frac{16}{\delta}} + \sqrt{\frac{4L_\phi^4 C_{\text{out}}^5}{\widetilde{n}}\log\frac{16}{\delta}} \\
&\leq (\sqrt{2}C_{\text{loss}} + 2L_\phi^2 C_{\text{out}}^{5/2})\sqrt{\frac{1}{\widetilde{n}}\log\frac{16}{\delta}} + (\sqrt{2}C_{\text{loss}} + 2L_\phi^{1/2} C_{\text{out}}^{5/2})\sqrt{\frac{1}{\widetilde{n}'}\log\frac{16}{\delta}}
\end{aligned}$$

for both $s = \{-1, 1\}$ at the same time.