Information and Inference: A Journal of the IMA (2012) **1**, 68–97
doi:10.1093/imaiai/ias003
Advance Access publication on 8 November 2012

Semi-supervised single- and multi-domain regression with multi-domain training Ⓡ

Tomer Michaeli* and Yonina C. Eldar
Technion–Israel Institute of Technology, Haifa, Israel
tomermic@tx.technion.ac.il
yonina@ee.technion.ac.il

and

Guillermo Sapiro
University of Minnesota, Minneapolis, MN, USA
guille@umn.edu

[Received on 22 February 2012; revised on 29 August 2012; accepted on 29 August 2012]

We address the problems of multi- and single-domain regression based on distinct and unpaired labeled training sets for each of the domains and a large unlabeled training set from all domains. We formulate these problems as a Bayesian estimation with partial knowledge of statistical relations. We propose a worst-case design strategy and study the resulting estimators. Our analysis explicitly accounts for the cardinality of the labeled sets and includes the special cases in which one of the labeled sets is very large or, in the other extreme, completely missing. We demonstrate our estimators in the context of removing expressions from facial images and in the context of audio-visual word recognition, and provide comparisons to several recently proposed multi-modal learning algorithms.

Keywords: Bayesian estimation; partial knowledge; multi- and single-domain regression; learning; hidden relationships; Bayesian networks; minimum mean squared error..

1. Introduction

There are many applications in which one can access data from multiple domains in order to perform a task. For example, word recognition may greatly benefit from the availability of joint audio-visual measurements [20]. Person recognition and verification may be performed much more accurately by fusing information from several modalities such as facial images, iris scans, voice recordings and handwritings.

A major difficulty in fusing multiple sources is that one can often access only distinct labeled training sets for the different domains and does not have paired labeled examples from all domains. Suppose, for instance, we wish to perform audio-visual gender recognition. There are numerous existing datasets of labeled voice recordings as well as labeled datasets of facial images. However, there are only a few jointly labeled audio-visual datasets, with a limited number of different subjects each. Thus, although it is straightforward to train a classifier based on audio or image data alone, it is not clear how to best fuse the two modalities, in particular, when they are unpaired. While paired multi-domain labeled examples are typically scarce, paired unlabeled examples are often abundant. For instance, enormous amounts of speaker video sequences (together with audio) can be easily collected.

Ⓡ symbol indicates reproducible data.

These videos, though, often do not come with labels. Nonetheless, they can be used to unveil the statistical relations between audio and video. An important question is how to best fuse audio- and image-based predictors, given these relations.

An even more interesting and practical question is whether the availability of multiple data sources can aid a machine learning algorithm during training, even if not all are measured during testing. For example, suppose we want to predict the age of a person based on an audio recording of him/her. Assume we have a labeled audio training set, a labeled image training set and a large amount of unlabeled audio-visual examples. Can the visual examples help construct a predictor, which is solely based on audio?

In this paper, we address the problem of multi-domain as well as single-domain regression based on distinct (unpaired) labeled training sets for each of the domains and an unlabeled multi-domain training set. Specifically, focusing on two domains for simplicity, we consider the situation in which we have at our disposal a very large unlabeled training set $\{\boldsymbol{x}_1^i, \boldsymbol{x}_2^i\}_{i \in \mathcal{U}}$ as well as two labeled sets $\{\boldsymbol{x}_1^i, \boldsymbol{y}^i\}_{i \in \mathcal{L}_1}$ and $\{\boldsymbol{x}_2^i, \boldsymbol{y}^i\}_{i \in \mathcal{L}_2}$. Using this multi-domain training data, we treat the problems of designing a predictor of $\boldsymbol{y}$ based on $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ (multi-domain regression) and a predictor of $\boldsymbol{y}$ based on $\boldsymbol{x}_1$ alone (single-domain regression). Our analysis is general in that it explicitly accounts for the cardinality of the labeled sets. In particular, it includes the special cases in which one or both labeled sets are very large as well as the cases in which one of the labeled sets is completely missing.

Several problems of similar nature have been treated in the literature. Perhaps the most widely studied of these is *multi-view learning* [2] in general and multi-view regression [10] in particular. These techniques make use of a large training set of data from multiple domains (views), which contains only a few labeled examples. It has been shown that if the views tend to agree in some sense, then the unlabeled examples are useful in constructing a single-view estimator [2, 10]. In our setting, however, we do not observe even a single multi-domain labeled example $\{\boldsymbol{x}_1^i, \boldsymbol{x}_2^i, \boldsymbol{y}^i\}$ and also make no assumptions on the underlying distribution. A multi-view framework for distinct labeled training sets, recently proposed in [1], assumes the availability of a mapping function that can generate a good estimate of the unobserved view from the observed one. In our setting, we do not assume that such a mapping is known or even exists.

Situations in which labeled samples $\{\boldsymbol{x}_2^i, \boldsymbol{y}^i\}$ from a source domain are used to construct a predictor of $\boldsymbol{y}$ from a target domain $\boldsymbol{x}_1$ fall under the category of *transfer learning* [21]. In some cases, unlabeled examples, as well as a few labeled examples $\{\boldsymbol{x}_1^i, \boldsymbol{y}^i\}$ from the target domain are also available. Situations in which the domains do not admit a common feature representation may be handled via the *multiple-outlook learning* framework [9]. These classes of problems, however, are different than ours in that they do not assume the availability of paired unlabeled examples $\{\boldsymbol{x}_1^i, \boldsymbol{x}_2^i\}$ from the two domains.

More related to our problem are the *cross-modality* and *shared-representation* learning scenarios recently studied in [20] in the context of multi-modal learning. In both settings, unlabeled training data $\{\boldsymbol{x}_1^i, \boldsymbol{x}_2^i\}$ from multiple modalities, such as audio and video, are used to perform a *feature learning* stage. In cross-modality learning, one constructs a predictor based on $\boldsymbol{x}_1$ alone using a labeled training set $\{\boldsymbol{x}_1^i, \boldsymbol{y}^i\}$. For example, we may want to build a classifier operating on audio features by observing labeled audio examples in addition to unlabeled audio-visual instances. In shared-representation learning, one constructs a predictor based on $\boldsymbol{x}_1$ alone using a labeled training set $\{\boldsymbol{x}_2^i, \boldsymbol{y}^i\}$. For instance, we may want to train an audio classifier by observing only labeled visual examples in addition to unlabeled audio-visual instances. In [20], both the cross-modality and the shared-representation settings were treated by learning multi-modal feature representations using deep networks. Shared-representation regression was recently studied from a Bayesian estimation perspective in [16], in which a link to instrumental variable regression [3] was also highlighted. As we show, both cross-modality and shared-representation learning are special cases of our approach, corresponding to the situation in which there are zero examples in one of the labeled sets.

In this paper, we formulate regression from unpaired datasets as a Bayesian estimation problem with partial knowledge of statistical relations. Specifically, we assume that, for each domain, we can determine the predictor that minimizes the mean square error (MSE) among some class of estimators. This can be done using the labeled training examples from the associated domain. Furthermore, we assume that we can determine the joint probability distribution of the data from the two domains using the unlabeled examples. Now, every joint distribution of labels and (multi-domain) data that is consistent with this knowledge is considered valid. The performance of any estimator depends, of course, on the unknown distribution. Thus, our approach in this paper is to seek estimators whose worst-case MSE over the set of valid distributions is the smallest possible. This strategy matches that used in [16] in the context of shared-representation regression. The methods we develop here constitute generalizations of the results of [16] to arbitrary single- and multi-domain regression settings.[1]

We show that the minimax problems we obtain have simple, yet non-trivial, closed form solutions which can be easily approximated from the available training examples. These expressions also provide insight into how data from multiple domains should be taken into account. In particular, we show that, from a worst-case standpoint, a domain with no labeled examples cannot help if it is not available at test time. Thus, it is impossible to perform cross-modality regression without making any assumptions on the underlying distributions. We illustrate our approach in the contexts of face normalization and audio-visual word recognition. In the former application, we demonstrate how an image of a smiling face can be converted into one with a neutral expression, without observing paired examples of neutral and smiling faces. In the latter setting, we show how spoken digits can be recognized from silent video (lipreading) when only labeled audio examples are available. We also show how they can be recognized from audio, when there is access only to labeled video examples. The experiments indicate that our approach is preferable to that of [20].

The remainder of this paper is organized as follows. In Section 2, we present the setting of interest in detail and discuss several special cases. We provide a mathematical formulation of our regression problems in Section 3. The minimax multi- and single-domain estimators are derived in Sections 4 and 5, respectively. Finally, experimental results are provided in Section 6.

2. Problem formulation

We denote random variables (RVs) by capital letters (e.g. X_1, X_2, Y) and the values that they take by bold lower-case letters (e.g. $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}$). The pseudo-inverse of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^\dagger$. The second-order moment matrix of an RV X is denoted by $\Gamma_{XX} = \mathbb{E}[XX^T]$, where $\mathbb{E}[\cdot]$ is the mathematical expectation operator. Similarly, the cross second-order moment matrix of two RVs X and Y is denoted by $\Gamma_{XY} = \mathbb{E}[XY^\top]$. The joint cumulative distribution function of the RVs X and Y is written as $F_{XY}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{P}(X \leqslant \boldsymbol{x}, Y \leqslant \boldsymbol{y})$, where the inequalities are element-wise. By definition, the marginal distribution of X is $F_X(\boldsymbol{x}) = F_{XY}(\boldsymbol{x}, \infty)$. In our setting, Y is the quantity to be estimated, and X_1 and X_2 are two sets of measurements (features). The RVs X_1, X_2 and Y take values in $\mathbb{R}^{M_1}$, $\mathbb{R}^{M_2}$ and $\mathbb{R}^N$, respectively.

Our goal in this paper is to propose an estimation theoretic approach for solving certain regression problems in which several distinct training sets are available during training. More specifically, we assume we are given access to three possible datasets as follows:

[1] Some of the results in this paper were reported without proof in [17]. This conference version did not include the most general formulation of the theoretical results as well as some of the experiments we present here.

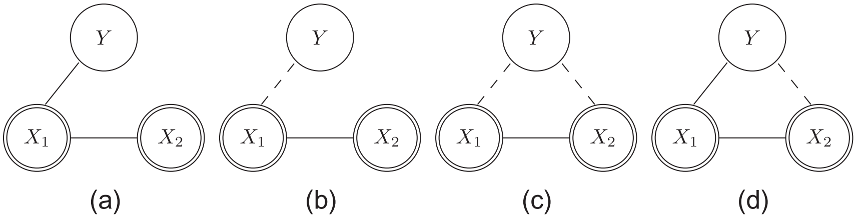

FIG. 1. Multi-domain regression. (a and b) Single-domain training with many/few labeled examples (Section 4.1). (c) Multi-domain training with few labeled examples (Sections 4.2 and 4.3). (d) Multi-domain training with many unpaired labeled examples from one domain and a few from the other domain (Sections 4.4 and 4.5).

(1) labeled examples $\{(\boldsymbol{x}_1^\ell, \boldsymbol{y}^\ell)\}_{\ell\in\mathcal{L}_1}$ from domain 1;

(2) labeled examples $\{(\boldsymbol{x}_2^\ell, \boldsymbol{y}^\ell)\}_{\ell\in\mathcal{L}_2}$ from domain 2;

(3) paired unlabeled examples $\{(\boldsymbol{x}_1^u, \boldsymbol{x}_2^u)\}_{u\in\mathcal{U}}$.

Here the index sets of the labeled and unlabeled examples do not intersect, namely $\mathcal{L}_1 = \{1, \ldots, L_1\}$, $\mathcal{L}_2 = \{L_1 + 1, \ldots, L_1 + L_2\}$ and $\mathcal{U} = \{L_1 + L_2 + 1, \ldots, L_1 + L_2 + U\}$. These training sets correspond to independent draws from the distributions F_{X_1Y}, F_{X_2Y} and $F_{X_1X_2}$, respectively. Our focus is on situations in which U is very large, so that the joint distribution $F_{X_1X_2}$ can be assumed known (or very well approximated, for example, by non-parametric methods). The cardinalities L_1 and L_2 of the labeled sets are arbitrary. In particular, one of them can be zero. In this case no knowledge whatsoever is available regarding the statistical relation between Y and the associated domain. At the other extreme, one (or both) of the labeled sets may be very large, in which case the associated single-domain minimum MSE (MMSE) estimator, say $\mathbb{E}[Y|X_1]$, can be assumed known (or accurately approximated).

In terms of testing, we treat two tasks. The first is *multi-domain regression*, in which the algorithm is asked to predict $\boldsymbol{y}$ based on an observation of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The second is *single-domain regression*, where prediction should be based solely on $\boldsymbol{x}_1$ (including the case where no $\boldsymbol{x}_1$ labeled data are available for training, that is, $L_1 = 0$). Several archetypical situations are depicted in Figs 1 and 2. Here, single- and double-lined circles correspond, respectively, to RVs that are unobserved and observed during testing. A continuous line, a dashed line and lack of a line between circles corresponds, respectively, to many, few and zero training examples.

3. Estimation theoretic formulation

In this paper, we adopt and generalize the framework proposed in [16] by posing our problem as one of estimation with partial knowledge of statistical relations. Before formalizing our multi-domain semi-supervised problem in estimation theoretic terms, we first recall the common practice for regression from one domain with a limited number of examples.

3.1 *Single-domain regression*

Suppose we are given a sample $\{\boldsymbol{x}^\ell, \boldsymbol{y}^\ell\}_{\ell=1}^L$, $\boldsymbol{x} \in \mathbb{R}^M$, independently drawn from the joint distribution of X and Y. If L is very large, then non-parametric methods can be used to approximate the conditional expectation estimator $\varphi(\boldsymbol{x}) = \mathbb{E}[Y|X = \boldsymbol{x}]$ with great accuracy at any $\boldsymbol{x}$. For example, one may use the Nadaraya–Watson non-parametric estimator [19, 25]

$$\hat{\varphi}(\boldsymbol{x}) = \frac{\sum_{\ell=1}^L \boldsymbol{y}^\ell K((\boldsymbol{x} - \boldsymbol{x}^\ell)/h)}{\sum_{\ell=1}^L K((\boldsymbol{x} - \boldsymbol{x}^\ell)/h)}, \tag{3.1}$$

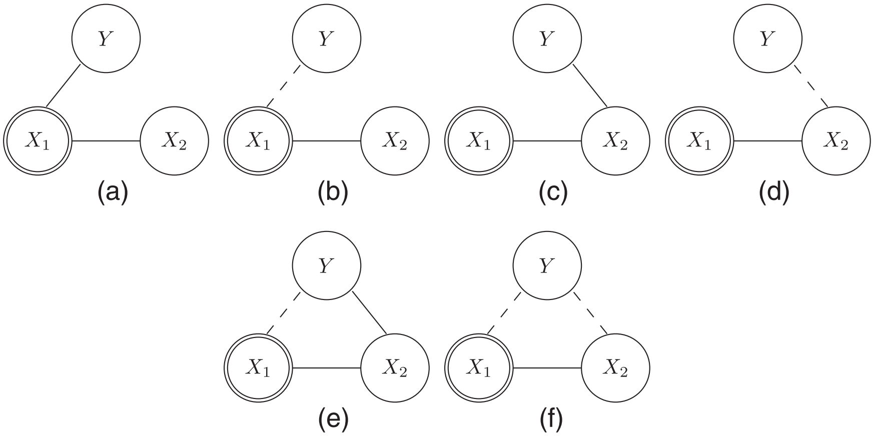

FIG. 2. Single-domain regression. (a and b) Cross-modality learning [20] with many/few labeled examples (Section 5.1). (c and d) Shared-representation regression [20], also referred to as estimation with partial knowledge [16], with many/few labeled examples (Section 5.2). (e and f) Multi-domain training with many/few labeled examples from the unobserved domain (Section 5.3).

where $K(\boldsymbol{x})$ is some density function with mean 0, called kernel, and $h > 0$ is a scalar called the bandwidth. For example, the Gaussian kernel $K(\boldsymbol{x}) = (2\pi)^{-1/2}\exp\{-\|x\|^2/2\}$ is in common use. Under mild conditions on $K(\boldsymbol{x})$, various converges properties of $\hat{\varphi}(\boldsymbol{x})$ to $\mathbb{E}[Y|X = \boldsymbol{x}]$ are known when $L \to \infty$ and $h \to 0$ at an appropriate rate. Such estimates, however, are often far from accurate when L is small. Common practice in such situations is to use parametric or semi-parametric methods that impose some structure on the sought predictor. In other words, rather than trying to approximate the regression function $\varphi(\boldsymbol{x}) = \mathbb{E}[Y|X = \boldsymbol{x}]$, which minimizes the MSE among all functions of X, we settle for approximating the optimal predictor among some family $\mathcal{A}$ of functions:

$$\varphi_{\mathcal{A}} = \arg\min_{\varphi\in\mathcal{A}} \mathbb{E}[\|Y - \varphi(X)\|^2]. \tag{3.2}$$

The less rich the class $\mathcal{A}$ is, the more accurate we can typically approximate $\varphi_{\mathcal{A}}(X)$ from the training data. This comes, of course, at the cost that the (theoretical) MSE that $\varphi_{\mathcal{A}}(X)$ achieves is higher. This is the well-known bias-variance tradeoff. In the sequel, we term the function $\varphi_{\mathcal{A}}(X)$ of (3.2) the $\mathcal{A}$-optimal estimator of Y from X.

Problem (3.2) can be given the following geometric interpretation. Let the RVs X and Y be defined over the probability triplet $(\Omega, \mathcal{F}, P)$ and let $L^2(\Omega, \mathcal{F}, P)$ denote the Hilbert space of RVs that take values in $\mathbb{R}^N$ and satisfy $\mathbb{E}[\|U\|^2] < \infty$, equipped with the inner product $\langle U, V\rangle_{L^2} = \mathbb{E}[V^\top U]$. Then problem (3.2) is equivalent to

$$\hat{Y}_{\tilde{\mathcal{A}}} = \arg\min_{\hat{Y}\in\tilde{\mathcal{A}}} \|Y - \hat{Y}\|^2_{L^2}. \tag{3.3}$$

Here, $\tilde{\mathcal{A}}$ denotes the set of all RVs in L^2, which can be expressed as $\varphi(X)$, for some function $\varphi : \mathbb{R}^M \to \mathbb{R}^N$ in $\mathcal{A}$ and $\hat{Y}_{\tilde{\mathcal{A}}} = \varphi_{\mathcal{A}}(X)$ is the RV in $\tilde{\mathcal{A}}$ which is closest to Y. This shows that the RV $\hat{Y}_{\mathcal{A}}$ is the orthogonal projection $\Pi_{\tilde{\mathcal{A}}}Y$ of the RV Y onto the set $\tilde{\mathcal{A}}$. If $\tilde{\mathcal{A}}$ is closed, then a solution is guaranteed to exist and if, in addition, $\tilde{\mathcal{A}}$ is convex (or, in particular, a linear subspace), then the solution is unique.

One of the simplest structural restrictions corresponds to linear estimation, so that $\mathcal{A}$ is the set of all linear functions from $\mathbb{R}^M$ to $\mathbb{R}^N$. In this case,

$$\varphi_{\mathcal{A}}(X) = \Gamma_{YX}\Gamma_{XX}^{\dagger}X. \tag{3.4}$$

The second-order moment matrices Γ_{YX}, Γ_{XX} can be estimated from the training set, for example, by using sample moments. A more general model corresponds to functions of the form

$$\varphi(X) = \sum_{k=1}^{K} a_k\varphi_k(X), \tag{3.5}$$

where $\{\varphi_k\}_{k=1}^K$ is a predefined set of functions and the coefficients $\{a_k\}_{k=1}^K$ are arbitrary. The optimal set of coefficients $\boldsymbol{a} = (a_1 \ \cdots \ a_K)^\top$ is given in this case by

$$\boldsymbol{a} = \Gamma_{\Phi\Phi}^{\dagger}\Gamma_{\Phi Y}, \tag{3.6}$$

where $\Gamma_{\Phi\Phi}$ denotes the $K \times K$ matrix whose (i,j)th entry is $\mathbb{E}[\varphi_i^\top(X)\varphi_j(X)]$ and $\Gamma_{\Phi Y}$ is a $K \times 1$ vector whose ith component is $\mathbb{E}[\varphi_i^\top(X)Y]$. These quantities can be estimated from the training data similar to the linear setting.

In both examples above, the set $\mathcal{A}$ of functions is linear in the sense that, for every $\varphi^1, \varphi^2 \in \mathcal{A}$ and $\alpha, \beta \in \mathbb{R}$, the function $\alpha\varphi^1 + \beta\varphi^2$ also belongs to $\mathcal{A}$. For future reference, we note that this claim is also trivially true when $\mathcal{A}$ is taken to be the set of all (Borel-measurable) functions, in which case $\varphi_{\mathcal{A}}(X) = \mathbb{E}[Y|X]$, and when $\mathcal{A}$ contains only the zero function, in which case $\varphi_{\mathcal{A}}(X) = 0$. From a geometric standpoint, the linearity of $\mathcal{A}$ implies that the set of RVs $\tilde{\mathcal{A}}$ of (3.3), onto which Y is projected, is a linear subspace of $L^2(\Omega, \mathcal{F}, P)$.

3.2 *Statistical knowledge deduced from separate training sets*

In our setting, we have access to two separate unpaired sets of labeled examples, one for each domain. Consequently, besides the standard uncertainty in statistics, which has to do with the fact that the underlying distributions are not known but rather only samples are observed, here there is another degree of uncertainty. Specifically, even if the number of training examples is taken to infinity in all three sets, we can only hope to be able to determine the joint distributions F_{X_1Y}, F_{X_2Y} and $F_{X_1X_2}$. These do not suffice in general for computing the MMSE estimate $\mathbb{E}[Y|X_1, X_2]$. To focus only on the second type of uncertainty, we assume that we are able to perform single-domain regression from each of the training sets with very small variance (at the expense of possible bias). Specifically, we assume that we can determine the $\mathcal{A}$-optimal predictor of Y given X_1 as well as the $\mathcal{B}$-optimal predictor of Y from X_2, where $\mathcal{A}$ and $\mathcal{B}$ are classes of functions chosen in accordance with the cardinality of the two sets. Note that each of the single-domain predictors may be very poor. In particular, if there are no labeled training examples from one of the domains, then we choose the corresponding class of valid predictors to contain only the zero function. Therefore, if, for instance, we have $L_1 = 0$ labeled examples from domain X_1, then we set $\mathcal{A} = \{0\}$ so that the $\mathcal{A}$-optimal predictor of Y from X_1 is simply $\varphi_{\mathcal{A}}(X_1) = 0$.

We further assume that the existence of many unlabeled examples (X_1, X_2) allows accurately determining the joint distribution of X_1 and X_2, for example, using non-parametric methods. Finally, we assume that there are enough labeled examples from at least one of the domains such that the second-order moment of Y can be accurately estimated.[2] The statistical relationships assumed known are depicted in Fig. 3.

[2] The predictors we derive turn out to be independent of the actual value of the second-order moment of Y, so that, in practice, it does not actually have to be estimated. Nevertheless, our solutions are optimal only under the assumption that the quadratic moment is known and finite.

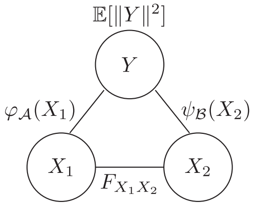

FIG. 3. Known statistical relationships. Each of the single-domain predictors may perform arbitrarily poorly (in particular, it is possible that $\varphi_{\mathcal{A}}(X_1) = 0$ or $\psi_{\mathcal{B}}(X_2) = 0$).

In a more mathematical language, assume that we are given two functions $\varphi_{\mathcal{A}} : \mathbb{R}^{M_1} \to \mathbb{R}^N$ and $\psi_{\mathcal{B}} : \mathbb{R}^{M_2} \to \mathbb{R}^N$, a cumulative probability function $F_{X_1X_2}$ over $\mathbb{R}^{M_1 \times M_2}$ and a scalar $c > 0$. Then, what we know regarding the RVs X_1, X_2 and Y is that their distribution $F_{X_1X_2Y}$ belongs to the set $\mathcal{D}$ of distributions satisfying

$$\begin{aligned}&\varphi_{\mathcal{A}} = \arg\min_{\varphi\in\mathcal{A}} \mathbb{E}[\|Y - \phi(X_1)\|^2], \quad \psi_{\mathcal{B}} = \arg\min_{\psi\in\mathcal{B}} \mathbb{E}[\|Y - \psi(X_2)\|^2],\\ &F_{X_1X_2Y}(\boldsymbol{x}_1, \boldsymbol{x}_2, \infty) = F_{X_1X_2}(\boldsymbol{x}_1, \boldsymbol{x}_2), \quad \mathbb{E}[\|Y\|^2] = c. \end{aligned} \tag{3.7}$$

We assume throughout the paper that $\mathcal{A}$ and $\mathcal{B}$ form linear sets of functions, as discussed in Section 3.1, so that the associated sets of RVs $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$ form linear subspaces of L^2. We also assume that $\mathcal{A}$ and $\mathcal{B}$ are such that the subspaces $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$ are closed in L^2. This is always the case for subspaces spanned by a finite number of functions, as in the parametric examples presented in Section 3.1, or for subspaces generated by all (Borel-measurable) functions of some RV.

As an illustrative example, suppose that X_1, X_2 and Y are scalar RVs, and that $\mathcal{A}$ and $\mathcal{B}$ are the sets of all linear functions from $\mathbb{R}$ to $\mathbb{R}$. Assume further that we know that the best linear estimator of Y from X_1 is $\varphi_{\mathcal{A}}(X_1) = 0.1X_1$, the best linear estimator of Y from X_2 is $\psi_{\mathcal{B}}(X_2) = 0.2X_2$, the probability density function (pdf) of (X_1, X_2) is $f_{X_1X_2}(x_1, x_2) \propto \exp\{-(x_1^2 + x_2^2)/2\}$ and that $\mathbb{E}[Y^2] = 1$. Then the normal density

$$f_{X_1X_2Y}(x_1, x_2, y) \propto \exp\left\{-\frac{1}{2}(x_1 \quad x_2 \quad y)\begin{pmatrix}1 & 0 & 0.1\\ 0 & 1 & 0.2\\ 0.1 & 0.2 & 1\end{pmatrix}^{-1}\begin{pmatrix}x_1\\ x_2\\ y\end{pmatrix}\right\} \tag{3.8}$$

is consistent with all these restrictions and is thus valid. In fact, there is an infinite number (a continuum) of other feasible densities. For instance, it can be easily verified that the Gaussian mixture pdf

$$\begin{aligned}f_{X_1X_2Y}(x_1, x_2, y) \propto {}& \exp\left\{-\frac{1}{2}(x_1 \quad x_2 \quad y)\begin{pmatrix}1 & 0 & 0.2\\ 0 & 1 & 0\\ 0.2 & 0 & 1\end{pmatrix}^{-1}\begin{pmatrix}x_1\\ x_2\\ y\end{pmatrix}\right\}\\ &+ \exp\left\{-\frac{1}{2}(x_1 \quad x_2 \quad y)\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0.4\\ 0 & 0.4 & 1\end{pmatrix}^{-1}\begin{pmatrix}x_1\\ x_2\\ y\end{pmatrix}\right\}\end{aligned} \tag{3.9}$$

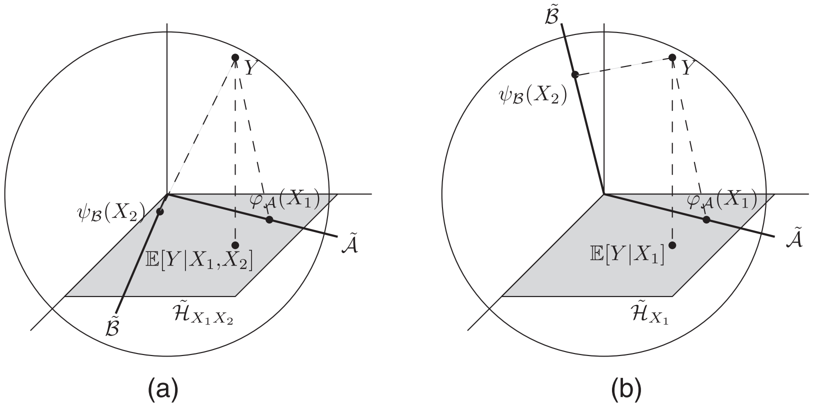
Y
$\psi_{\mathcal{B}}(X_2)$
$\varphi_{\mathcal{A}}(X_1)$
$\mathbb{E}[Y|X_1,X_2]$
$\tilde{\mathcal{A}}$
$\tilde{\mathcal{B}}$
$\tilde{\mathcal{H}}_{X_1X_2}$
(a)
$\tilde{\mathcal{B}}$
Y
$\psi_{\mathcal{B}}(X_2)$
$\varphi_{\mathcal{A}}(X_1)$
$\mathbb{E}[Y|X_1]$
$\tilde{\mathcal{A}}$
$\tilde{\mathcal{H}}_{X_1}$
(b)

FIG. 4. Geometric interpretation of multi- and single-domain estimation. We seek an estimate $\hat{Y}$ lying in the gray plane, which is as close as possible to Y. All we know is the norm of Y and its projections onto $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$. (a) Multi-domain prediction. Here, the best possible estimate is $\mathbb{E}[Y|X_1, X_2]$. (b) Single-domain prediction. Here the best possible estimate is $\mathbb{E}[Y|X_1]$.

is also consistent with all the constraints, making it a valid candidate as well. On the other hand, the density

$$f_{X_1X_2Y}(x_1, x_2, y) \propto \exp\left\{-\frac{1}{2}(x_1 \quad x_2 \quad y)\begin{pmatrix} 2 & 0 & 0.2 \\ 0 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix}^{-1}\begin{pmatrix} x_1 \\ x_2 \\ y \end{pmatrix}\right\} \tag{3.10}$$

satisfies all requirements except for the demand that it be consistent with the given marginal distribution $f_{X_1X_2}(x_1, x_2)$. Therefore, it is not feasible.

From a geometric standpoint, the lack of training samples drawn from $F_{X_1X_2Y}$ prevents us from being able to determine the position of the RV Y with respect to the subspace $\tilde{\mathcal{H}}_{X_1X_2}$ of all RVs that are functions of X_1 and X_2 or with respect to the subspace $\tilde{\mathcal{H}}_{X_1}$ of all RVs that are functions of X_1 alone. Our knowledge of the $\mathcal{A}$-optimal estimate of Y from X_1 and the $\mathcal{B}$-optimal estimate of Y from X_2 corresponds to knowledge of the projections $\Pi_{\tilde{\mathcal{A}}}Y$ and $\Pi_{\tilde{\mathcal{B}}}$ of Y onto the subspaces $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$, respectively. Additionally, knowledge of the joint distribution $F_{X_1X_2}$ can be interpreted as knowledge regarding how the spaces $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$ are situated with respect to one another. Finally, the fact that $\mathbb{E}[\|Y\|^2] = c$ corresponds to the knowledge that Y lies on a sphere of radius c. This set of constraints defines a set $\tilde{\mathcal{D}}$ of RVs to which Y must belong. A schematic illustration of this interpretation is depicted in Fig. 4.

3.3 *Goals*

The first problem we address in this paper is multi-domain regression. In this context, we would like to construct a predictor of Y from the two domains X_1 and X_2, where the only knowledge we have is that $F_{X_1X_2Y} \in \mathcal{D}$. The second problem we tackle is single-domain regression. Here, the goal is to construct an estimator of Y given X_1 alone based, again, only on the knowledge that $F_{X_1X_2Y} \in \mathcal{D}$. The special case of *shared-representation learning*, in which no labeled examples from the first domain are available (see Fig. 2(c, d)), corresponds to setting $\mathcal{A} = \{0\}$. The setting of *cross modality learning*, in which there is no access to training examples from the second domain (see Fig. 2(a, b)), can be addressed by setting

$\mathcal{B} = \{0\}$. The general case we treat here can account for a wide spectrum of possibilities, including these two extremes.

Any predictor of Y, whether a function of X_1 and X_2 or of X_1 alone, may perform well under certain distributions $F_{X_1X_2Y} \in \mathcal{D}$ and worse under others. Our goal is therefore to uniformly minimize the MSE over $\mathcal{D}$. As we will see, this minimax approach leads to simple closed form solutions, which can be easily applied to the various settings discussed in Section 2.

4. Multi-domain regression

Assume that the joint distribution of the triplet (X_1, X_2, Y) is known to belong to the family $\mathcal{D}$ of (3.7), where $\mathcal{A}$ and $\mathcal{B}$ are linear sets of prediction functions. For any distribution $F_{X_1X_2Y}$, the MSE attained by an estimator $\hat{Y} = \rho(X_1, X_2)$ is defined as

$$\mathrm{MSE}(F_{X_1X_2Y}, \rho) = \mathbb{E}[\|Y - \rho(X_1, X_2)\|^2], \tag{4.1}$$

where the expectation is with respect to $F_{X_1X_2Y}$. Since the MSE depends on $F_{X_1X_2Y}$, which is unknown, our approach is to seek the estimator whose worst-case MSE over $\mathcal{D}$ is minimal. This minimax concept is widely practiced in deterministic parameter estimation [6, 5] as well as in random parameter estimation [7, 8]. More concretely, we are interested in[3]

$$\rho_{\mathrm{M}} = \arg\min_{\rho} \sup_{F_{X_1X_2Y} \in \mathcal{D}} \mathrm{MSE}(F_{X_1X_2Y}, \rho). \tag{4.2}$$

Geometrically speaking, we saw that the knowledge we have does not suffice to determine the position of the RV Y with respect to the subspace $\tilde{\mathcal{H}}_{X_1X_2}$ of RVs that are functions of X_1 and X_2. Rather, it only suffices for determining a set $\tilde{\mathcal{D}}$ of RVs, to which Y must belong. Problem (4.2) can be interpreted as the search of an RV $\hat{Y}$ in $\tilde{\mathcal{H}}_{X_1X_2}$, whose distance to the farthest point in $\tilde{\mathcal{D}}$ is minimal. Note that the subspace $\tilde{\mathcal{H}}_{X_1X_2}$ clearly contains the subspaces $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$, onto which the projection of Y is known. However, $\tilde{\mathcal{H}}_{X_1X_2}$ is generally much larger than $\tilde{\mathcal{A}}$ or $\tilde{\mathcal{B}}$ or, even, larger than $\tilde{\mathcal{A}} + \tilde{\mathcal{B}}$ (not every function $\rho(X_1, X_2)$ can be written as $\varphi(X_1) + \psi(X_2)$ with $\varphi \in \mathcal{A}$ and $\psi \in \mathcal{B}$). Thus, the projection $\mathbb{E}[Y|X_1, X_2]$ of Y onto $\tilde{\mathcal{H}}_{X_1X_2}$, which is the best possible multi-domain predictor, cannot generally be determined from mere knowledge of the projections of Y onto $\tilde{\mathcal{A}}$ and onto $\tilde{\mathcal{B}}$.

The next theorem, whose proof can be found in Appendix A, provides a means for solving (4.2).

THEOREM 4.1 (Multi-domain minimax-MSE prediction) Choose any distribution $F_{X_1X_2Y} \in \mathcal{D}$ and consider the estimator[4]

$$\rho_{\mathcal{C}} = \arg\min_{\rho \in \mathcal{C}} \mathrm{MSE}(F_{X_1X_2Y}, \rho), \tag{4.3}$$

where $\mathcal{C} = \mathcal{A} + \mathcal{B}$, namely

$$\mathcal{C} = \{\rho : \rho(\boldsymbol{x}_1, \boldsymbol{x}_2) = \varphi(\boldsymbol{x}_1) + \psi(\boldsymbol{x}_2),\ \varphi \in \mathcal{A},\ \psi \in \mathcal{B}\}. \tag{4.4}$$

[3] The subscript ‘M’ stands for ‘multi-domain’.

[4] The minimum exists since the set of RVs $\tilde{\mathcal{C}}$ corresponding to the set of predictors $\mathcal{C}$ equals $\tilde{\mathcal{A}} + \tilde{\mathcal{B}}$ and is thus a closed subspace in $L^2(\Omega, \mathcal{F}, P)$.

Then

(1) the function $\rho_{\mathcal{C}}$ does not depend on the choice of $F_{X_1X_2Y} \in \mathcal{D}$;

(2) the value $\mathrm{MSE}(F_{X_1X_2Y}, \rho_{\mathcal{C}})$ does not depend on the choice of $F_{X_1X_2Y} \in \mathcal{D}$;

(3) the estimator $\rho_{\mathcal{C}}$ of (4.3) is also the solution ρ_{M} to (4.2).

Theorem 4.1 shows that instead of solving the minimax problem (4.2), we can equivalently solve the minimization problem (4.3). Namely, all we need to do is determine the MMSE estimator of Y among all functions of the form $\phi(X_1) + \psi(X_2)$ with $\phi \in \mathcal{A}$ and $\psi \in \mathcal{B}$. In other words, the minimax multi-domain estimate $\hat{Y}_{\mathrm{M}}$ is the projection of Y onto $\tilde{\mathcal{C}} = \tilde{\mathcal{A}} + \tilde{\mathcal{B}}$, which can be determined from the individual projections of Y onto $\tilde{\mathcal{A}}$ and onto $\tilde{\mathcal{B}}$. The importance of this observation follows from the fact that, as we show below, for many practical cases, $\rho_{\mathcal{C}}(X_1, X_2)$ possesses a simple closed-form solution.

Before demonstrating the utility of the minimax MSE approach, we note that optimizing the worst-case performance of an estimator is very conservative and may sometimes lead to over-pessimistic solutions. As an alternative, researchers in many application areas have proposed minimizing the worst-case *regret* [**6**, **7**, **15**, **16**]. The regret of an estimator $\rho(X_1, X_2)$ is defined as the difference between the MSE it achieves and the MSE of the MMSE solution, namely

$$\mathrm{REG}(F_{X_1X_2Y}, \rho) = \mathbb{E}[\|Y - \rho(X_1, X_2)\|^2] - \mathbb{E}[\|Y - \mathbb{E}[Y|X_1, X_2]\|^2]. \tag{4.5}$$

In this expression, both terms depend on $F_{X_1X_2Y}$, so that minimization of the worst-case regret is generally not equivalent to minimization of the worst-case MSE. Additional insight into the regret can be obtained from its equivalent characterization [**16**] as the MSE between $\rho(X_1, X_2)$ and $\mathbb{E}[Y|X_1, X_2]$, namely

$$\mathrm{REG}(F_{X_1X_2Y}, \rho) = \mathbb{E}[\|\rho(X_1, X_2) - \mathbb{E}[Y|X_1, X_2]\|^2]. \tag{4.6}$$

As we show in the following theorem, however, in the multi-domain prediction setting, the minimax-regret estimator coincides with the minimax-MSE solution. The proof of the theorem is provided in Appendix B.

THEOREM 4.2 (Multi-domain minimax-regret prediction) Consider the problem

$$\rho_{\mathrm{R}} = \arg\min_{\rho} \sup_{F_{X_1X_2Y} \in \mathcal{D}} \mathrm{REG}(F_{X_1X_2Y}, \rho), \tag{4.7}$$

where minimization is performed over all functions ρ of X_1 and X_2. Then its solution ρ_{R} coincides with ρ_{M} of (4.2).

We now apply Theorem 4.1 in several scenarios.

4.1 *Single-domain training*

Consider the situation of Fig. 1(a and b), where we have at our disposal only labeled examples from one domain, say X_1. In this case, $\mathcal{B} = \{0\}$ so that $\mathcal{C} = \mathcal{A}$. Consequently, the solution to (4.3) is simply

$$\rho_{\mathcal{C}}(X_1, X_2) = \varphi_{\mathcal{A}}(X_1). \tag{4.8}$$

This shows that when labeling unseen examples, there is no gain in basing the prediction on the domain X_2 for which we have no labeled training examples. Furthermore, at least from a worst-case perspective, there is no better strategy than using our initial predictor based on X_1 alone. More concretely, for any estimator that differs from $\varphi_{\mathcal{A}}(X_1)$ (and, in particular, one that is a function of X_2), there exist distributions $F_{X_1X_2Y} \in \mathcal{D}$ (one maybe being the true underlying distribution) under which the predictor $\varphi_{\mathcal{A}}(X_1)$ performs better.

This result does not stand in contrast to the basic observation in multi-view learning that unlabeled data help [2]. This is because in our setting, we do not assume that the two views are "coherent" or tend to agree in any sense, as done, for instance, in [10] in the context of multi-view regression.

4.2 *Multi-domain linear regression*

Suppose, as in Fig. 1(c), that we have a limited amount of labeled examples from both domains, which only suffice for identifying (with very high precision) the optimal linear predictor from each view. In this case, $\mathcal{A}$ and $\mathcal{B}$ correspond to the collection of all linear functions from $\mathbb{R}^{M_1}$ to $\mathbb{R}^N$ and from $\mathbb{R}^{M_2}$ to $\mathbb{R}^N$, respectively. Consequently, $\mathcal{C}$ is the set of all linear functions from $\mathbb{R}^{M_1} \times \mathbb{R}^{M_2}$ to $\mathbb{R}^N$. This implies that the solution to (4.3) is simply the best linear predictor of Y based on X_1 and X_2, namely

$$\rho_{\mathcal{C}}(X_1, X_2) = (\Gamma_{YX_1} \quad \Gamma_{YX_2}) \begin{pmatrix} \Gamma_{X_1X_1} & \Gamma_{X_1X_2} \\ \Gamma_{X_2X_1} & \Gamma_{X_2X_2} \end{pmatrix}^{\dagger} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}. \tag{4.9}$$

The second-order moments $\Gamma_{X_iX_j}$, $i, j \in \{1, 2\}$, can be estimated from the unlabeled training set. Similarly, the matrices Γ_{YX_j}, $j \in \{1, 2\}$, can be determined from the labeled sets.

The dependence of the multi-domain predictor $\rho_{\mathcal{C}}$ on the single-domain estimators $\phi_{\mathcal{A}}$ and $\psi_{\mathcal{B}}$ is not apparent at first sight. However, recall that the orthogonality principle states that $\mathbb{E}[(Y - \phi_{\mathcal{A}}(X_1))X_1^\top] = 0$ and $\mathbb{E}[(Y - \psi_{\mathcal{B}}(X_2))X_2^\top] = 0$. Therefore, the terms Γ_{YX_1} and Γ_{YX_2} in (4.9) can be replaced by $\mathbb{E}[\phi_{\mathcal{A}}(X_1)X_1^\top]$ and $\mathbb{E}[\psi_{\mathcal{B}}(X_2)X_2^\top]$, respectively. As these expectations are with respect to F_{X_1} and F_{X_2}, their computation can be carried out based only on the knowledge of $F_{X_1X_2}$, $\phi_{\mathcal{A}}$ and $\psi_{\mathcal{B}}$, which is available according to our problem formulation.

4.3 *Multi-domain parametric regression*

The above observation naturally extends to the case in which the training sets suffice for identifying the optimal parametric predictors of the forms

$$\varphi(X_1) = \sum_{k=1}^{K_1} a_k^1 \varphi_k(X_1), \qquad \psi(X_2) = \sum_{k=1}^{K_2} a_k^2 \psi_k(X_2), \tag{4.10}$$

where $\{\varphi_k\}_{k=1}^{K_1}$ and $\{\psi_k\}_{k=1}^{K_2}$ are given functions and $\{a_k^1\}_{k=1}^{K_1}$ and $\{a_k^2\}_{k=1}^{K_2}$ are arbitrary parameters. In this situation, $\mathcal{C}$ corresponds to the family of functions having the form

$$\rho(X_1, X_2) = \sum_{k=1}^{K_1} a_k^1 \varphi_k(X_1) + \sum_{k=1}^{K_2} a_k^2 \psi_k(X_2). \tag{4.11}$$

Thus, the optimal set of parameters $\boldsymbol{a} = (a_1^1 \ \cdots \ a_{K_1}^1 \ a_1^2 \ \cdots \ a_{K_2}^2)^\top$ is given by

$$\boldsymbol{a}^* = \begin{pmatrix} \Gamma_{\Phi\Phi} & \Gamma_{\Phi\Psi} \\ \Gamma_{\Psi\Phi} & \Gamma_{\Psi\Psi} \end{pmatrix}^\dagger \begin{pmatrix} \Gamma_{\Phi Y} \\ \Gamma_{\Psi Y} \end{pmatrix}, \tag{4.12}$$

with $\Gamma_{\Phi\Phi}$, $\Gamma_{\Psi\Psi}$, $\Gamma_{\Phi Y}$ and $\Gamma_{\Psi Y}$ being as in (3.6) and $\Gamma_{\Phi\Psi}$ being a $K_1 \times K_2$ matrix whose (i,j)th entry is $\mathbb{E}[\varphi_i(Y)^\top \psi_j(Z)]$. Similar to linear regression, the vectors $\Gamma_{\Phi Y}$ and $\Gamma_{\Psi Y}$ can be replaced, due to the orthogonality principle, by vectors whose jth entries are $\mathbb{E}[\varphi_j^\top(X_1)\varphi_{\mathcal{A}}(X_1)]$ and $\mathbb{E}[\psi_j^\top(X_1)\psi_{\mathcal{B}}(X_2)]$, respectively.

4.4 *Multi-domain partially linear regression*

Suppose, as in Fig. 1(d), that we have numerous labeled examples from the first domain, allowing us to determine $\mathbb{E}[Y|X_1]$, and only a limited amount of examples from the second domain, so that we can only determine the best linear predictor of Y from X_2. In this setting, Theorem 4.1 implies that the minimax-optimal predictor based on X_1 and X_2 is the estimator minimizing the MSE among all functions of the form

$$\rho(X_1, X_2) = \boldsymbol{a}(X_1) + \boldsymbol{B}X_2, \tag{4.13}$$

where $\boldsymbol{a} : \mathbb{R}^{M_1} \to \mathbb{R}^N$ is an arbitrary function and $\boldsymbol{B} \in \mathbb{R}^{N \times M_2}$ is some matrix. It was shown in [18] that the solution to this particular case is given by

$$\rho_{\mathrm{M}}(X_1, X_2) = \mathbb{E}[Y|X_1] + \Gamma_{YW}\Gamma_{WW}^\dagger W, \tag{4.14}$$

where $W = X_2 - \mathbb{E}[X_2|X_1]$.

The intuition here is that we need to make sure that we do not account for variations in Y twice when fusing information from X_1 and X_2. Thus, we start with the estimate $\varphi_{\mathcal{A}}(X_1) = \mathbb{E}[Y|X_1]$, and then update it with the linear MMSE estimate of Y based on the error $W = X_2 - \mathbb{E}[X_2|X_1]$ in predicting X_2 from X_1. A similar interpretation arises in the Kalman filter [11] (the measurement update stage) and in Kolmogorov's theory of prediction of wide-sense stationary sequences [23]. The main difference is that here we treat non-linear estimation, so that W is a non-linear function of X_1. Nevertheless, borrowing the terminology of recursive filtering and prediction, we refer here to the RV W as the *innovation* of domain X_2 with respect to the estimate $\varphi_{\mathcal{A}}(X_1)$ of Y from domain X_1.

In practice, the term $\mathbb{E}[Y|X_1]$ can be approximated from the labeled training examples of the first domain, e.g. using non-parametric methods. The second term in (4.14) can be obtained via a three-stage procedure. Specifically, we first employ a non-parametric technique to approximate $\xi(\boldsymbol{x}_1) = \mathbb{E}[X_2|X_1 = \boldsymbol{x}_1]$ from the unlabeled set. Next, we use the unlabeled samples to form the set $\{\xi(\boldsymbol{x}_1^u), \boldsymbol{x}_2^u\}_{u\in\mathcal{U}}$, from which we approximate the covariance matrix Γ_{WW} of $W = X_2 - \mathbb{E}[X_2|X_1]$. Lastly, we approximate Γ_{YX_2} from the labeled examples $\{\boldsymbol{x}_2^\ell, \boldsymbol{y}^\ell\}_{\ell\in\mathcal{L}_2}$ and $\Gamma_{Y\xi(X_1)}$ from the labeled examples $\{\xi(\boldsymbol{x}_1^\ell), \boldsymbol{y}^\ell\}_{\ell\in\mathcal{L}_1}$ in order to compute $\Gamma_{YW} = \Gamma_{YX_2} - \Gamma_{Y\xi(X_1)}$.

4.5 *Multi-domain semi-parametric regression*

Suppose, as above, that we know $\mathbb{E}[Y|X_1]$; however, we can also determine the best estimator of Y from X_2 among the parametric family

$$\psi(X_2) = \sum_{k=1}^{K} a_k \psi_k(X_2). \tag{4.15}$$

In this case, according to Theorem 4.1, the minimax-optimal estimator of Y based on X_1 and X_2 is the one minimizing the MSE among all functions of the form

$$\rho(X_1, X_2) = \boldsymbol{a}(X_1) + \sum_{k=1}^{K} a_k \psi_k(X_2). \tag{4.16}$$

The solution to this problem can be deduced by relying on the concept of $(\mathcal{A}, \mathcal{B})$-innovation, as we now define.

Definition 4.3 The $(\mathcal{A}, \mathcal{B})$-innovation of X_2 with respect to X_1 in predicting Y, which we denote by $\rho_{\mathcal{B}|\mathcal{A}}(X_2|X_1)$, is the MMSE estimator of Y among all functions of the form

$$\psi(X_2) - \eta_\psi(X_1), \tag{4.17}$$

with ψ being some function in $\mathcal{B}$ and $\eta_\psi(X_1)$ denoting the $\mathcal{A}$-optimal estimator of $\psi(X_2)$ from X_1.

Note that $\rho_{\mathcal{B}|\mathcal{A}}(X_2|X_1)$ is a function of both X_1 and X_2 (not to be confused with expressions such as $\mathbb{E}[X_2|X_1]$, which are only functions of X_1). Using this definition, we make the following observation regarding the structure of the minimax estimator, the proof of which is given in Appendix C.

Theorem 4.4 The solution to problem (4.3) can be expressed as

$$\rho_{\mathcal{C}}(X_1, X_2) = \varphi_{\mathcal{A}}(X_1) + \rho_{\mathcal{B}|\mathcal{A}}(X_2|X_1), \tag{4.18}$$

where $\rho_{\mathcal{B}|\mathcal{A}}(X_2|X_1)$ is the $(\mathcal{A}, \mathcal{B})$-innovation of X_2 with respect to X_1.

From a geometric perspective, the $(\mathcal{A}, \mathcal{B})$ innovation of X_2 with respect to X_1 is the projection of Y onto the subspace $\tilde{\mathcal{E}} = \Pi_{\tilde{\mathcal{A}}^\perp} \tilde{\mathcal{B}}$. Every RV in $\tilde{\mathcal{E}}$ is the projection of some RV in $\tilde{\mathcal{B}}$ onto $\tilde{\mathcal{A}}^\perp$. Thus, what Theorem 4.4 actually shows is that if a subspace $\tilde{\mathcal{C}}$ equals the sum of subspaces $\tilde{\mathcal{A}} + \tilde{\mathcal{B}}$, then the projection operator $\Pi_{\tilde{\mathcal{C}}}$ can be expressed as $\Pi_{\tilde{\mathcal{A}}} + \Pi_{\tilde{\mathcal{E}}}$ with $\tilde{\mathcal{E}}$ denoting the projection of $\tilde{\mathcal{B}}$ onto $\tilde{\mathcal{A}}^\perp$.

In our setting, $\mathcal{A}$ corresponds to the set of all functions from $\mathbb{R}^{M_1}$ to $\mathbb{R}^N$ so that $\varphi_{\mathcal{A}}(X_1) = \mathbb{E}[Y|X_1]$. Furthermore, $\mathcal{B}$ is the family of functions from $\mathbb{R}^{M_2}$ to $\mathbb{R}^N$ having the form (4.15). Therefore, for any $\psi \in \mathcal{B}$, the $\mathcal{A}$-optimal estimator of $\psi(X_2)$ based on X_1 is given by

$$\eta_\psi(X_1) = \mathbb{E}[\psi(X_2)|X_1] = \mathbb{E}\left[\sum_{k=1}^{K} a_k \psi_k(X_2) \middle| X_1\right] = \sum_{k=1}^{K} a_k \mathbb{E}[\psi_k(X_2)|X_1]. \tag{4.19}$$

Consequently, $\rho_{\mathcal{B}|\mathcal{A}}(X_2|X_1)$ in (4.18) is of the form

$$\psi(X_2) - \eta_\psi(X_1) = \sum_{k=1}^{K} a_k \psi_k(X_2) - \sum_{k=1}^{K} a_k \mathbb{E}[\psi_k(X_2)|X_1] = \sum_{k=1}^{K} a_k \rho_k(X_1, X_2), \tag{4.20}$$

where we defined $\rho_k(X_1, X_2) = \psi_k(X_2) - \mathbb{E}[\psi_k(X_2)|X_1]$. The optimal set of coefficients is given by

$$\boldsymbol{a}^* = \Gamma_{\rho\rho}^{\dagger} \Gamma_{\rho Y}, \tag{4.21}$$

where $\Gamma_{\rho\rho}$ and $\Gamma_{\rho Y}$ are as in (3.6) with $\varphi_i(X_1)$ replaced by $\rho_i(X_1, X_2)$.

To conclude, the optimal estimator of the form (4.16) is

$$\rho_{\mathrm{M}}(X_1, X_2) = \mathbb{E}[Y|X_1] + \sum_{k=1}^{K} a_k(\psi_k(X_2) - \mathbb{E}[\psi_k(X_2)|X_1]), \tag{4.22}$$

with coefficients $\{a_k\}$ given by (4.21). The first term in this expression can be approximated via non-parametric regression techniques from the labeled training examples of the first domain. The second term can be computed in two stages. First, each of the functions $\{\psi_k(X_2)\}_{k=1}^{K}$ is regressed on X_1 using the unlabeled dataset, to obtain an approximation of $\mathbb{E}[\psi_k(X_2)|X_1]$. Then, Y is linearly regressed against $\{\psi_k(X_2) - \mathbb{E}[\psi_k(X_2)|X_1]\}_{k=1}^{K}$, using the two labeled sets, as discussed in Section 4.4.

5. Single-domain prediction with multi-domain training

Next, we address the setting in which, at the testing stage, our predictor is only supplied with one type of features, say X_1. The interesting question in this context is how to take into account the training sets of both domains in order to design an improved estimator of Y based on X_1 alone.

Since our estimator operates on X_1 and is judged by the proximity of its output to Y, its performance is only affected by the joint distribution of Y and X_1. It may thus seem at first that the second set of features X_2 cannot be of help in improving the estimation accuracy. However, note that F_{X_1Y} is not fully known in our setting. Thus, being told the statistical relations between Y and X_2 and between X_1 and X_2 might help to narrow down the set of candidate distributions F_{X_1Y} for which we need to design an estimator.

The statistical relations known to us are the same as in Section 4. Namely, we know that $F_{X_1X_2Y}$ belongs to the class $\mathcal{D}$ of (3.7). Therefore, as in Section 4, our goal is to optimize the worst case performance of our estimator over $\mathcal{D}$. As it turns out, in contrast with the multi-domain problem, in the single-domain setting the minimax MSE and minimax regret solutions no longer coincide. A simple example demonstrating this phenomenon is discussed in Appendix D. Here, we focus on minimizing the worst-case regret. As will be clear from the proof provided in Appendix B, determining the minimax-MSE estimator in the single-domain setting is generally much harder than minimizing the worst-case regret. The former remains an open problem.

In single-domain regression, whatever we do, our estimator will not achieve a lower MSE than the conditional expectation $\mathbb{E}[Y|X_1]$. Therefore, the *regret* of interest is now

$$\mathrm{REG}(F_{X_1X_2Y}, \rho) = \mathbb{E}[\|Y - \rho(X_1)\|^2] - \mathbb{E}[\|Y - \mathbb{E}[Y|X_1]\|^2]. \tag{5.1}$$

As in the multi-domain setting, this regret can be written as [16]

$$\mathrm{REG}(F_{X_1X_2Y}, \rho) = \mathbb{E}[\|\rho(X_1) - \mathbb{E}[Y|X_1]\|^2]. \tag{5.2}$$

Our goal is to determine the minimax-regret estimator[5]

$$\rho_{\mathrm{S}} = \arg\min_{\rho} \sup_{F_{X_1X_2Y} \in \mathcal{D}} \mathrm{REG}(F_{X_1X_2Y}, \rho), \tag{5.3}$$

where now minimization is performed only over functions ρ of X_1.

[5] The subscript ‘S’ stands for ‘single-domain’.

TABLE 1 *Single-domain prediction scenarios*

	Training		
	Unlabeled	Labeled	Testing
Cross-domain regression	$X_1 + X_2$	X_1	X_1
Shared representation regression	$X_1 + X_2$	X_2	X_1
Regression with side information	$X_1 + X_2$	X_1, X_2	X_1

The next theorem, whose proof may be found in Appendix B, describes the single-domain minimax-regret estimator in terms of the multi-domain minimax-MSE solution.

THEOREM 5.1 (Single-domain minimax-regret prediction) The solution to problem (5.3) is given by

$$\rho_{\mathrm{S}}(X_1) = \mathbb{E}[\rho_{\mathrm{M}}(X_1, X_2)|X_1], \tag{5.4}$$

where $\rho_{\mathrm{M}}(X_1, X_2)$ is the multi-domain minimax estimator (4.2).

This result has a very simple and intuitive explanation. We know that $F_{X_1X_2Y}$ belongs to the set $\mathcal{D}$, and therefore $\rho_{\mathrm{M}}(X_1, X_2)$ is the optimal estimate of Y in a minimax-MSE sense. However, we cannot use this estimate as it is a function of X_2, which is not measured in our setting. What Theorem 5.1 shows is that the optimal strategy is to estimate $\rho_{\mathrm{M}}(X_1, X_2)$ based on the available measurements, which are X_1 alone. Computation of the conditional expectation $\mathbb{E}[\rho_{\mathrm{M}}(X_1, X_2)|X_1]$ only requires knowledge of the marginal distribution $F_{X_1X_2}$, which is available in our setting.

In geometric terms, the single-domain minimax-regret estimate $\hat{Y}_{\mathrm{S}}$ is simply the projection $\Pi_{\tilde{\mathcal{H}}_{X_1}} \hat{Y}_{\mathrm{M}}$ of the multi-domain minimax estimate $\hat{Y}_{\mathrm{M}}$ onto the subspace $\tilde{\mathcal{H}}_{X_1}$ of RVs that are functions of X_1.

We now apply this result to three interesting special cases, as shown in Table 1.

5.1 *Cross-domain regression*

In cross-modality learning (coined in [**20**]), we only have labeled examples from domain X_1 and not from X_2, as illustrated in Figs 2(a, b) and in the first row of Table 1. The basic intuition here, as presented in [**20**], is that the unlabeled data may be used to boost the performance of the best single-domain estimator $\varphi_{\mathcal{A}}(X_1)$ that can be designed based solely on labeled examples from the domain X_1. This is supposedly possible by learning better single-modality feature representations given the unlabeled data from the multiple modalities, thus the term cross-modality.

This setting can be treated within our framework by setting $\psi_{\mathcal{B}}(X_2) = 0$. As we have seen in Section 4.1, in this situation $\rho_{\mathrm{M}}(X_1, X_2) = \varphi_{\mathcal{A}}(X_1)$. Therefore, the single-domain minimax-regret predictor of Y from X_1 is given by

$$\rho_{\mathrm{S}}(X_1) = \mathbb{E}[\varphi_{\mathcal{A}}(X_1)|X_1] = \varphi_{\mathcal{A}}(X_1). \tag{5.5}$$

We see that despite the fact that we know $F_{X_1X_2}$, there is no better strategy than using the estimator $\varphi_{\mathcal{A}}(X_1)$ here. This implies that cross-modality learning is not useful unless additional knowledge on the underlying distributions is available.

The authors of [20] used cross-modality learning to classify isolated words from either audio or video (lipreading). It was reported that unlabeled audio-visual examples helped improve visual recognition but failed to boost the performance of an audio classifier. This empirical result aligns with our theoretical analysis, which states that, in the worst-case scenario, there is nothing better to do than disregarding the modality for which no labeled examples are available.

5.2 *Shared-representation regression*

In shared-representation learning (coined in [20]), also referred to as estimation with partial knowledge [16], we have no labeled examples from domain X_1 but rather only from X_2. This is illustrated in Figs 2(c, d) and in the second row of Table 1. The mechanism used in [20] to construct a predictor (classifier) was based on learning multi-modal features from the unlabeled data, thus the term shared representation. More specifically, the deep architectures proposed in [20] were designed to learn to reconstruct both modalities (audio and video, in the case of [20]) given only one modality as the input.

Since we can learn a predictor $\psi_{\mathcal{B}}(X_2)$ from the second domain, and only measure an instance X_1 from the first domain, a naive approach would be to feed the predictor $\psi_{\mathcal{B}}$ with an estimate of X_2, which is based on X_1, rather than with X_2 itself. For example, we can use the predictor $\psi_{\mathcal{B}}(\mathbb{E}[X_2|X_1])$, where the MMSE estimate $\mathbb{E}[X_2|X_1]$ is approximated by non-parametric methods from the unlabeled training set. However, as we now show, this strategy is generally *not* minimax-optimal.

Recall from Section 4.1 that the multi-domain predictor corresponding to the setting in which $\mathcal{A} = \{0\}$ is $\rho_{\mathrm{M}}(X_1, X_2) = \psi_{\mathcal{B}}(X_2)$. Therefore, the single-domain minimax-regret predictor of Y from X_1 is given by

$$\rho_{\mathrm{S}}(X_1) = \mathbb{E}[\psi_{\mathcal{B}}(X_2)|X_1] \tag{5.6}$$

in this case. This solution generalizes the estimator of [16, Theorem 8], which was developed for the case in which $\mathcal{B}$ is the set of all functions. In the latter scenario $\psi_{\mathcal{B}}(X_2) = \mathbb{E}[Y|X_2]$ so that $\rho_{\mathrm{S}}(X_1) = \mathbb{E}[\mathbb{E}[Y|X_2]|X_1]$, and the two methods coincide.

Geometrically, in the shared-representation setting, we seek an RV in $\tilde{\mathcal{H}}_{X_1}$ that best approximates Y while all we know is the projection $\Pi_{\tilde{\mathcal{B}}} Y$ of Y onto $\tilde{\mathcal{B}}$. The minimax strategy dictates that we should project $\Pi_{\tilde{\mathcal{B}}} Y$ onto $\tilde{\mathcal{H}}_{X_1}$, so that the minimax single-domain predictor is $\hat{Y}_{\mathrm{S}} = \Pi_{\tilde{\mathcal{H}}_{X_1}} \Pi_{\tilde{\mathcal{B}}} Y$. This interpretation is illustrated in Fig. 5.

As an example, consider the setting in which we have a limited number of labeled examples from domain X_2, which only allows one to determine the best linear predictor of Y from X_2. In this case, $\psi_{\mathcal{B}}(X_2) = \Gamma_{YX_2} \Gamma^{\dagger}_{X_2X_2} X_2$, implying that $\rho_{\mathrm{S}}(X_1) = \mathbb{E}[\Gamma_{YX_2} \Gamma^{\dagger}_{X_2X_2} X_2 | X_1] = \Gamma_{YX_2} \Gamma^{\dagger}_{X_2X_2} \mathbb{E}[X_2|X_1]$. Namely, minimax-regret estimation does boil down, in this setting, to the naive strategy of applying $\psi_{\mathcal{B}}$ on $\mathbb{E}[X_2|X_1]$. This, however, is not always the case. Suppose, for instance, that we have numerous examples from domain X_2, so that $\mathcal{B}$ is the set of all functions from $\mathbb{R}^{M_2}$ to $\mathbb{R}^N$. In this situation, $\psi_{\mathcal{B}}(X_2) = \mathbb{E}[Y|X_2]$, so that $\rho_{\mathrm{S}}(X_1) = \mathbb{E}[\mathbb{E}[Y|X_2]|X_1]$. This solution does not generally coincide with the naive estimator $\mathbb{E}[Y|\mathbb{E}[X_2|X_1]]$.

The estimator (5.6) can be approximated from the available training data by first determining the function $\psi_{\mathcal{B}}(\boldsymbol{x}_2)$ from the labeled set of the second domain and then using non-parametric regression on the set $\{\boldsymbol{x}_1^u, \psi_{\mathcal{B}}(\boldsymbol{x}_2^u)\}_{u \in \mathcal{U}}$.

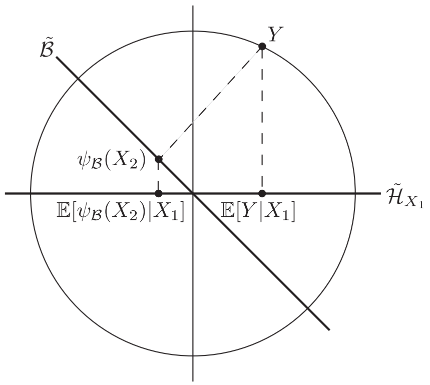
$\tilde{\mathcal{B}}$
Y
$\psi_{\mathcal{B}}(X_2)$
$\mathbb{E}[\psi_{\mathcal{B}}(X_2)|X_1]$
$\mathbb{E}[Y|X_1]$
$\tilde{\mathcal{H}}_{X_1}$

FIG. 5. Geometric interpretation of shared representation estimation. Among all RVs in $\tilde{\mathcal{H}}_{X_1}$, the best possible estimate is $\mathbb{E}[Y|X_1]$, the projection of Y onto $\tilde{\mathcal{H}}_{X_1}$. We only know, however, $\Psi_{\mathcal{B}}(X_2)$, the projection of Y onto $\tilde{\mathcal{B}}$. The minimax-regret estimator is the projection of $\Psi_{\mathcal{B}}(X_2)$ onto $\tilde{\mathcal{H}}_{X_1}$.

5.3 *Regression with side information*

The general setting in which we have training data from both domains can be treated by employing Theorem 4.4. Specifically, when $\mathcal{A}$ and $\mathcal{B}$ are two arbitrary spaces of prediction functions, $\rho_{\mathrm{M}}(X_1, X_2)$ is given by (4.18), and therefore

$$\rho_{\mathrm{S}}(X_1) = \varphi_{\mathcal{A}}(X_1) + \mathbb{E}[\rho_{\mathcal{B}|\mathcal{A}}(X_2|X_1)|X_1], \tag{5.7}$$

where $\rho_{\mathcal{B},\mathcal{A}}(X_2|X_1)$ is the $(\mathcal{A}, \mathcal{B})$ innovation of X_2 with respect to X_1. This representation highlights the fact that the second labeled set and the unlabeled set come into play in the term $\mathbb{E}[\rho_{\mathcal{B},\mathcal{A}}(X_2|X_1)|X_1]$.

To understand when training data from an unobserved domain cannot help, we recall from Definition 4.3 that $\rho_{\mathcal{B},\mathcal{A}}(X_2|X_1)$ is of the form $\psi(X_2) - \eta_{\psi}(X_1)$, with $\psi \in \mathcal{B}$ and $\eta_{\psi}(X_1)$ being the $\mathcal{A}$-optimal estimate of $\psi(X_2)$ from X_1. Therefore, the second term in (5.7) vanishes if, for example,

$$\mathbb{E}[\psi(X_2)|X_1] = \eta_{\psi}(X_1) \tag{5.8}$$

for every $\psi \in \mathcal{B}$. Intuitively, this can happen if the class $\mathcal{A}$ of functions is very rich and/or the class $\mathcal{B}$ is not. As an example, if $\mathcal{A}$ is the set of all functions from $\mathbb{R}^{M_1}$ to $\mathbb{R}^N$, then $\eta_{\psi}(X_1) = \mathbb{E}[\psi(X_2)|X_1]$, so that (5.8) is satisfied, indicating that the training set from the second domain is not needed. Indeed, in this situation $\varphi_{\mathcal{A}}(X_1) = \mathbb{E}[Y|X_1]$, meaning that we can already determine the MMSE predictor of Y from X_1 using the first training set so that no potential improvement can be obtained using the second set.

As a more interesting example, suppose that the RVs X_1 and X_2 are jointly Gaussian, that $\mathcal{B}$ is the set of all linear functions from $\mathbb{R}^{M_2}$ to $\mathbb{R}^N$, and that $\mathcal{A}$ contains the set of all linear functions from $\mathbb{R}^{M_1}$ to $\mathbb{R}^N$. In this case, every $\psi \in \mathcal{B}$ corresponds to some matrix $\boldsymbol{A}$ such that $\psi(X_2) = \boldsymbol{A}X_2$. Consequently, using the fact that the MMSE estimate is linear in the Gaussian setting,

$$\mathbb{E}[\psi(X_2)|X_1] = \mathbb{E}[\boldsymbol{A}X_2|X_1] = \boldsymbol{A}\mathbb{E}[X_2|X_1] = \boldsymbol{A}\Gamma_{X_2X_1}\Gamma^{\dagger}_{X_1X_1}X_1. \tag{5.9}$$

Moreover, X_1 and $\psi(X_2)$ are jointly Gaussian, implying that

$$\eta_{\psi}(X_1) = \Gamma_{\psi(X_2)X_1}\Gamma^{\dagger}_{X_1X_1}X_1 = \boldsymbol{A}\Gamma_{X_2X_1}\Gamma^{\dagger}_{X_1X_1}X_1. \tag{5.10}$$

Thus, (5.9) and (5.10) coincide and (5.8) is satisfied, indicating that the second training set is not required here as well.

Another interesting viewpoint can be obtained by switching the roles of X_1 and X_2 in the representation (4.18) of $\rho_{\mathrm{M}}(X_1, X_2)$. This leads to the expression

$$\rho_{\mathrm{S}}(X_1) = \mathbb{E}[\psi_{\mathcal{B}}(X_2)|X_1] + \mathbb{E}[\rho_{\mathcal{A}|\mathcal{B}}(X_1, X_2)|X_1]. \tag{5.11}$$

Here, we recognize the first term as being the shared-representation estimator (5.6) of Y from X_1, which does not use labeled examples from the domain X_1. Therefore, we see that the training set from the first (observed) domain is not needed if the second term in (5.11) vanishes. Using the fact that $\rho_{\mathcal{A}|\mathcal{B}}(X_1|X_2) = \varphi(X_1) - \eta_\varphi(X_2)$ with $\varphi \in \mathcal{A}$ and $\eta_\varphi(X_2)$ being the $\mathcal{B}$-optimal estimate of $\varphi(X_1)$ from X_2, we conclude that this happens if, for example,

$$\varphi(X_1) = \mathbb{E}[\eta_\varphi(X_2)|X_1] \tag{5.12}$$

for every $\varphi \in \mathcal{A}$. As a concrete example, consider again the setting in which the RVs X_1 and X_2 are jointly Gaussian and $\mathcal{A}$ and $\mathcal{B}$ are classes of linear functions. In this situation, $\varphi(X_1) = \boldsymbol{A}X_1$ for some matrix $\boldsymbol{A}$, so that $\eta_\varphi(X_2) = \Gamma_{\varphi(X_1)X_2}\Gamma^\dagger_{X_2X_2}X_2 = \boldsymbol{A}\Gamma_{X_1X_2}\Gamma^\dagger_{X_2X_2}X_2$ and, consequently,

$$\mathbb{E}[\eta_\varphi(X_2)|X_1] = \boldsymbol{A}\Gamma_{X_1X_2}\Gamma^\dagger_{X_2X_2}\mathbb{E}[X_2|X_1] = \boldsymbol{A}\Gamma_{X_1X_2}\Gamma^\dagger_{X_2X_2}\Gamma_{X_2X_1}\Gamma^\dagger_{X_1X_1}X_1. \tag{5.13}$$

Therefore, (5.12) is satisfied if $\Gamma_{X_1X_2}\Gamma^\dagger_{X_2X_2}\Gamma_{X_2X_1}\Gamma^\dagger_{X_1X_1} = \boldsymbol{I}$, or, equivalently if $\Gamma_{X_1X_1} - \Gamma_{X_1X_2}\Gamma^\dagger_{X_2X_2}\Gamma_{X_2X_1} = 0$. The latter expression is none other than the error covariance of the MMSE estimate of X_1 from X_2. Therefore, condition (5.12) is satisfied in this setting if X_1 can be estimated from X_2 with no error. Indeed, in this scenario, we do not need to observe training examples from the domain X_1, as these can be synthetically generated from the examples of the second domain.

To approximate the resulting estimators from sets of points, it is often more convenient to use the form (5.11) rather than (5.7). As a concrete example, consider linear regression with non-linear side information, namely where $\mathcal{A}$ is the set of all linear functions and $\mathcal{B}$ is the family of all (not necessarily linear) functions. Then, from Theorem 5.1 and (4.14), we conclude that

$$\rho_{\mathrm{S}}(X_1) = \mathbb{E}[\mathbb{E}[Y|X_2]|X_1] + \Gamma_{YW}\Gamma^\dagger_{WW}(X_1 - \mathbb{E}[\mathbb{E}[X_1|X_2]|X_1]), \tag{5.14}$$

where here $W = X_1 - \mathbb{E}[X_1|X_2]$. The terms $\mathbb{E}[\mathbb{E}[Y|X_2]|X_1]$ and $\mathbb{E}[\mathbb{E}[X_1|X_2]|X_1]$ can be approximated using non-parametric methods, similar to the discussion in Section 5.2, and the covariance matrices Γ_{YW} and Γ_{WW} can be approximated as in Section 4.4.

6. Experimental results

We now demonstrate our regression approach, that derives from the theoretical results just presented, in a toy example and two illustrative applications.

6.1 *Toy example*

To gain intuition into the difficulties associated with learning from distinct training sets, we begin with a toy example incorporating scalar RVs. For the simplicity of visualization, we concentrate on the shared-representation learning setting described in Section 5.2 (Fig. 2(c, d), second row of Table 1), which, on one hand, is simple, and, on the other hand, has a non-trivial solution.

Figure 6(a, b) depict two joint distributions of X_1, X_2 and Y, as point clouds. Recall that, in the shared-representation setting, the datasets available to us correspond to independent draws from $F_{X_1X_2}$ (unlabeled multi-domain examples) and from F_{X_2Y} (labeled examples from the second domain). These datasets are plotted on the $X_1 - X_2$ plane and on the $X_2 - Y$ plane, respectively. From the labeled examples, we can determine an estimator $\psi_{\mathcal{B}}(X_2)$ of Y based on X_2. This estimator is plotted on the $X_2 - Y$ plane. Here, we used a large number of training examples, which allowed us to approximate $\psi_{\mathcal{B}}(X_2) = \mathbb{E}[Y|X_2]$ using the Nadaraya–Watson non-parametric regression method (see (3.1)). Based on knowledge of the estimator $\psi_{\mathcal{B}}(X_2)$ and on the distribution $F_{X_1X_2}$, our method constructs an estimator $\rho_{\mathrm{S}}(X_1)$ of Y based on X_1.

Figure 6(c, d) depict the marginal distribution F_{X_1Y} as well as the MMSE estimate $\mathbb{E}[Y|X_1]$ and our minimax-regret solution, $\rho_{\mathrm{S}}(X_1) = \mathbb{E}[\mathbb{E}[Y|X_2]|X_1]$, corresponding to the settings of Fig. 6(a, b), respectively. As apparent from these plots, although the distributions of Fig. 6(a, b) are different, the marginal F_{X_1Y} and, consequently, also the MMSE estimator $\mathbb{E}[Y|X_1]$ are the same in both situations. Nevertheless, the solution $\rho_{\mathrm{S}}(X_1)$, which does not have access to draws from F_{X_1Y} is completely different in the two scenarios. Indeed, in Fig. 6(c), $\rho_{\mathrm{S}}(X_1)$ is close to the MMSE solution while in Fig. 6(d) it is not. This is rooted in the fact that F_{X_2Y} in the latter case is such that it is hard to estimate Y from X_2 (the performance of $\psi_{\mathcal{B}}(X_2) = \mathbb{E}[Y|X_2]$ is poor). Therefore, as the shared-representation estimator relies only on the knowledge of $\psi_{\mathcal{B}}(X_2)$ and of the statistical relation between X_1 and X_2, it performs worse in the situation of Fig. 6(b).

6.2 *Face normalization*

Many facial recognition methods rely on a preprocessing stage, coined *normalization*, which is aimed at removing variations that were not observed in the training database. These may include variations due to illumination, pose, facial expressions and more. To demonstrate the utility of our approach, we now focus on the problem of producing a neutral expression face from a smiling one.

A straightforward way of tackling this problem is to learn a regression function from pairs of training images. This requires a database in which each subject appears at least twice, one time with a neutral expression and one time with a smile. Unfortunately, large datasets of this sort are hard to collect. In many practical situations one only has access to a database in which each subject appears only once. While different subjects may be wearing different expressions, direct inference of the statistical relation between a smiling and a neutral expression face is virtually impossible in such scenarios. To bypass this obstacle, we can use a second domain, or view, for which it is easy to obtain examples that are paired with the images in the database. This can be done, for example, by manually marking a set of points in several predefined locations on all images in the database. Thus, denoting by (X_1, X_2, Y) a triplet of a smiling face, its point annotations and the corresponding neutral expression image, we may construct an unlabeled set of annotated smiling faces $\{\boldsymbol{x}_1^u, \boldsymbol{x}_2^u\}$ and a set of annotated neutral expression faces $\{\boldsymbol{x}_2^\ell, \boldsymbol{y}^\ell\}$. This allows employing our shared-representation regression technique for designing a predictor of Y based on X_1. If, in addition, several subjects were photographed more than once, then we may construct a third set $\{\boldsymbol{x}_1^\ell, \boldsymbol{y}^\ell\}$, containing pairs of images of smiling and neutral expression faces. In this case, we can apply regression with side information, as discussed in Section 5.3.

Figure 7 depicts several manually annotated neutral and smiling facial images taken from the AR database [13]. The point annotations were taken from http://www-prima.inrialpes.fr/FGnet/data/05-ARFace/tarfd_markup.html. The images were scaled, rotated and cropped into an ellipsoidal template such that the eyes appear at predefined locations. In practice, this can be performed automatically [24, 14]. To apply our methods, we normalized the images to be of zero mean and unity norm

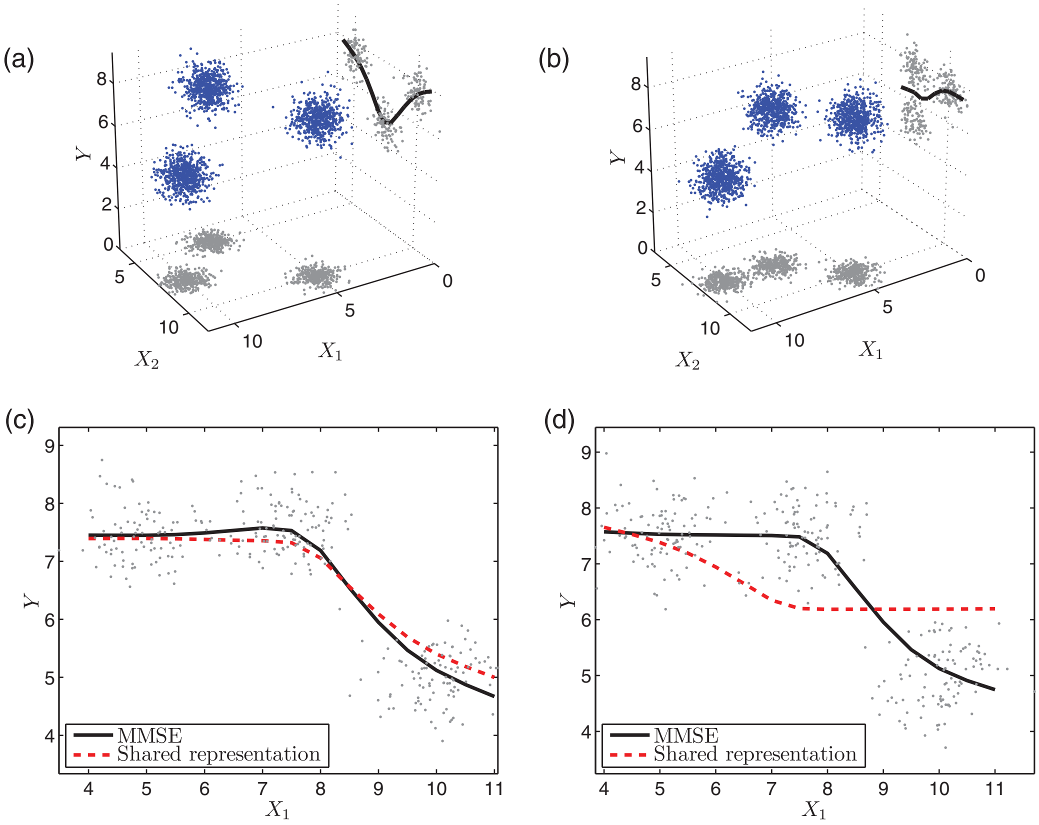

FIG. 6. Visualization of shared-representation regression. (a and b) Two examples of distributions $F_{X_1X_2Y}$ (dark point clouds) together with their corresponding marginal distributions $F_{X_1X_2}$ and F_{X_2Y} (gray point clouds) and the corresponding estimator $\mathbb{E}[Y|X_2]$ (curve on the $X_2 - Y$ plane). (c and d) The marginal distribution F_{X_1Y} (point cloud), the MMSE estimate $\mathbb{E}[Y|X_1]$ (solid line) and the minimax-regret solution $\rho_S(X_1)$ (dashed line) corresponding to the settings of Fig. 6(a, b), respectively.

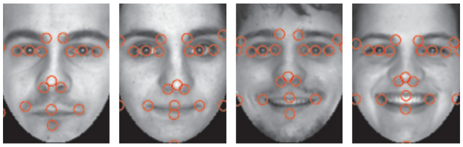

FIG. 7. Annotated images from the AR database.

and reduced them to 86 dimensions using principal component analysis (PCA). The non-linear regression scheme we used as a building block in our methods was first-order polynomial regression with a Gaussian kernel. The bandwidth of the kernel was adaptively tuned to be a constant times the average squared distance between the query (test) and the training data points.[6]

[6] More specifically, for a labeled training set $\{\boldsymbol{x}^\ell, \boldsymbol{y}^\ell\}_{\ell=1}^L$ and a query (test) point $\boldsymbol{x}$, we chose $h = 0.2 \times \frac{1}{L}\sum_{\ell=1}^L \|\boldsymbol{x} - \boldsymbol{x}^\ell\|^2$.

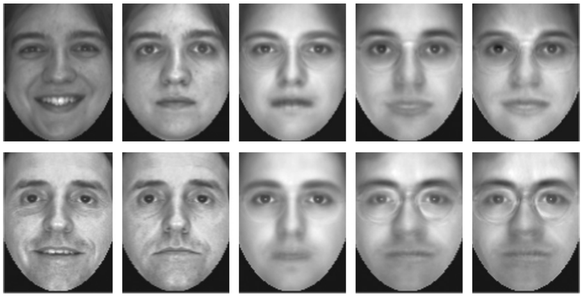

FIG. 8. Neutral expression synthesis from smiling images. From left to right: query, ground truth, direct non-linear regression, shared-representation non-linear regression (Section 5.2), linear regression with non-linear side information (Section 5.3).

TABLE 2 *Performance of neutral expression synthesis methods*

Setting	RMSE
Direct non-linear regression with 118 examples	0.187
Direct linear regression with 40 examples	0.200
Shared-representation non-linear regression	0.263
Linear regression with non-linear side information	0.247

Figure 8 demonstrates the results obtained with our approach in several settings. These results correspond to a leave-one-out experiment. Namely, each time we use one subject for testing and the rest for training. The two leftmost columns correspond to the query smiling face and the corresponding desired (unobserved) neutral expression image. The third column shows the result of directly performing regression using 118 pairs of smile/neutral images. The fourth column is the result of performing shared representation regression via (5.6), using a training set of 38 annotated smiling faces and a set of 40 annotated neutral images (of different subjects). The rightmost column uses, in addition to these two sets, a training set comprising 40 pairs of images of neutral and smiling expressions to perform linear regression with non-linear side information (equation (5.14)).

Table 2 shows the root MSE (RMSE), $(\mathbb{E}[\|Y - \hat{Y}\|^2])^{\frac{1}{2}}$, attained in each of the settings. The results are averaged over all subjects. As expected, using direct training with 118 examples yields the best results (lowest RMSE). Interestingly, in this situation, using direct linear regression with 40 examples results in only slightly inferior results. It can be seen that employing two sets with roughly 40 examples each, instead of direct training, leads to an increase in the RMSE of 41%. This gap is reduced to 32% with the aid of an additional set of 40 direct training pairs.

A somewhat counterintuitive phenomenon is that the method of linear regression with non-linear side information, which employs 40 labeled smiling faces, performs worse than direct linear regression, which is based on the same 40 labeled examples in addition to 40 examples of annotated neutral faces. This suggests that the addition of information (labeled examples from domain X_2 in this case) may deteriorate the performance of our estimators. Indeed, this results from the fact that our methods are only optimal in a worst-case regret sense. Thus, it is only the worst-case regret that is assured to improve when adding information, not the MSE in a specific situation.

Perceptually, the images produced by the indirect methods do not seem to be much worse than those obtained with direct training. Note that the spatial smoothing apparent in all methods is due to the fact that all regression methods boil down at the end to some sort of averaging of many images from the training set. It is also important to note that the vague traces of glasses in the last two columns are no coincidence. Specifically, when there are no (or very few) joint examples of smile/neutral faces, no method can ever be able to determine whether the person wears glasses or not. This is because we only know how the smiling images (pixel values) relate to the geometry (point annotations) and how the geometry relates to the neutral images. Now, for every possible geometry, roughly half the people in the neutral database wear glasses and half do not.

6.3 *Audio-visual word recognition*

Although the entire discussion in this paper has focused on regression, we believe that similar methods can be developed for classification tasks. Such developments would surely require an entirely different arsenal of mathematical techniques. However, to support our claim, we now illustrate that shared representation classification can even be achieved by using the naive approach of performing regression and then quantizing the output in order to obtain a classification rule.

Specifically, we now consider the tasks of spoken digit classification from audio-only and video-only measurements. These are 10-class classification problems, with the classes corresponding to the digits $0, \ldots, 9$. To study this setting, we used the Grid Corpus [4], which consists of speakers saying simple structured sentences. Every sentence contains one digit, which we isolated using the supplied transcriptions. For every digit, there are 100 examples per speaker. We constructed three distinct training sets: one of labeled audio examples (4 males, 4 females), one of visual examples (4 males, 4 females) and one of unlabeled audio-visual examples (6 males, 4 females). Six speakers were used for testing (3 males, 3 females). After manual removal of examples in which our automatic lip detection failed (see description below), we ended up with 7793 labeled audio examples (with between 772 and 791 examples per digit), 7585 labeled video examples (with between 746 and 765 examples per digit), 9897 unlabeled audio-visual examples (with between 983 and 991 examples per digit) and 5611 test examples (with between 540 and 575 examples per digit).

To process the video, we converted the images to gray scale, used the face-detection method of [12] and then applied several mean-shift iterations on the gradient image map in order to extract the lip region in the first image of each frame-bunch. Segments of duration 320 ms were used for recognition. This corresponded to eight consecutive video frames (at a rate of 25 frames per second) and 1600 audio samples (at a sampling rate of 5 KHz). The image frames were reduced to 10 dimensions using PCA, resulting in an 80-dimensional video feature vector. The processing of the audio was performed by computing spectrograms with windows of duration 10 ms and an overlap of 2.5 ms. The dimension of the spectrogram was reduced using PCA to 180 to constitute the audio features. In all experiments Y was a 10-dimensional vector with 1 at the location corresponding to the spoken digit and 0 elsewhere. Figure 9 visualizes the basic audio-visual preprocessing.

As mentioned above, our approach is designed for regression, so that the predicted $\hat{Y}$ is a continuous variable. To perform classification, we chose the maximal element in $\hat{Y}$. For simplicity, $\mathcal{A}$ and $\mathcal{B}$ were taken as the sets of all linear functions (linear regression). This choice yields rather poor classification results based solely on audio or solely on video. Our goal, though, is to demonstrate that even with such naive single-domain predictors, we can attain good recognition accuracy by using our approach, which cleverly fuses the two domains.

Table 3 shows the accuracy of our approach and, for reference, also presents the results obtained with the deep restricted Boltzmann machine (RBM) of [20] on the Clemson University audio visual experiments (CUAVE) dataset [22]. The Grid corpus used here is more challenging in that the digits

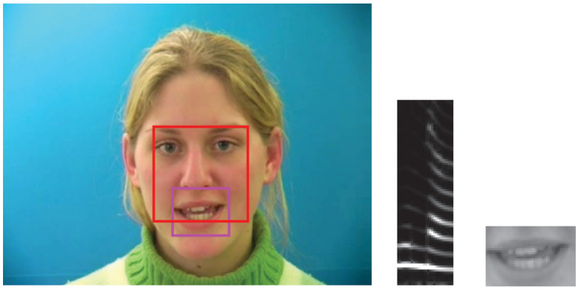

FIG. 9. Processing of the video and audio of a speaker saying the word ‘nine’. From left to right: lip detection, spectrogram and extracted lip region.

TABLE 3 *Audio-visual digit classification performance*

Features		Accuracy	
		Minimax, %	Deep RBM, %
Training	Testing	(Grid corpus)	(CUAVE)
Audio	Audio	69.3	95.8
Video	Video	52.0	69.7
Video	Audio	50.1	27.5
Audio	Video	44.6	29.4

appear within sentences, rather than individually. We note that Table 3 *does not* serve to draw conclusions regarding the benefit, in terms of absolute performance, of our solutions with respect to those of [20], as the datasets are different. Rather, it is brought here only to highlight the differences between the methods in terms of the relative performance deterioration when different modalities are available at training and test times.

As can be seen, the single-domain predictors we start with perform relatively poorly (rows 1 and 2). Nevertheless, in the shared-representation settings (rows 3 and 4), our predictors’ accuracy drops by only 7–20% with respect to the corresponding single-domain estimators (rows 1 and 2, respectively). In contrast, the difference in success rates for the RBM predictor is between 30 and 70%.

7. Conclusion

In this paper, we analyzed the problems of multi- and single-domain regression in settings involving distinct unpaired labeled training sets for the different domains and a large unlabeled set of paired examples from all domains. We derived minimax-optimal results and obtained closed-form solutions for many practical scenarios. We used the resulting expressions to study when training data from a domain, which is not available during testing, can help. In particular, we showed that in the setting of cross-modality learning, originally presented in [20], there is no advantage in using the training data from the unobserved domain, at least from a worst-case perspective. We demonstrated our methods in the context of synthesis of a neutral expression face from an image of a smiling subject and in the context of audio-visual spoken digit recognition. In the latter setting, we demonstrated that our approach may be more effective than that proposed in [20]. This is despite the fact that our method is designed for regression rather than classification and even though we applied it on a more challenging audio-visual sentence corpus.

Funding

This work was supported in part by the Office of Naval Research; Army Research Office; Defense Advanced Research Projects Agency; National Security Science and Engineering Faculty Fellowship; National Science Foundation; Israel Science Foundation (grant number 170/10); Ollendorff Minerva Foundation; and Google Research Award.

References

1. Amini, M. R., Usunier, N. & Goutte, C. (2009) Learning from multiple partially observed views–an application to multilingual text categorization. Advances in Neural Information Processing Systems. *Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS)*. British Columbia, Canada: MIT Press.

2. Blum, A. & Mitchell, T. (1998) Combining labeled and unlabeled data with co-training. *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. Madison, WI: ACM, pp. 92–100.

3. Bowden, R. J. & Turkington, D. A. (1984) *Instrumental Variables*. Cambridge: Cambridge University Press.

4. Cooke, M., Barker, J., Cunningham, S. & Shao, X. (2006) An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.*, **120**, 2421–2424.

5. Eldar, Y. C. (2008) Rethinking biased estimation: improving maximum likelihood and the Cramér–Rao bound. *Found. Trends Signal Processing*, **1**, 305–449.

6. Eldar, Y. C., Ben-Tal, A. & Nemirovski, A. (2004) Linear minimax regret estimation of deterministic parameters with bounded data uncertainties. *IEEE Trans. Signal Process.*, **52**, 2177–2188.

7. Eldar, Y. C. & Merhav, N. (2004) A competitive minimax approach to robust estimation of random parameters. *IEEE Trans. Signal Process.*, **52**, 1931–1946.

8. Eldar, Y. C. & Merhav, N. (2005) Minimax MSE-ratio estimation with signal covariance uncertainties. *IEEE Trans. Signal Process.*, **53**, 1335–1347.

9. Harel, M. & Mannor, S. (2011) Learning from multiple outlooks. *Proceedings of the International Conference on Machine Learning (ICML)*. Bellevue, Washington: ACM.

10. Kakade, S. M. & Foster, D. P. (2007) Multi-view regression via canonical correlation analysis. *Proceedings of the 20th Annual Conference on Learning Theory*. San Diego, CA: Springer, pp. 82–96.

11. Kalman, R. E. (1960) A new approach to linear filtering and prediction problems. *J. Basic Eng.*, **82**, 35–45.

12. Kienzle, W., Bakir, G. H., Franz, M. & Schölkopf, B. (2004) Face detection–efficient and rank deficient. *Advances in Neural Information Processing Systems (NIPS)*. British Columbia, Canada, pp. 673–680.

13. Martinez, A. M. & Benavente, R. (1998) The AR face database. (24), *CVC Technical Report*.

14. Michaeli, T. (2010) Face normalization for recognition and enrollment. *Patent*, EP 1872303 B1.

15. Michaeli, T. & Eldar, Y. C. (2010) Optimization techniques in modern sampling theory. *Convex Optimization in Signal Processing and Communications* (Y. C. Eldar, & D. Palomar eds). Cambridge: Cambridge University Press, pp. 266–314.

16. Michaeli, T. & Eldar, Y. C. (2011) Hidden relationships: Bayesian estimation with partial knowledge. *IEEE Trans. Signal Process.*, **59**.

17. Michaeli, T., Eldar, Y. C. & Sapiro, G. (2012) Semi-supervised multi-domain regression with distinct training sets. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kyoto, Japan: IEEE.

18. Michaeli, T., Sigalov, D. & Eldar, Y. C. (2012) Partially linear estimation with application to sparse signal recovery from measurement pairs. *IEEE Trans. Signal Process*, **60**, 2125–2137.

19. Nadaraya, E. A. (1964) On Estimating Regression. *Theory Probab. Appl.*, **9**, 141.

20. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. & Ng, A. Y. (2010) Multimodal deep learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*. British Columbia, Canada: Curran Associates, Inc.

21. Pan, S. J. & Yang, Q. (2010) A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, **22**, 1345–1359.

22. Patterson, E., Gurbuz, S., Tufekci, Z. & Gowdy, J. N. (2002) CUAVE: A new audio-visual database for multimodal human-computer interface research. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Orlando, Florida: IEEE, vol. 2, pp. 2017–2020.

23. Porat, B. (1994) *Digital Processing of Random Signals: Theory and Methods*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

24. Reisfeld, D. & Yeshurun, Y. (1998) Preprocessing of face images: Detection of features and pose normalization. *Comput. Vis. Image Underst.*, **71**, 413–430.

25. Watson, G. S. (1964) Smooth regression analysis. *Sankhya: Indian J. Stat. Ser. A*, **26**, 359–372.

Appendix A. Proof of Theorem 4.1

We begin by proving Claim (1). Since $\mathcal{A}$ is a linear subspace, the orthogonality principle implies that $\varphi_{\mathcal{A}}(X_1)$ is the unique estimator satisfying

$$\mathbb{E}[(Y-\varphi_{\mathcal{A}}(X_1))^\top \varphi(X_1)] = 0, \tag{A.1}$$

for every $\varphi \in \mathcal{A}$. Consequently, for every $\varphi \in \mathcal{A}$, we have that

$$\mathbb{E}[Y^\top \varphi(X_1)] = \mathbb{E}[\varphi_{\mathcal{A}}(X_1)^\top \varphi(X_1)]. \tag{A.2}$$

Similarly, for every $\psi \in \mathcal{B}$, we have that

$$\mathbb{E}[Y^\top \psi(X_2)] = \mathbb{E}[\psi_{\mathcal{B}}(X_2)^\top \psi(X_2)]. \tag{A.3}$$

Finally, as $\mathcal{C} = \mathcal{A} + \mathcal{B}$, the set $\mathcal{C}$ is a subspace as well. Therefore, $\rho_{\mathcal{C}}$ of (4.3) is the unique estimator satisfying

$$\mathbb{E}[Y^\top (\varphi(X_1) + \psi(X_2))] = \mathbb{E}[\rho_{\mathcal{C}}(X_1, X_2)^\top (\varphi(X_1) + \psi(X_2))], \tag{A.4}$$

for every $\varphi \in \mathcal{A}$ and $\psi \in \mathcal{B}$. Substituting (A.2) and (A.3), condition (A.4) reduces to the requirement that

$$\mathbb{E}[\varphi_{\mathcal{A}}(X_1)^\top \varphi(X_1)] + \mathbb{E}[\psi_{\mathcal{B}}(X_2)^\top \psi(X_2)] = \mathbb{E}[\rho_{\mathcal{C}}(X_1, X_2)^\top (\varphi(X_1) + \psi(X_2))], \tag{A.5}$$

for every $\varphi \in \mathcal{A}$ and $\psi \in \mathcal{B}$. Now, the $\mathcal{A}$- and $\mathcal{B}$-optimal estimators of Y from X_1 and X_2 are fixed over $\mathcal{D}$ (given by $\varphi_{\mathcal{A}}$ and $\psi_{\mathcal{B}}$, respectively). Furthermore, all expectations in (A.5) are with respect to $F_{X_1X_2}$, which is also fixed over $\mathcal{D}$. This implies that the function $\rho_{\mathcal{C}}$ does not depend on the choice of $F_{X_1X_2Y} \in \mathcal{D}$, completing the proof of Claim (1).

To prove Claim (2), we note that from the orthogonality principle (A.4) follows the Pythagorean relation

$$\mathbb{E}[\|Y-\rho_{\mathcal{C}}(X_1, X_2)\|^2] = \mathbb{E}[\|Y\|^2] - \mathbb{E}[\|\rho_{\mathcal{C}}(X_1, X_2)\|^2]. \tag{A.6}$$

The first term on the right-hand side equals c for every $F_{X_1X_2Y} \in \mathcal{D}$. We have also seen that $\rho_{\mathcal{C}}(X_1, X_2)$ is fixed over $\mathcal{D}$. Moreover, the expectation in the second term is with respect to $F_{X_1X_2}$, which is fixed over $\mathcal{D}$. Therefore, the second term, as well, does not depend on the choice of $F_{X_1X_2Y} \in \mathcal{D}$. This completes the proof of Claim (2).

Lastly, we prove Claim (3). To do so, we first note that $\varphi_{\mathcal{A}}(X_1)$ and $\psi_{\mathcal{B}}(X_2)$ are not only the $\mathcal{A}$- and $\mathcal{B}$-optimal estimators of Y based on X_1 and X_2, respectively; they are also the $\mathcal{A}$- and $\mathcal{B}$-optimal estimators of $\rho_{\mathcal{C}}(X_1,X_2)$. To see this, note that both $\mathcal{A}$ and $\mathcal{B}$ are contained in $\mathcal{C}$. Consequently, the orthogonality principle implies that, for every $\varphi \in \mathcal{A}$ (which is also in $\mathcal{C}$), we have

$$\mathbb{E}[\|Y-\varphi(X_1)\|^2] = \mathbb{E}[\|Y-\rho_{\mathcal{C}}(X_1,X_2)\|^2] + \mathbb{E}[\|\rho_{\mathcal{C}}(X_1,X_2)-\varphi(X_1)\|^2]. \tag{A.7}$$

As the first term does not depend on φ, we see that minimization of the MSE over $\varphi \in \mathcal{A}$ is equivalent to minimization of the second term alone. Thus, $\varphi_{\mathcal{A}}(X_1)$ is the $\mathcal{A}$-optimal estimate of $\rho_{\mathcal{C}}(X_1,X_2)$ given X_1. The same argument can be invoked to deduce that $\psi_{\mathcal{B}}(X_2)$ is the $\mathcal{B}$-optimal estimate of $\rho_{\mathcal{C}}(X_1,X_2)$ from X_2.

A second observation we need for proving Claim (3) follows from the fact that $\mathcal{A}$ and $\mathcal{B}$ are linear sets. Specifically, this implies that if $\varphi_1^*(V)$ and $\varphi_2^*(V)$ are the $\mathcal{A}$-optimal estimates of the two RVs W_1 and W_2, respectively, based on the RV V, then the $\mathcal{A}$-optimal estimate of W_1+W_2 is $\varphi_1^*(V)+\varphi_2^*(V)$. This can be seen by noting that the estimator $\varphi_1^*(V)+\varphi_2^*(V)$ satisfies the orthogonality principle, namely, for any $\varphi \in \mathcal{A}$, we have that

$$\begin{aligned}\mathbb{E}[(W_1+W_2-\varphi_1^*(W_1)-\varphi_2^*(W_1))^\top\varphi(W_1)] &= \mathbb{E}[(W_1-\varphi_1^*(W_1))^\top\varphi(W_1)] + \mathbb{E}[(W_2-\varphi_2^*(W_1))^\top\varphi(W_1)]\\ &= 0.\end{aligned} \tag{A.8}$$

The statement also holds, of course, with respect to $\mathcal{B}$-optimal estimates.

Following these two observations, for any $F_{X_1X_2Y} \in \mathcal{D}$, setting $\tilde{Y} = 2\rho_{\mathcal{C}}(X_1,X_2) - Y$ results in a distribution $F_{X_1X_2\tilde{Y}}$ that also belongs to $\mathcal{D}$. This is because the $\mathcal{A}$-optimal estimate of $\tilde{Y}$ from X_1 equals twice the $\mathcal{A}$-optimal estimate of $\rho_{\mathcal{C}}(X_1,X_2)$ from X_1 (which is $\varphi_{\mathcal{A}}(X_1)$) minus the $\mathcal{A}$-optimal estimate of Y from X_1 (which is also $\varphi_{\mathcal{A}}(X_1)$). Namely, the $\mathcal{A}$-optimal estimate of $\tilde{Y}$ from X_1 is $\varphi_{\mathcal{A}}(X_1)$. Similarly, the $\mathcal{B}$-optimal estimate of $\tilde{Y}$ from X_2 is $\psi_{\mathcal{B}}(X_2)$. Finally, due to the orthogonality principle, the second-order moment of $\tilde{Y}$ is given by

$$\begin{aligned}\mathbb{E}[\|\tilde{Y}\|^2] &= \mathbb{E}[\|\rho_{\mathcal{C}}(X_1,X_2)\|^2] + \mathbb{E}[\|Y-\rho_{\mathcal{C}}(X_1,X_2)\|^2]\\ &= \mathbb{E}[\|\rho_{\mathcal{C}}(X_1,X_2)\|^2] + \mathbb{E}[\|Y\|^2] - \mathbb{E}[\rho_{\mathcal{C}}(X_1,X_2)\|^2]\\ &= c.\end{aligned} \tag{A.9}$$

We now use this fact to prove Claim (3). The orthogonality principle (A.4) implies that the MSE attained by any estimator ρ satisfies

$$\begin{aligned}\mathbb{E}[\|Y-\rho(X_1,X_2)\|^2] &= \mathbb{E}[\|Y-\rho_{\mathcal{C}}(X_1,X_2)\|^2] + \mathbb{E}[\|\rho_{\mathcal{C}}(X_1,X_2)-\rho(X_1,X_2)\|^2]\\ &\quad + 2\mathbb{E}[(Y-\rho_{\mathcal{C}}(X_1,X_2))^\top(\rho_{\mathcal{C}}(X_1,X_2)-\rho(X_1,X_2))]\\ &= \mathbb{E}[\|Y-\rho_{\mathcal{C}}(X_1,X_2)\|^2] + \mathbb{E}[\|\rho_{\mathcal{C}}(X_1,X_2)-\rho(X_1,X_2)\|^2]\\ &\quad + 2\mathbb{E}[(\rho_{\mathcal{C}}(X_1,X_2)-Y)^\top\rho(X_1,X_2)].\end{aligned} \tag{A.10}$$

The first term in this expression is not a function of ρ and, as we have seen in (A.6), is constant as a function of $F_{X_1X_2Y}$ over $\mathcal{D}$. The second term is a function of ρ, but since the expectation is with respect to $F_{X_1X_2}$, it is constant as a function of $F_{X_1X_2Y}$ over $\mathcal{D}$. Therefore,

$$\min_{\rho} \sup_{F_{X_1X_2Y}\in\mathcal{D}} \text{MSE}(F_{X_1X_2Y},\rho) = \mathbb{E}[\|Y-\rho_{\mathcal{C}}(X_1,X_2)\|^2] + \min_{\rho}\{\mathbb{E}[\|\rho_{\mathcal{C}}(X_1,X_2)-\rho(X_1,X_2)\|^2] + \sup_{F_{X_1X_2Y}\in\mathcal{D}} 2\mathbb{E}[(\rho_{\mathcal{C}}(X_1,X_2)-Y)^\top\rho(X_1,X_2)]\}. \tag{A.11}$$

We saw that, for every $F_{X_1X_2Y}\in\mathcal{D}$, setting $\tilde{Y}=2\rho_{\mathcal{C}}(X_1,X_2)-Y$ results in a distribution $F_{X_1X_2\tilde{Y}}$ that also belongs to $\mathcal{D}$. Now, with $F_{X_1X_2\tilde{Y}}$, the expression $2\mathbb{E}[(\rho_{\mathcal{C}}(X_1,X_2)-\tilde{Y})^\top\rho(X_1,X_2)]$ equals $-2\mathbb{E}[(\rho_{\mathcal{C}}(X_1,X_2)-Y)^\top\rho(X_1,X_2)]$. Consequently, the maximum of this term over $F_{X_1X_2Y}\in\mathcal{D}$ is necessarily non-negative. We thus have that

$$\begin{aligned}\min_{\rho} \sup_{F_{X_1X_2Y}\in\mathcal{D}} \text{MSE}(F_{X_1X_2Y},\rho) &\geqslant \mathbb{E}[\|Y-\rho_{\mathcal{C}}(X_1,X_2)\|^2] + \min_{\rho}\mathbb{E}[\|\rho_{\mathcal{C}}(X_1,X_2)-\rho(X_1,X_2)\|^2]\\ &= \mathbb{E}[\|Y-\rho_{\mathcal{C}}(X_1,X_2)\|^2],\end{aligned} \tag{A.12}$$

where we used the fact that the minimal value of 0 is attained with $\rho(X_1,X_2)=\rho_{\mathcal{C}}(X_1,X_2)$.

We have established a lower bound on the worst-case MSE of any estimator. Next, we show that the estimator $\rho(X_1,X_2)=\rho_{\mathcal{C}}(X_1,X_2)$ attains this bound, which proves that it is minimax-optimal. Indeed, substituting this solution into (A.10), we find that

$$\sup_{F_{X_1X_2Y}\in\mathcal{D}} \text{MSE}(F_{X_1X_2Y},\rho_{\mathcal{C}}) = \mathbb{E}[\|Y-\rho_{\mathcal{C}}(X_1,X_2)\|^2], \tag{A.13}$$

completing the proof.

Appendix B. Proof of Theorems 4.2 and 5.1

We simultaneously prove Theorems 4.2 and 5.1 by using an auxiliary RV Z, which can be any (fixed) function of X_1 and X_2. Therewith, we will study the solution to

$$\arg\min_{\rho} \sup_{F_{X_1X_2Y}\in\mathcal{D}} \text{REG}(F_{X_1X_2Y},\rho), \tag{B.1}$$

where minimization is performed over all functions ρ of Z and the regret is with respect to $\mathbb{E}[Y|Z]$. Specifically, we will show that the solution to this problem is given by $\mathbb{E}[\rho_{\text{M}}(X_1,X_2)|Z]$. Setting, $Z=(X_1^\top,X_2^\top)^\top$, we get $\mathbb{E}[\rho_{\text{M}}(X_1,X_2)|Z]=\rho_{\text{M}}(X_1,X_2)$, proving Theorem 4.2. Setting $Z=X_1$, the solution becomes $\mathbb{E}[\rho_{\text{M}}(X_1,X_2)|X_1]$, proving Theorem 5.1.

Expressing $Y=\rho_{\text{M}}(X_1,X_2)+(Y-\rho_{\text{M}}(X_1,X_2))$, the regret of any estimator $\rho(Z)$ can be written as

$$\mathbb{E}[\|\mathbb{E}[Y|Z]-\rho(Z)\|^2] = \mathbb{E}[\|\mathbb{E}[\rho_{\text{M}}(X_1,X_2)|Z]-\rho(Z)\|^2] + \mathbb{E}[\|\mathbb{E}[Y-\rho_{\text{M}}(X_1,X_2)|Z]\|^2] + 2\mathbb{E}[\mathbb{E}[Y-\rho_{\text{M}}(X_1,X_2)|Z]^\top(\mathbb{E}[\rho_{\text{M}}(X_1,X_2)|Z]-\rho(Z))]. \tag{B.2}$$

Since the marginal distribution $F_{X_1X_2}$ is fixed over $\mathcal{D}$, the first term in the above expression does not depend on the choice of $F_{X_1X_2Y} \in \mathcal{D}$. Consequently,

$$\sup_{F_{X_1X_2Y}\in\mathcal{D}} \mathrm{REG}(F_{X_1X_2Y},\rho) = \mathbb{E}[\|\mathbb{E}[\rho_{\mathrm{M}}(X_1,X_2)|Z] - \rho(Z)\|^2] + \sup_{F_{X_1X_2Y}\in\mathcal{D}} \{\mathbb{E}[\|\mathbb{E}[Y-\rho_{\mathrm{M}}(X_1,X_2)|Z]\|^2] \\ + 2\mathbb{E}[\mathbb{E}[Y-\rho_{\mathrm{M}}(X_1,X_2)|Z]^\top(\mathbb{E}[\rho_{\mathrm{M}}(X_1,X_2)|Z]-\rho(Z))]\}. \tag{B.3}$$

As we have seen in Appendix A, for every $F_{X_1X_2Y}\in\mathcal{D}$, setting $\tilde{Y} = 2\rho_{\mathrm{M}}(X_1,X_2) - Y$ results in a distribution $F_{X_1X_2\tilde{Y}}$ that also belongs to $\mathcal{D}$. Now, $\tilde{Y} - \rho_{\mathrm{M}}(X_1,X_2) = -(Y - \rho_{\mathrm{M}}(X_1,X_2))$, implying that if $F_{X_1X_2Y}$ maximizes the first term within the braces, then either $F_{X_1X_2Y}$ or $F_{X_1X_2\tilde{Y}}$ yields at least the same value for the objective comprising both terms. Therefore,

$$\begin{aligned}\min_{\rho} \sup_{F_{X_1X_2Y}\in\mathcal{D}} \mathrm{REG}(F_{X_1X_2Y},\rho) &\geqslant \min_{\rho} \mathbb{E}[\|\mathbb{E}[\rho_{\mathrm{M}}(X_1,X_2)|Z]-\rho(Z)\|^2] \\ &\quad + \sup_{F_{X_1X_2Y}\in\mathcal{D}} \mathbb{E}[\|\mathbb{E}[Y-\rho_{\mathrm{M}}(X_1,X_2)|Z]\|^2] \\ &= \sup_{F_{X_1X_2Y}\in\mathcal{D}} \mathbb{E}[\|\mathbb{E}[Y-\rho_{\mathrm{M}}(X_1,X_2)|Z]\|^2],\end{aligned} \tag{B.4}$$

where the last equality is due to the fact that $\rho(Z) = \mathbb{E}[\rho_{\mathrm{M}}(X_1,X_2)|Z]$ achieves the minimal value of 0 in the first term.

We established a lower bound on the worst-case regret of any estimator. Next, we show that the estimator $\rho^*(Z) = \mathbb{E}[\rho_{\mathrm{M}}(X_1,X_2)|Z]$ attains this bound, which proves that it is minimax-optimal. Indeed, substituting this solution into (B.3), we find that

$$\sup_{F_{X_1X_2Y}\in\mathcal{D}} \mathrm{REG}(F_{X_1X_2Y},\rho_{\mathrm{M}}) = \sup_{F_{X_1X_2Y}\in\mathcal{D}} \mathbb{E}[\|\mathbb{E}[Y-\rho_{\mathrm{M}}(X_1,X_2)|Z]\|^2], \tag{B.5}$$

completing the proof.

Appendix C. Proof of Theorem 4.4

To prove the claim, we show that the estimation error corresponding to $\rho_{\mathcal{C}}(X_1,X_2)$ of (4.18) is uncorrelated with every RV of the form $\varphi(X_1)+\psi(X_2)$ with $\varphi\in\mathcal{A}$ and $\psi\in\mathcal{B}$. Indeed, for every $\varphi\in\mathcal{A}$, the estimator $\rho_{\mathcal{C}}(X_1,X_2)$ of (4.18) satisfies

$$\begin{aligned}\mathbb{E}[(Y-\rho_{\mathcal{C}}(X_1,X_2))^\top\varphi(X_1)] &= \mathbb{E}[(Y-\varphi_{\mathcal{A}}(X_1))^\top\varphi(X_1)] - \mathbb{E}[\rho_{\mathcal{B}|\mathcal{A}}^\top(X_2|X_1)\varphi(X_1)] \\ &= \mathbb{E}[(\psi(X_2)-\eta_\psi(X_1))^\top\varphi(X_1)] \\ &= 0,\end{aligned} \tag{C.1}$$

where we used the orthogonality principle. To prove orthogonality with respect to RVs of the form $\psi(X_2)$, with $\psi\in\mathcal{B}$, we write $\psi(X_2) = \psi(X_2) - \eta_\psi(X_1) + \eta_\psi(X_1)$, where $\eta_\psi(X_1)$ is the $\mathcal{A}$-optimal estimate of $\psi(X_2)$ based on X_1. By the orthogonality principle, the errors $Y-\varphi_{\mathcal{A}}(X_1)$ and $\rho_{\mathcal{B}|\mathcal{A}}(X_2|X_1) = \psi(X_2)-\eta_\psi(X_1)$ are uncorrelated with any RV $\eta(X_1)$, where $\eta\in\mathcal{A}$, and thus, in particular, with the term

$\eta_\psi(X_1)$. Therefore, we have that

$$\begin{aligned}\mathbb{E}[(Y-\hat{Y})^\top\psi(X_2)] &= \mathbb{E}[(Y-\varphi_{\mathcal{A}}(X_1)-\rho_{\mathcal{B}|\mathcal{A}}(X_2|X_1))^\top(\psi(X_2)-\eta_\psi(X_1))]\\ &= \mathbb{E}[(Y-\rho_{\mathcal{B}|\mathcal{A}}(X_2|X_1))^\top(\psi(X_2)-\eta_\psi(X_1))]\\ &= 0. \end{aligned} \tag{C.2}$$

Here, the second equality results from the fact that the term $\psi(X_2)-\eta_\psi(X_1)$ is orthogonal to every RV $\varphi(X_1)$, where $\varphi\in\mathcal{A}$, and, in particular, to $\varphi_{\mathcal{A}}(X_1)$. The third equality follows from the fact that $\rho_{\mathcal{B}|\mathcal{A}}(X_2|X_1)$ is the MMSE estimate of Y among all functions of the form $\psi(X_2)-\eta_\psi(X_1)$, with ψ being some function in $\mathcal{B}$ and $\eta_\psi(X_1)$ being the $\mathcal{A}$-optimal estimator of $\psi(X_2)$ from X_1. Consequently, the error $Y-\rho_{\mathcal{B}|\mathcal{A}}(X_2|X_1)$ is orthogonal to every RV of the form $\psi(X_2)-\eta_\psi(X_1)$, and, in particular, to $\psi(X_2)-\eta_\psi(X_1)$.

Appendix D. An example of a trivial single-domain minimax MSE solution

The MSE of a single-domain estimator $\rho(X_1)$ is defined as

$$\mathrm{MSE}(F_{X_1X_2Y},\rho)=\mathbb{E}[\|Y-\rho(X_1)\|^2]. \tag{D.1}$$

Consider the single-domain minimax MSE problem

$$\arg\min_{\rho}\sup_{F_{X_1X_2Y}\in\mathcal{D}}\mathrm{MSE}(F_{X_1X_2Y},\rho), \tag{D.2}$$

where $\mathcal{D}$ is the set of all joint distributions $F_{X_1X_2Y}$ such that

$$\mathbb{E}[\|Y\|^2]=1,\quad \mathbb{E}[Y|X_2]=\frac{\sqrt{2}}{2}X_2,\quad \begin{pmatrix}X_1\\X_2\end{pmatrix}\sim\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & \frac{\sqrt{2}}{2}\\ \frac{\sqrt{2}}{2} & 1\end{pmatrix}\right). \tag{D.3}$$

We will show that the minimax MSE estimator in this case is given by $\rho_{\mathrm{MM}}(X_1)=0$. Note that the worst-case MSE of this estimator is

$$\sup_{F_{X_1X_2Y}\in\mathcal{D}}\mathrm{MSE}(F_{X_1X_2Y},\rho_{\mathrm{MM}})=\sup_{F_{X_1X_2Y}\in\mathcal{D}}\mathbb{E}[\|Y\|^2]=1. \tag{D.4}$$

To demonstrate the minimax-optimality of ρ_{MM}, we note that one of the feasible distributions $F_{X_1X_2Y}$ in this setting corresponds to the case $Y=\sqrt{2}X_2-X_1$. Indeed, it can be easily verified that $\mathbb{E}[(\sqrt{2}X_2-X_1)^2]=1$ and $\mathbb{E}[\sqrt{2}X_2-X_1|X_2]=\sqrt{2}X_2/2$. Let us denote this distribution by $\mathcal{F}^*_{X_1X_2Y}$. Now, the worst-case MSE of any estimator ρ can be lower bounded by the MSE attained by ρ under the distribution $\mathcal{F}^*_{X_1X_2Y}$:

$$\sup_{F_{X_1X_2Y}\in\mathcal{D}}\mathrm{MSE}(F_{X_1X_2Y},\rho)\geqslant\mathrm{MSE}(F^*_{X_1X_2Y},\rho)=\mathbb{E}[(\sqrt{2}X_2-X_1-\rho(X_1))^2]. \tag{D.5}$$

This value can be further lower bounded by the MSE attained by the MMSE estimator under $\mathcal{F}^*_{X_1X_2Y}$, which is $\rho(X_1) = \mathbb{E}[\sqrt{2}X_2 - X_1|X_1] = 0$, leading to

$$\sup_{F_{X_1X_2Y}\in\mathcal{D}} \mathrm{MSE}(F_{X_1X_2Y}, \rho) \geqslant \mathbb{E}[(\sqrt{2}X_2 - X_1)^2] = 1. \tag{D.6}$$

We have thus established that the worst-case MSE of any estimator is lower bounded by 1, whereas the worst-case MSE of the estimator $\rho_{\mathrm{MM}}(X_1) = 0$ is exactly 1, demonstrating that ρ_{MM} is minimax-optimal in an MSE sense.

As we show in Section 5.2, the minimax-regret estimator in this setting is given by

$$\rho_{\mathrm{MR}} = \mathbb{E}[\mathbb{E}[Y|X_2]|X_1] = \tfrac{1}{2}X_1. \tag{D.7}$$

Therefore, we see that the minimax-regret and minimax MSE strategies generally lead to two different solutions in the single-domain estimation situation.