Journal of Machine Learning Research 23 (2022) 1-49 Submitted 9/21; Revised 8/22; Published 11/22

Fundamental Limits and Tradeoffs in Invariant Representation Learning

Han Zhao HANZHAO@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Chen Dan CDAN@CS.CMU.EDU
Carnegie Mellon University

Bryon Aragam BRYON@CHICAGOBOOTH.EDU
University of Chicago

Tommi S. Jaakkola TOMMI@CSAIL.MIT.EDU
Massachusetts Institute of Technology

Geoffrey J. Gordon GGORDON@CS.CMU.EDU
Carnegie Mellon University

Pradeep Ravikumar PRADEEPR@CS.CMU.EDU
Carnegie Mellon University

Editor: Adrian Weller

Abstract

A wide range of machine learning applications such as privacy-preserving learning, algorithmic fairness, and domain adaptation/generalization among others, involve learning invariant representations of the data that aim to achieve two competing goals: (a) maximize information or accuracy with respect to a target response, and (b) maximize invariance or independence with respect to a set of protected features (e.g. for fairness, privacy, etc). Despite their wide applicability, theoretical understanding of the optimal tradeoffs — with respect to accuracy, and invariance — achievable by invariant representations is still severely lacking. In this paper, we provide an information theoretic analysis of such tradeoffs under both classification and regression settings. More precisely, we provide a geometric characterization of the accuracy and invariance achievable by any representation of the data; we term this feasible region the information plane. We provide an inner bound for this feasible region for the classification case, and an exact characterization for the regression case, which allows us to either bound or exactly characterize the Pareto optimal frontier between accuracy and invariance. Although our contributions are mainly theoretical, a key practical application of our results is in certifying the potential sub-optimality of any given representation learning algorithm for either classification or regression tasks. Our results shed new light on the fundamental interplay between accuracy and invariance, and may be useful in guiding the design of future representation learning algorithms.

Keywords: Invariant representation learning, domain adaptation, fairness, privacy-preservation, information theory

. Equal contribution

1. Introduction

A key initial step in any machine learning pipeline is to find a suitable feature representation of the raw data. When we have a specific target response in mind, a common approach is to obtain data representations that maximize predictive accuracy with respect to that target response. In practice, however, one typically faces many competing criteria beyond just accuracy. Examples of such criteria include invariance (typically to a group action), statistical independence (with respect to some feature), privacy (so that certain sensitive information cannot be inferred or otherwise exploited), and out-of-domain generalization (across multiple domains and/or tasks). Along these lines, there has been a surge of recent interest in learning invariant representations (Zemel et al., 2013; Hamm, 2015; Anselmi et al., 2016a; Ganin et al., 2016; Johansson et al., 2016; Zhao et al., 2019a) i.e. feature transformations of the input data that aim to balance two goals simultaneously. First, the features should preserve enough information with respect to the target task of interest, e.g., achieve high predictive accuracy. Second, the representations should be invariant to changes in some attribute. Often there is a tension between these two competing goals of accuracy and invariance maximization. Despite wide applicability of invariant representations however, a theoretical understanding of the limits and tradeoffs between accuracy and invariance achievable by any representation learning algorithm is severely lacking.

In this paper, towards such an understanding, we provide an inner bound on the "feasible region" of the tuple of accuracy and invariance of any data representation. The geometric properties of this feasible region can then be related to the Pareto optimal tradeoffs between accuracy and invariance (cf. Section 3 and Figure 1). Interestingly, given existing representation learning algorithms, our proof technique also implies a constructive approach using randomization to achieve any desired interpolation among the given accuracy and invariance. Although our contributions are mainly theoretical, we also demonstrate a practical application of our results in certifying the suboptimality of several existing representation learning algorithms in both classification and regression tasks. Overall, we believe our results take an important step towards a better understanding of the interplay between accuracy and invariance.

1.1 Our Contributions

Our main contributions can be summarized as follows:

- We provide an explicit characterization of the feasible region of accuracy and invariance (which we term the information plane; see Section 3 for definitions) for any representation of the data (Sections 4, 5). This can be used to characterize the tradeoffs between accuracy and invariance, as well as the sub-optimality of any learned representation.

- While we provide an inner bound of the feasible region for classification, we are able to exactly characterize this feasible region for regression tasks. Moreover, we derive an analytic solution to characterize the Pareto optimal frontier with respect to accuracy and invariance, and discuss sufficient conditions under which this frontier could be achievable.

- Finally, we demonstrate a practical application of our theoretical results to certify the sub-optimality of existing representation learning algorithms. In particular, we compare several existing invariant representation learning algorithms on real datasets for both classification and regression tasks using our characterization of the information plane.

We first illustrate our results in the discrete, noiseless setting, which already presents some nontrivial analytical challenges (Section 4). We then generalize to the noisy, continuous setting, where we can exploit additional machinery in order to refine these results further (Section 5).

1.2 Related Work

There are abundant applications of learning invariant representations in various downstream tasks, including domain adaptation (Ben-David et al., 2007, 2010; Ganin et al., 2016; Zhao et al., 2018), algorithmic fairness (Edwards and Storkey, 2015; Zemel et al., 2013; Zhang et al., 2018; Zhao et al., 2019b), privacy-preserving learning (Hamm, 2015, 2017; Coavoux et al., 2018; Xiao et al., 2019), invariant visual representations (Quiroga et al., 2005; Mallat, 2012; Anselmi et al., 2016b), causal inference (Johansson et al., 2016; Shalit et al., 2017; Johansson et al., 2020), and multilingual machine translation (Johnson et al., 2017; Aharoni et al., 2019; Zhao et al., 2020).

To the best of our knowledge, no previous work studies the particular tradeoff problem in this paper. Closest to our work are results in domain adaptation (Zhao et al., 2019a) and algorithmic fairness (Menon and Williamson, 2018; Zhao and Gordon, 2019), where the authors have shown a lower bound on the classification accuracy on two groups. Compared to these previous results, our work directly characterizes the tradeoff between accuracy and invariance in both classification and regression settings. Furthermore, we also give an approximation to the Pareto frontier between accuracy and invariance in both cases. Another line of related work is the information bottleneck method (Tishby et al., 2000; Tishby and Zaslavsky, 2015), where the goal is to learn representations that preserve sufficient statistics w.r.t. the target task of interest while at the same time are compressed as much as possible. The main difference between the information bottleneck and the invariant representations analyzed in this paper is that in the former there is no feature that learned representations are required to be invariant to.

2. Preliminaries

Consider the typical supervised learning setting, where we have an input random vector $X \in \mathcal{X}$, and a target response random variable $Y \in \mathcal{Y}$. In either classification or regression, we then aim to estimate a prediction function $f : \mathcal{X} \to \mathcal{Y}$ that minimizes $\mathbf{E}\ell(f(X), Y)$ for some loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbf{R}$. In addition to these input and response variables, in our setting there is a third random variable A, with respect to which we desire that our prediction function be invariant. As some examples, A could correspond to potential protected attributes (such as ethnicity or gender of an individual) in algorithmic fairness, or could index the environment or domain in domain adaptation. We let D denote the joint distribution over the triple (X, A, Y), from which our observational data are sampled from.

Throughout the paper, we will use $H()$ to denote the entropy of a random variable and $I(;)$ to denote the mutual information between a pair of random variables. For any two random variables X, Y over the same sample space, we also use $X \perp Y$ to denote their statistical independence. As usual, we use $\mathbf{E}[]$ and $\mathrm{Var}()$ to denote the expectation and variance of a random variable, respectively.

Upon receiving the data, the goal of the learner is twofold. On the one hand, the learner aims to accurately predict the target Y. On the other hand, it also tries to be insensitive to variation in A. To achieve these dual goals, a standard approach in the literature (Zemel et al., 2013; Edwards and Storkey, 2015; Hamm, 2015; Ganin et al., 2016; Zhao et al., 2018) is invariant representation

learning. Formally, we are interested in learning a (possibly randomized) transformation function $Z = g(X)$ that contains as much information as possible about the target Y, while at the same time filtering out information related to A.

To make these notions more precise, we first define $R_Y(g) := \inf_h \mathbf{E}_D \ell(h(g(X)), Y)$, as the optimal risk in predicting Y using the feature encoding $Z = g(X)$ under loss ℓ. We similarly define $R_A(g)$ to be the optimal risk in predicting A using the feature encoding $Z = g(X)$ under some loss ℓ. Phrased in terms of these risk functions $R_Y(g)$ and $R_A(g)$, the goal of invariant representation learning is to find a representation g such that the response prediction loss $R_Y(g)$ is minimized while the "invariance variable" prediction loss $R_A(g)$ is maximized.

In this paper, we consider two canonical choices of ℓ:

1. (Classification) The cross entropy loss, i.e. $\ell(y,y') = -y \log(y') - (1-y)\log(1-y')$, typically used in classification when Y is a discrete variable. Under cross-entropy loss, using standard information-theoretic identities (see Appendix A for more detailed derivations), the risk can be written as

$$R_Y(g) = H(Y \mid Z), \quad R_A(g) = H(A \mid Z). \tag{1}$$

2. (Regression) The squared loss, i.e. $\ell(y,y') = (y - y')^2$, which is suitable for regression tasks with continuous Y. Under the squared loss, by the law of total variance (see Appendix A for more detailed derivations), the risk functions can be written as

$$R_Y(g) = \mathbf{E}\,\mathrm{Var}(Y \mid Z), \quad R_A(g) = \mathbf{E}\,\mathrm{Var}(A \mid Z). \tag{2}$$

It is worth pointing out that by using the above risks directly in our framework of analysis, we are focused on the information-theoretic limits of the population errors incurred by the learned representations. Put it in another way, this means that our analysis and results will be oblivious to the optimization methods used to learn the representations, as well as the finite sample effects that exist in practical applications. This is a significant advantage of our approach since our results apply to any algorithm used to learn invariant representations, including future developments in this rapidly growing area.

2.1 Motivating Examples

We next discuss several motivating examples to which our framework can be applied. As can be seen from the wide range of these examples, the framework is very generally applicable, and analyzing the tradeoffs discussed in the previous section is of considerable interest.

Example 2.1 (Privacy-Preservation) In privacy applications, the goal is to make it difficult to predict sensitive data, represented by the attribute A, while retaining information about Y. One way to achieve this is to pass information through Z, the "privatized" data, while simultaneously to preserve the information related to the target application.

Example 2.2 (Algorithmic Fairness) In fairness applications, we seek to make predictions about the response Y without discriminating based on the information contained in the protected attributes A. For example, A may represent a protected class of individuals defined by, e.g. race or gender. This definition of fairness is also known as statistical parity in the literature, and has received increasing

attention recently from an information-theoretic perspective (McNamara et al., 2019; Zhao and Gordon, 2019; Dutta et al., 2019). In particular, one way to achieve this goal is through learning fair representations (Zemel et al., 2013; Madras et al., 2018; Zhao et al., 2019b).

Example 2.3 (Domain Adaptation / Domain Generalization) In domain adaptation and domain generalization, the goal is to train a predictor using labeled data from the source domain that generalizes to the target domain. In this case, A corresponds to the identity (or index) of domains, and the hope here is to learn domain-invariant representations Z that are informative about the target Y (Ben-David et al., 2007, 2010; Zhao et al., 2018; Muandet et al., 2013; Li et al., 2018).

Example 2.4 (Group Invariance) In computer vision, it is desirable to learn predictors that are invariant to the action of a group G on the input space. Typical examples include rotation, translation, and scale. By considering random variables A that take their values in G, one approach to this problem is to learn representations Z that "ignore" changes in A while preserving discriminative features. For example, the recent proposal of contrastive learning (Chen et al., 2020; He et al., 2020; Khosla et al., 2020) falls into this category.

Example 2.5 (Information Bottleneck) The information bottleneck (Tishby et al., 2000) is the problem of finding a representation Z that maximizes the objective $I(Z;Y) \; / I(Z;X)$ in an unsupervised manner. This is related to, but not the same as the problem we study, owing to the additional invariant attribute A in our setting. Instead, the goal of information bottleneck methods are to obtain compressed representations that also preserve target-related information. Hence at a high-level, information bottleneck seeks to find a tradeoff between accuracy and compression.

3. Feasible Region for Accuracy & Invariance

Every representation $Z = g(X)$ can be represented by a point in a two-dimensional information plane, which is a notion we introduce in the sequel. The information plane not only facilitates our quantitative analysis, but also enables visualizing the quality of a learned representation with respect to the two competing imperatives of accuracy and invariance, and further, makes it easier to compare different representations.

Let us first consider the classification setting. We have that the risk can be expressed as:

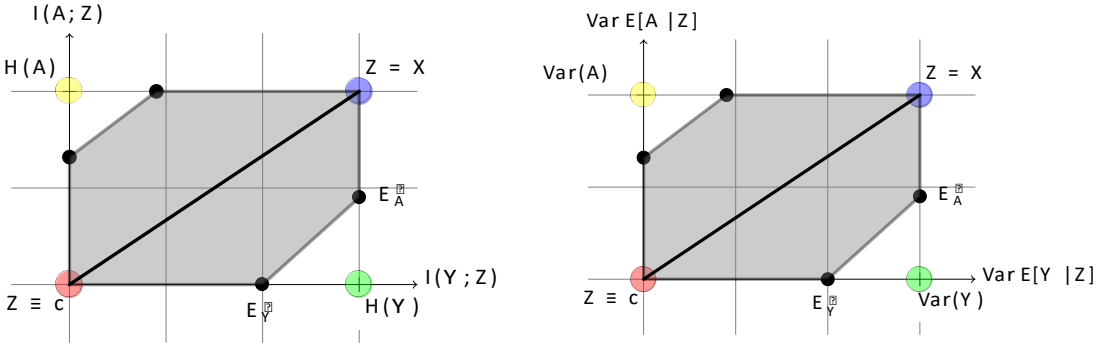
$$R_Y(g) = H(Y \mid Z) = H(Y) \quad I(Y;Z).$$

Noting that $H(Y)$ does not depend on Z, we can thus characterize the quality of the representations $Z = g(X)$ with respect to accuracy, i.e. the $R_Y(g)$, by simply measuring the non-constant term $I(Y;Z)$. Similarly, instead of $R_A(g)$, we could gauge the quality of the representation $Z = g(X)$ with respect to invariance by simply measuring the non-constant term $I(A;Z)$.

Similarly, for the regression setting, we have the risk defined as:

$$R_Y(g) = \mathbb{E}\,\mathrm{Var}(Y \mid Z) = \mathrm{Var}(Y) \quad \mathrm{Var}\,\mathbb{E}[Y \mid Z].$$

As before, since $\mathrm{Var}(Y)$ does not depend on the representations Z, we can then characterize the quality of the representations $Z = g(X)$ with respect to accuracy, i.e. the $R_Y(g)$, by measuring the non-constant term $\mathrm{Var}\,\mathbb{E}[Y \mid Z]$. Likewise, instead of working with $R_A(g)$ directly, we could



(a) Classification (b) Regression

Figure 1: 2D information plane. The shaded area correspond to the feasible region.

measure the invariance of the representations $Z = g(X)$ by looking at the non-constant term $\mathrm{Var}\,\mathbf{E}[A \mid Z]$.

We refer to the space of all possible pairs $(I(Y;Z), I(A;Z))$, as we range over possible random vectors Z, the information plane. With a slight abuse of terminology, we also refer to the space of all possible pairs $(\mathrm{Var}\,\mathbf{E}[Y \mid Z], \mathrm{Var}\,\mathbf{E}[A \mid Z])$, as we range over possible random vectors Z, as the information plane. But we are interested in a particular class of random vectors $Z = g(X)$ that are transformations of the input X; these then specify the following regions:

$$\text{(Classification):}\quad R_{CE} := \{(I(Y;Z), I(A;Z)) : Z = g(X)\},$$
$$\text{(Regression):}\quad R_{LS} := \{(\mathrm{Var}\,\mathbf{E}[Y \mid Z], \mathrm{Var}\,\mathbf{E}[A \mid Z]) : Z = g(X)\}.$$

Both R_{CE} and R_{LS} are subsets of the information plane, which we refer to as feasible regions for the classification and regression settings, respectively, since each point in this feasible region corresponds to the quality of any representation $Z = g(X)$ with respect to accuracy, and the quality with respect to invariance. We provide an illustration of the information planes and the feasible regions for both cases in Figure 1.

The primary goal of invariant representation learning is to find representations Z such that $I(Y;Z)$ (resp. $\mathrm{Var}\,\mathbf{E}[Y \mid Z]$) is large and $I(A;Z)$ (resp. $\mathrm{Var}\,\mathbf{E}[A \mid Z]$) is small, corresponding to the lower right corner of Figure 1. By the previous discussion, this is equivalent to finding representations Z that maximize accuracy (i.e. $\mathbf{E}_D[\ell(f(Z), Y)]$) while simultaneously maximizing invariance (i.e. $\mathbf{E}_D[\ell(f'(Z), A)]$). More precisely, consider the four vertices (not necessarily part of the feasible region) in Figure 1. These four corners have intuitive interpretations:

(Red; Bottom Left) The so-called "informationless" regime, in which all of the information regarding both Y and A is destroyed. This is trivially achieved by choosing a constant representation $Z \equiv c$.

(Yellow; Top Left) Here, we retain all of the information in A while removing all of the information about Y. This is not a particularly interesting regime for the aforementioned applications.

(Blue; Top Right) The full information regime, where $Z = X$ and no information is lost. This is the "classical" setting, wherein information about A is allowed to leak into Y.

(Green; Bottom Right) This is the ideal representation that we would like to attain: we preserve all the relevant information about Y while simultaneously removing all the information about A.

Unfortunately, in general, the ideal representation above (the green corner) may not be attainable due to potential dependency between Y and A. That is, there can never exist a representation $Z = g(X)$ that will have invariance and accuracy quality measures corresponding to the green corner. We are thus interested in characterizing how "close" we can get to attaining this ideal transformation via any possible representation $Z = g(X)$. Towards this, it is useful to consider the following extreme points on the boundary of the feasible region:

E_Y: This point corresponds to a representation Z that maximizes accuracy subject to a hard constraint on the invariance that there be no leakage of information about A into the representation Z. In classification, we enforce this via the mutual information constraint $I(A;Z) = 0$ and in regression via the conditional variance constraint $\mathrm{Var}\,\mathbf{E}[A \mid Z] = 0$.

E_A: This point corresponds to a representation Z that maximizes invariance subject to a hard constraint on the accuracy that there be no loss of information about Y in the representation Z. In classification, we enforce this via the mutual information constraint $I(Y;Z) = H(Y)$ and in regression we enforce this via the conditional variance constraint $\mathrm{Var}\,\mathbf{E}[Y \mid Z] = \mathrm{Var}(Y)$.

Now that we have formalized the feasible region, we next geometrically characterize this region, and analytically find the solutions to the extremal points of the region. This characterization will imply a variety of upper and lower bounds on constrained accuracy and invariance, as discussed in Sections 4 and 5.

4. Classification

We focus on the case where the original input X contains sufficient information to predict both Y and A, so that any loss of accuracy or invariance from a representation $Z = g(X)$ is not due to the non-informative input X, but due to the representation itself. This allows us to delineate the effects of the tradeoffs between the two competing goals of accuracy and invariance. Our analysis thus follows the classical setup of Probably Approximately Correct (PAC) learning framework (Valiant, 1984):

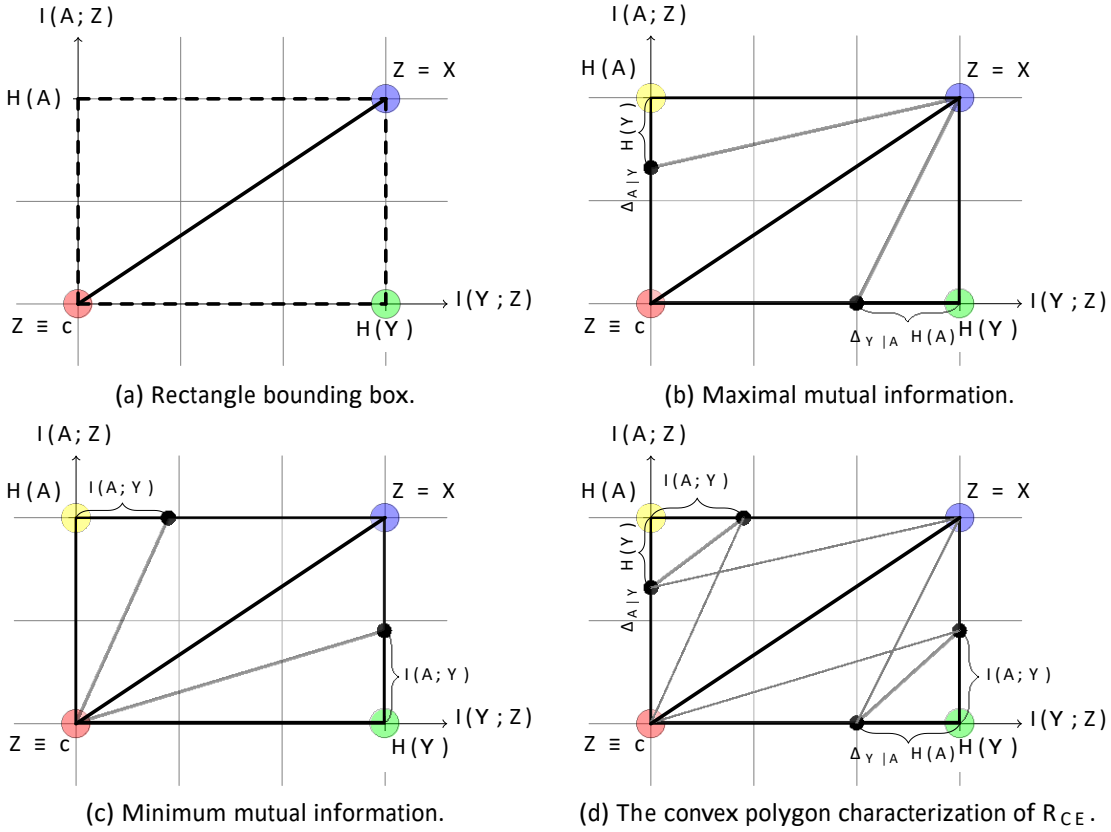
Assumption 4.1 There exist functions $f_Y()$ and $f_A()$, such that $Y = f_Y(X)$ and $A = f_A(X)$.

The perfect predictors f_Y and f_A are also known as concepts in the PAC learning framework. Alternatively, this assumption corresponds to the noiseless setting with Bayes optimal classification error being 0.[1] In order to characterize the feasible region R_{CE}, first note that due to the data processing inequality, the following inequalities hold:

$$0 \quad I(Y;Z) \quad I(Y;X) = H(Y), \quad 0 \quad I(A;Z) \quad I(A;X) = H(A).$$

Thus given any Z, the point $(I(Y;Z), I(A;Z))$ must lie within a rectangle shown in Figure 2a. The following proposition shows that the feasible region R_{CE} is convex:

1. This assumption is not necessary; see Appendix C. Essentially, accounting for the Bayes error rate complicates the analysis but does not fundamentally change the conclusions.



(a) Rectangle bounding box.

(b) Maximal mutual information.

(c) Minimum mutual information.

(d) The convex polygon characterization of R_{CE}.

Figure 2: Information plane in classification. Shaded area corresponds to the known feasible region.

Proposition 4.1 The feasible region R_{CE} for the classification setting is convex.

The convexity of R_{CE} can be shown by constructing randomized feature transformations to interpolate between any two given representations. Given two feature transformations $Z_0 = g_0(X)$, $Z_1 = g_1(X)$, we can define a uniform random variable $S \sim \text{Bernoulli}(u)$, for some constant $u \in [0,1]$, such that S is independent of Z_0 and Z_1. Now consider the randomized feature $Z = SZ_0 + (1 - S)Z_1$, i.e.,

$$Z = \begin{cases} Z_0 & \text{If } S \leq u, \\ Z_1 & \text{otherwise.} \end{cases} \tag{3}$$

It can then be seen that the invariance and accuracy tuple corresponding to Z is a convex combination of the two points in R_{CE} corresponding to Z_0 and Z_1, where the constant $u \in [0,1]$ specifies the convex combination weights. From an algorithmic perspective, this construction also suggests a simple way to achieve a new accuracy-invariance trade-off given existing representations by simply randomizing between them. This geometric nature, convexity, of the feasible region has natural consequences for feasible tradeoffs achievable by some representation. For instance, this entails that all the points on the diagonal of the bounding rectangle are also attainable by some feature representation.

In the next two sections, we characterize this feasible region in further detail.

4.1 E_Y: Maximal Mutual Information under the Independence Constraint

In this section we explore the first extreme point E_Y (Figure 1), which corresponds to maximizing the mutual information of Z w.r.t. Y while simultaneously being independent of A:

$$E_Y := \underset{Z : I(A;Z)=0}{\text{maximize}}\ I(Y;Z)\ . \tag{4}$$

First of all, realize that the optimal solution of (4) clearly depends on the coupling between A and Y. To see this, consider the following two extreme cases:

Example 4.1 If $A = Y$ almost surely, then $I(A;Z) = 0$ directly implies $I(Y;Z) = 0$, hence $\max_Z I(Y;Z) = 0$ under the constraint that $I(A;Z) = 0$.

Example 4.2 If $A \perp Y$, then $Z = f_Y(X) = Y$ satisfies the constraint that $I(A;Z) = I(A;Y) = 0$. Furthermore, $I(Y;Z) = I(Y;Y) = H(Y) \geq I(Y;Z')$, $\forall Z'$ s.t. $Z' = Y$. Hence $\max_Z I(Y;Z) = H(Y)$.

These two examples show that the optimal solution of (4) must involve a quantity that characterizes the dependency between A and Y. This motivates the following definition:

Definition 4.1 Define $D_{Y|A} := |\Pr_D(Y = 1 \mid A = 0) - \Pr_D(Y = 1 \mid A = 1)|$.

It is easy to verify the following claims: $0 \leq D_{Y|A} \leq 1$ and

$$\begin{aligned} D_{Y|A} = 0 &\iff A \perp Y, \\ D_{Y|A} = 1 &\iff A = Y \text{ or } A = 1 - Y. \end{aligned} \tag{5}$$

Armed with this notation, we can now analyze the solution to (4).

Theorem 4.1 The optimal value E_Y of the program in Eqn. (4) i.e. the maximal mutual information under the independence constraint is given as:

$$\max_{Z : I(A;Z)=0} I(Y;Z) = H(Y) - D_{Y|A}\, H(A). \tag{6}$$

As a sanity check of this result, note that if $A \perp Y$, then $D_{Y|A} = 0$, and in this case the optimal value given by Theorem 4.1 reduces to $H(Y) - 0 \cdot H(A) = H(Y)$, which is consistent with Example 4.2. Next, consider the other extreme case where $A = Y$. In this case $D_{Y|A} = 1$ and $H(Y) = H(A)$, therefore the optimal solution given by Theorem 4.1 becomes $H(Y) - 1 \cdot H(A) = H(Y) - H(Y) = 0$. This is consistent with Example 4.1.

Furthermore, due to the symmetry between A and Y, we can now characterize the locations of the two extremal points on the lower and left boundaries (Figure 2b). In Figure 2b, $D_{A|Y}$ is defined analogously as $D_{Y|A}$ by swapping Y and A.

4.2 E_A: Minimal Mutual Information under the Sufficient Statistics Constraint

Next, we characterize E_A in Figure 1. As before, it suffices to solve the following optimization problem, whose optimal solution is E_A.

$$E_A := \underset{Z : I(Y;Z) = H(Y)}{\text{minimize}} \; I(A;Z) \; . \tag{7}$$

Theorem 4.2 The optimal value E_A of the program in Eqn. (7) i.e. the Minimum Mutual Information under the Sufficient Statistics Constraint is

$$\min_{Z : I(Y;Z) = H(Y)} I(A;Z) = I(A;Y). \tag{8}$$

Clearly, if A and Y are independent, then the gap $I(A;Y) = 0$, meaning that we can simultaneously preserve all the target related information and filter out all the information related to A. We can now characterize the locations of the remaining two extremal points on the top and right boundaries of bounding rectangle (Figure 2c).

4.3 Application: Bounds on the Classification Loss

One application of the optimal solutions derived in Theorems 4.1 and 4.2 is to derive bounds on the risk functions:

Corollary 4.1 Let $Z = g(X)$ be any representation. If $I(Z;A) = 0$ then

$$\inf_h \mathbf{E}[\ell(Y, h(Z))] \geq D_{Y|A} - H(A).$$

If $I(Y;Z) = H(Y)$, then

$$\inf_h \mathbf{E}[\ell(A, h(Z))] \geq H(A|Y).$$

In the context of learning fair representations where A corresponds to a protected attribute, the first inequality implies that any fair classifier has to incur a population cross-entropy error of at least $D_{Y|A} - H(A)$, even if there exists a perfect classifier f_Y over the input X.[2] Furthermore, the term $H(A)$ also takes into account the population ratio between different demographic subgroups, in the sense that the more balanced the subgroups in the population, the larger the error lower bound. At first glance this may seem counter-intuitive, but the key is to observe that if the subpopulation proportions are unbalanced, then a classifier could try to achieve invariance by increasing the classification error on the smaller subpopulation. In the limit where one subpopulation has probability zero, then there is no loss in classification error.

In the context of privacy, if we interpret A as a sensitive attribute, and recall that $H(A|Y) = \inf_{h^0} \mathbf{E}[\ell(A, h^0(Y))]$ (see Appendix A for more details on this variational form of the conditional entropy), then the second inequality shows that, based on such representation Z, the worst case adversary could steal more sensitive information about A by looking at the feature Z instead of the target variable Y.

2. If Assumption 4.1 is violated, the lower bound will involve the Bayes error rate as well. In either case, $D_{Y|A} - H(A)$ represents an additional cost that is incurred by any learning algorithm.

4.4 Application: Certifying Sub-optimality

For the aforementioned applications in Section 2.1, our characterization of the feasible region is critical in order to be able to certify that a given model is not optimal. For example, given some practically computed representation Z and by using known estimators of the mutual information, it is possible to estimate $I(A;Y)$, $I(Y;Z)$, $I(A;Z)$ in order to directly bound the (sub)-optimality of Z using Figure 2d. In particular, we could compute how far away the point $(I(Y;Z), I(A;Z))$ is from the line segment between E_Y and E_A. This distance lower bounds the distance to the optimal representations on the Pareto frontier.

More concretely, the following corollary illustrates one example of applying these ideas. We say that Z is suboptimal if there exists a representation Z^0 such that $I(Y;Z^0) > I(Y;Z)$ and $I(A;Z^0) = I(A;Z)$ (resp. $I(A;Z^0) < I(A;Z)$ and $I(Y;Z^0) = I(Y;Z)$), i.e. Z^0 contains strictly more information about Y without leaking additional information about A (resp. Z^0 contains strictly less information about A without losing any information about Y). In other words, Z^0 achieves a better accuracy-invariance tradeoff.

Corollary 4.2 Suppose that the representation Z satisfies

$$\frac{I(A;Z)}{I(A;Y)} + \frac{H(Y|Z)}{D_{Y|A}\, H(A)} > 1. \tag{9}$$

Then Z is suboptimal.

Geometrically, (9) states that if $(I(Y;Z), I(A;Z))$ lies in the strict interior of the feasible region in Figure 2d, then the corresponding representation Z is suboptimal. Since each of the quantities in (9) is estimable, this provides a practical certificate of suboptimality for a learned representation Z.

In practice, while we do not have access to the population quantities like $I(A,Z)$, we can estimate them using finite samples. The key idea is to use entropy estimators in a plug-in fashion. For example, due to the identity $I(A,Z) = H(A) + H(Z) - H(A,Z)$, any consistent entropy estimator $\hat{H}(A)$, $\hat{H}(Z)$, $\hat{H}(A,Z)$ can be turned into a consistent estimator of $I(A,Z)$. The associated estimators and rates are standard, see the work of (Gao et al., 2017; Ross, 2014; Wu and Yang, 2016). Besides, $D_{Y|A}$ can also be easily estimated via relative frequencies (i.e. empirical probabilities) for discrete data. Therefore, with any existing entropy estimator, one can construct an estimator of the quantity in Eq. (9) with finite sample guarantee. For concreteness, here is an example where one can use (Wu and Yang, 2016) when Z is discrete with alphabet size at most k:

Proposition 4.2 When Z is discrete with alphabet size at most k, with the plug-in estimator (Wu and Yang, 2016) for mutual information applied to the left-hand side of (9) on a sample of size n from the joint distribution, if

$$\frac{\hat{I}(A;Z)}{\hat{I}(A;Y)} + \frac{\hat{H}(Y|Z)}{\hat{D}_{Y|A}\, \hat{H}(A)} > 1 + \#, \tag{10}$$

then the representation Z is certifiably suboptimal with probability at least $1 - \exp(-\Omega(n\#^2))$.

Furthermore, if $D_{Y|A}$, $H(Y)$, $I(A;Y)$ are all lower bounded by some positive constant c, the estimator converges at a rate of $1/\sqrt{n}$. Finite sample versions of other results in the paper follow by similar arguments, as well.

4.5 Application: Implications for Learning Invariant Representations in Domain Generalization

One potential application of particular interest of Theorem 4.1 is domain generalization. Domain generalization refers to the problem where we aim to train a model on data from a set of source domains so that the learned model can generalize to unseen target domains. Under this context, the protected attribute A could be understood to be the index of the training domains, where it further specifies from which domain/group the labeled data comes from.

A line of recent works have proposed to learn domain-invariant representations (Albuquerque et al., 2019; Deng et al., 2020; Nguyen et al., 2021; Dong et al., 2022) for domain generalization, where intuitively the goal is to learn representations Z from which one cannot infer the domain index A. Clearly, in this case if the domain index A also contains discriminative information about the target attribute Y, there is a potential loss of accuracy by using domain-invariant representations for domain generalization. More precisely, applying Theorem 4.1 to this context, we have that for domain-invariant representations Z ($I(A;Z) = 0$), the weighted error of any classifier h over Z has the following error lower bound:

$$\inf_h \mathbf{E}[\ell(Y, h(Z))] = \inf_h \sum_a \Pr(A = a)\, \mathbf{E}[\ell(Y, h(Z)) \mid A = a] \geq D_{Y|A}\, H(A). \tag{11}$$

Note that although we present our theory for binary classification, the implications here we obtained for domain generalization also hold more generally for multi-class classification problems with more than two source domains. This kind of impossibility result for domain generalization is similar in nature to the one in domain adaptation (Zhao et al., 2019a, Theorem 4.3), where the main difference lies in that Zhao et al. (2019a, Theorem 4.3) considers a sum of balanced errors of the source and target domains whereas the one in (11) is concerned with weighted (given by the size of each source domain) errors among the source domains.

5. Regression

We now present analogous results in the regression setting, where entropy is replaced by conditional variance. To simplify the presentation, we assume a similar noiseless setting:

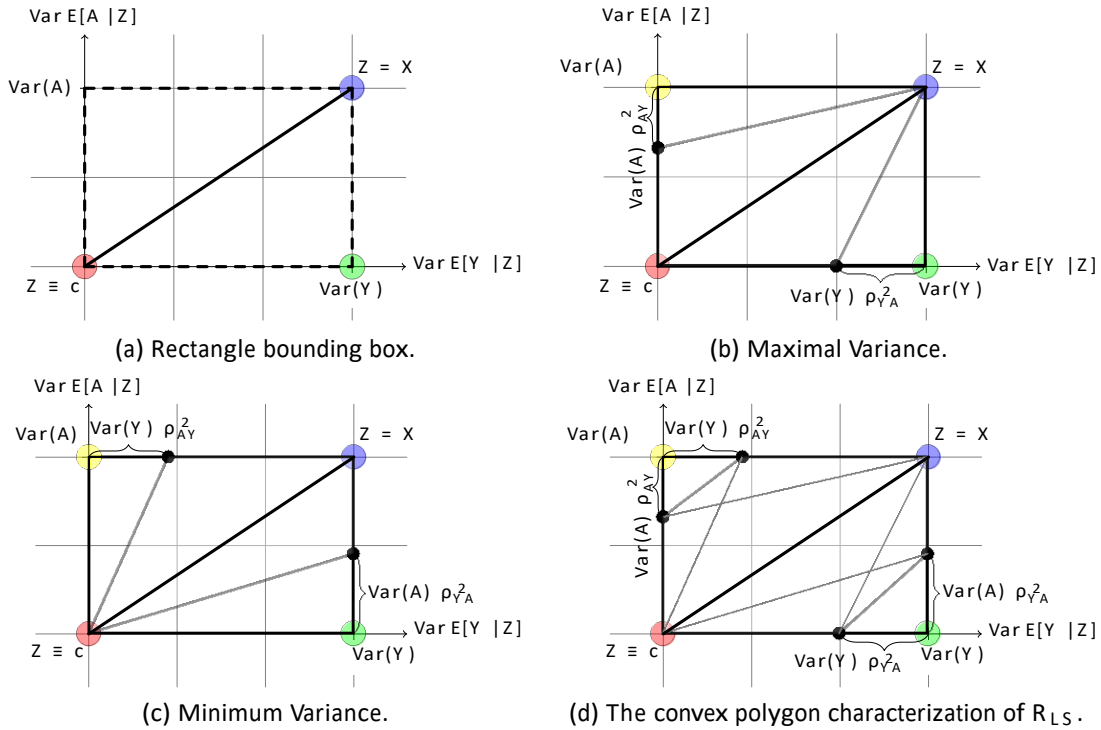
Assumption 5.1 There exist functions $f_Y, f_A \in \mathbf{H}$, such that $Y = f_Y(X)$ and $A = f_A(X)$, where $\mathbf{H}$ is a given reproducing kernel Hilbert space (RKHS).

This assumption is not necessary; the results presented here are special cases of more general results for the noisy setting presented in Appendix C.

Let $\langle \cdot, \cdot \rangle$ be the canonical inner product in RKHS $\mathbf{H}$. Under this assumption, there exists a feature map $\phi(X)$ and $a = 0, y = 0$, such that $Y = f_Y(X) = \langle \phi(X), y \rangle$ and $A = f_A(X) = \langle \phi(X), a \rangle$. This feature map does not have to be finite-dimensional; our analysis applies to infinite-dimensional feature maps as well.

Similar to our analysis in the classification setting, by the law of total variance, the following inequalities hold:

$$0 \leq \mathrm{Var}\,\mathbf{E}[Y \mid Z] \leq \mathrm{Var}\,\mathbf{E}[Y \mid X] = \mathrm{Var}(Y), \leq 0$$
$$\mathrm{Var}\,\mathbf{E}[A \mid Z] \leq \mathrm{Var}\,\mathbf{E}[A \mid X] = \mathrm{Var}(A),$$



(a) Rectangle bounding box.

(b) Maximal Variance.

(c) Minimum Variance.

(d) The convex polygon characterization of R_{LS}.

Figure 3: Information plane in regression. Shaded area corresponds to the known feasible region.

which means that for any transformation $Z = g(X)$, the point $(\mathrm{Var}[Y \mid Z], \mathrm{Var}[A \mid Z])$ must lie within a rectangle shown in Figure 3a. To simplify the notation, we define $S := \mathrm{Cov}(f(X), f(X))$ to be the covariance operator.

If we consider all the possible feature transformations $Z = g(X)$, then the points $(\mathrm{Var}[Y \mid Z], \mathrm{Var}[A \mid Z])$ will form a feasible region R_{LS}. The following lemma shows that the feasible region R_{LS} is convex:

Lemma 5.1 R_{LS} is convex.

Again, the convexity of R_{LS} is guaranteed by a construction of randomized feature transformation. One direct implication of the convexity is that we could interpolate between the performance tuples of existing representation learning methods in the regression setting by allowing randomized representations, as we discussed in the classification setting. Also, both the "informationless" origin and the "full information" diagonal vertex are attainable. Hence, all the points on the diagonal of the bounding rectangle are also attainable.

5.1 E_Y: Maximal Variance under the Invariance Constraint

In this section we explore the first extreme point E_Y (Figure 1) in the regression setting, which corresponds to maximizing the variance of Z w.r.t. Y while simultaneously minimizing that of A:

$$E_Y := \underset{Z \,:\, \mathrm{Var}[A \mid Z]=0}{\text{maximize}} \ \mathrm{Var}[Y \mid Z] \,. \tag{12}$$

It is clear that the optimal solution of (12) depends on the coupling between A and Y, and the following theorem precisely characterizes this relationship:

Theorem 5.1 The optimal solution of optimization problem (12) is upper bounded by

$$\max_{Z, \mathrm{Var}\,\mathbb{E}[A \mid Z]=0} \mathrm{Var}\,\mathbb{E}[Y \mid Z] \;\; \mathrm{Var}(Y) \;\; \frac{\mathrm{Cov}(A,Y)^2}{\mathrm{Var}(A)} = \mathrm{Var}(Y)(1 \;\; r_{YA}^2),$$

where r_{YA} is the correlation coefficient of Y and A.

Let us do a sanity check of this result: If A and Y are uncorrelated, then $r_{YA} = 0$ and hence the gap is 0, and the optimal solution becomes $\mathrm{Var}(Y)$. Next, consider the other extreme case where A is perfectly correlated with A: In this case $r_{YA} = 1$ so the optimal solution reduces to 0. For readers who are familiar with regression analysis, it can be readily observed that the gap, i.e., r_{YA}^2 $\mathrm{Var}(Y)$, due to the constraint $\mathrm{Var}\,\mathbb{E}[A \mid Z] = 0$ essentially corresponds to the explained variance of Y by performing an optimal linear regression from A, since r_{YA}^2 equals the R-square of linearly regressing from A to Y.

With Lemma 5.1 and Theorem 5.1, we can now characterize the locations of the two extremal points on the bottom and left boundaries of bounding rectangle in Figure 3b.

5.2 E_A: Minimal Variance under the Sufficient Statistics Constraint

Next, we characterize E_A in the right panel of Figure 1. As before, it suffices to solve the following optimization problem, whose optimal solution is E_A.

$$E_A := \underset{Z \,:\, \mathrm{Var}\,\mathbb{E}[Y \mid Z]=\mathrm{Var}(Y)}{\text{minimize}} \mathrm{Var}\,\mathbb{E}[A \mid Z]. \tag{13}$$

Theorem 5.2 The optimal solution of optimization problem (13) is lower bounded by

$$\min_{Z, \mathrm{Var}\,\mathbb{E}[Y \mid Z]=\mathrm{Var}(Y)} \mathrm{Var}\,\mathbb{E}[A \mid Z] \;\; \frac{\mathrm{cov}(A, y_i)^2}{\mathrm{var}(y, y_i)} = \mathrm{Var}(A)\; r_{YA}^2, \tag{14}$$

where r_{YA} is the correlation coefficient of Y and A.

Again, if A is uncorrelated with Y, then the lower bound becomes 0, meaning that we can simultaneously preserve all the target variance and filter out all the variance related to A. This is the no tradeoff regime. On the other hand, if A is perfectly correlated with Y, then $r_{YA} = 1$, hence the lower bound becomes exactly $\mathrm{Var}(A)$. This extreme case shows that if we would like to preserve all the variation w.r.t. the target Y, we also have to retain all the information about A as well. This is expected since, again, the lower bound corresponds to the explained variance of A from Y due to the constraint $\mathrm{Var}\,\mathbb{E}[Y \mid Z] = \mathrm{Var}(Y)$.

By symmetry, the updated plot is shown in Figure 3c. We plot the full picture of R_{LS} in Figure 3d. Both the constrained accuracy optimal solution and the constrained invariance optimal solution can be readily read from Figure 3d as well. Finally, as in the discrete setting, the characterization of the feasible region and the extremal points E_Y and E_A have some important consequences, e.g., certifying the suboptimality of a given representation. Here, we note the analogues to Corollary 4.1 for the regression setting:

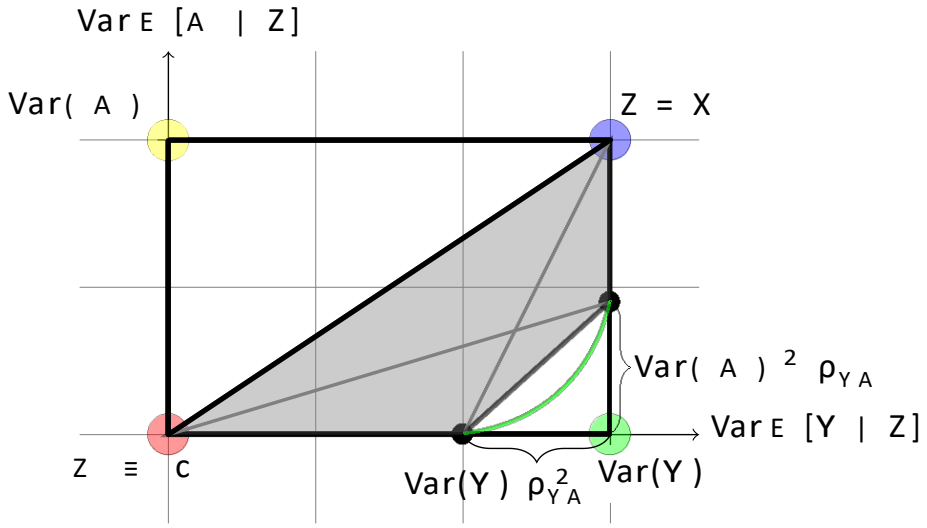


Figure 4: The exact characterization of R_{LS} in the regression setting. The green convex curve connecting the two black dots is the frontier between $\mathrm{Var}\,\mathbf{E}[Y|Z]$ and $\mathrm{Var}\,\mathbf{E}[A|Z]$, given by Theorem 5.3.

Corollary 5.1 Let $Z = g(X)$ be any representation. If $\mathrm{Var}\,\mathbf{E}[A \mid Z] = 0$ then

$$\inf_h \mathbf{E}[\ell(Y, h(Z))] \geq \mathrm{Var}(Y)\, r_{YA}^2. \tag{15}$$

Furthermore, if $\mathrm{Var}\,\mathbf{E}[Y \mid Z] = \mathrm{Var}\,Y$, then

$$\inf_h \mathbf{E}[\ell(A, h(Z))] \geq \mathrm{Var}(A)(1 - r_{YA}^2). \tag{16}$$

A similar analogue to Corollary 4.2 can be derived in order to certify suboptimality.

5.3 An Exact Characterization of the Frontier

In the regression setting, we can say even more: we can derive a tight upper bound to the constrained problem

$$\begin{aligned} \underset{Z}{\text{maximize}} \quad & \mathrm{Var}\,\mathbf{E}[Y \mid Z] \\ \text{subject to} \quad & \mathrm{Var}\,\mathbf{E}[A \mid Z] \leq c. \end{aligned} \tag{17}$$

Theorem 5.3 The optimal solution of the constrained problem (17) has the following upper bound: for any $c \in [0, r_{YA}^2\,\mathrm{Var}(A)]$, we have

$$\mathrm{Var}\,\mathbf{E}[Y \mid Z] \leq \mathrm{Var}(Y) - 2r_{YA}\sqrt{(1 - r_{YA}^2)a(1-a)} + 1 - a - r_{YA}^2 + 2ar_{YA}^2 \tag{18}$$

where $a := c/\mathrm{Var}(A)$.

When this bound is tight (see Section 5.4), this fully characterizes the feasible region R_{LS}. Before we move to the discussion of the tightness of these bounds, we pause to make a few comments on the relationship between Theorem 5.3 and previous results.

A notable special case is when $c = 0$, where the upper bound simplifies to:

$$\mathrm{Var}\,\mathbf{E}[Y|Z] \;\; \mathrm{Var}(Y)(1 \quad r^2_{YA}), \tag{19}$$

which reduces exactly to the bound in Theorem 5.1.

We also note that Theorem 5.2 is also closely related: Theorem 5.2 states that to achieve perfect utility (i.e. $\mathrm{Var}\,\mathbf{E}[Y|Z] = \mathrm{Var}(Y)$), one must have $\mathrm{Var}\,\mathbf{E}[A|Z] \;\; r^2_{YA}\mathrm{Var}(A)$. Setting $c = r^2_{YA}\mathrm{Var}(A)$ in Theorem 5.3, the upper bound exactly simplifies to $\mathrm{Var}(Y)$, the perfect utility. Therefore, Theorem 5.3 interpolates between the two extreme cases (Theorem 5.1 and 5.2). Furthermore, one can verify that the upper bound in Theorem 5.3 is a concave function of a, which means that it is a convex curve on the information plane. In fact, it exactly characterizes the frontier between $\mathrm{Var}\,\mathbf{E}[Y|Z]$ and $\mathrm{Var}\,\mathbf{E}[A|Z]$ in the regression setting. We visualize this result in Figure 4, where the convex curve connecting the two black dots (the two optimal solutions in the extreme cases) corresponds to the bound in Theorem 5.3.

The proof is based on a careful characterization of the Lagrangian form of the constrained problem. Define $\mathrm{OPT}(\lambda)$ to be the Lagrangian:

$$\mathrm{OPT}(\lambda) := \min_{Z} \; \lambda \,\mathrm{Var}\,\mathbf{E}[A|Z] \quad \mathrm{Var}\,\mathbf{E}[Y|Z] \tag{20}$$

By Lagrange duality, $\mathrm{OPT}(\lambda)$ provides an explicit upper bound on the optimal solution to the following problem, which is a relaxed version of the problem (12). More specifically, for any $c > 0$,

$$\mathrm{Var}\,\mathbf{E}[Y|Z] \quad \sup_{\lambda \geq 0}\{\mathrm{OPT}(\lambda) \quad \lambda c\}. \tag{21}$$

Therefore, the remaining work is to provide a lower bound of $\mathrm{OPT}(\lambda)$, for which we establish the following result:

Theorem 5.4 The optimal solution of the Lagrangian in regression has the following lower bound:

$$\mathrm{OPT}(\lambda) \;\; \frac{1}{2}\Big[\lambda\,\mathrm{Var}(A) \quad \mathrm{Var}(Y) \quad \sqrt{\mathrm{Var}^2(Y) + \lambda^2\,\mathrm{Var}^2(A) \quad 2\lambda\,\mathrm{Var}(A)\,\mathrm{Var}(Y)(2r_{YA}{}^2 \quad 1)}\Big]. \tag{22}$$

A few remarks are in order. The key quantity in the lower bound (22) is the correlation coefficient between A and Y: r^2_{YA}, which effectively measures the dependence between the Y and A under the feature covariance S. To see the connection between Theorem 5.4 and Theorem 5.2, Theorem 5.1, note that $\lambda = \infty$ corresponds to the hard constraint that $\mathrm{Var}\,\mathbf{E}[A|Z] = 0$, and $\lambda \;\; 0$ corresponds to the hard constraint that $\mathrm{Var}\,\mathbf{E}[Y|Z] = \mathrm{Var}(Y)$. Of course, $\lambda \in (0,\infty)$ corresponds to the soft constraint that $\mathrm{Var}\,\mathbf{E}[A|Z] \;\; c$.

5.4 When are These Bounds Tight?

The proofs of Theorem 5.1, Theorem 5.2 and Theorem 5.4 all rely on a semidefinite programming (SDP) relaxation, and construct an explicit optimal solution to this relaxation. At a high level, we

re-formulate the objective as a linear functional of $V := \mathrm{Var}\,\mathbb{E}[f(X) \mid Z]$, which satisfies the semi-definite constraint $0 \preceq V \preceq S = \mathrm{Var}\, f(X)$. Therefore, the optimal value of the SDP is an upper/lower bound of the corresponding objective. In this section we shall discuss when these SDP relaxations are tight. In particular, we show that under the following regularity condition on (X, f), the SDP relaxation is exact.

Definition 5.1 (X, f) is called regular, if for any positive semidefinite matrix $M: 0 \preceq M \preceq S$, there exists (possibly randomized) $Z = g(X)$, such that $\mathrm{Var}\,\mathbb{E}[f(X) \mid Z] = M$.

One particularly interesting setting where the regularity condition holds is when $f(X)$ follows a Gaussian distribution.

Theorem 5.5 (X, f) is regular if $f(X)$ follows a Gaussian distribution.

The proof of Theorem 5.5, again, depends on an ingenious construction of randomized features. We defer the proof of this theorem to Section C.7. Roughly speaking, we first show that under the Gaussian assumption, all the rank-1 matrices $0 \preceq M \preceq S$ could be attained via deterministic linear transformations. We then show that the rest of such matrices M could be attained by constructing randomized features, which corresponds to a convex combination of all the rank-1 matrices.

Under the regularity condition, we can show all the bounds in Theorem 5.1, Theorem 5.2 and Theorem 5.4 are tight.

Theorem 5.6 When (X, f) is regular, the upper bound in Theorem 5.1, the lower bound in Theorem 5.2 and the spectral lower bound in Theorem 5.4 are all achievable.

Combining Theorem 5.5 and Theorem 5.6, we now know that if the distribution over $f(X)$ in the RKHS is Gaussian, then all of the bounds in Theorem 5.1, Theorem 5.2 and Theorem 5.4 become tight. Although here we only show that the regularity condition holds under Gaussian assumption, we conjecture that it could be verified under broader settings. More discussions about the tightness of our bounds and the analysis are presented in Section C.7 of the appendix.

6. Numerical Experiments

In this section, we demonstrate the empirical application of our theoretical results in both classification and regression tasks to certify the suboptimality of certain representation learning algorithms. To this end, we conduct experiments on two real-world benchmark datasets, the UCI Adult dataset (Asuncion and Newman, 2007) for classification and the Law School dataset (Wightman, 1998) for regression. In what follows we first provide a brief description about these two datasets and the corresponding tasks.

6.1 Datasets

UCI Adult The Adult dataset contains 30,162/15,060 training/test instances for income prediction. Each instance in the dataset describes an adult from the 1994 US Census. Attributes of each individual include gender, education level, age, etc. In this experiment we use gender (binary) as the protected attribute, and we preprocess the dataset to convert all the categorical variables into corresponding one-hot representations. The processed data contains 114 attributes. The target variable (income) is

also binary: 1 if 50K/year otherwise 0. For the protected attribute A, A = 0 means Male otherwise Female. In this dataset, the base rates across groups are different: Pr(Y = 1 j A = 0) = 0.310 while Pr(Y = 1 j A = 1) = 0.113. The group ratio is also quite imbalanced, with Pr(A = 0) = 0.673 and Pr(A = 1) = 0.327.

Law School The Law School dataset contains 1,823 records for law students who took the bar passage study for Law School Admission.[3] The features in the dataset include variables such as undergraduate GPA, LSAT score, full-time status, family income, gender, etc. In this experiment, we use gender (treated as a continuous variable that takes value in [0, 1]) as the protected attribute and undergraduate GPA (continuous) as the target variable. For both variables, we use the mean-squared error as the loss function. We use 80 percent of the data as our training set and the rest 20 percent as the test set. In the Law School dataset, Pr(A = 1) = 0.452, which is quite balanced. The data distribution for different subgroups in the Law School dataset could be found in Figure 5. From Figure 5, we can see that the conditional distributions Pr(Y j A = a) are different across different subgroups A = a 2 f0, 1g.

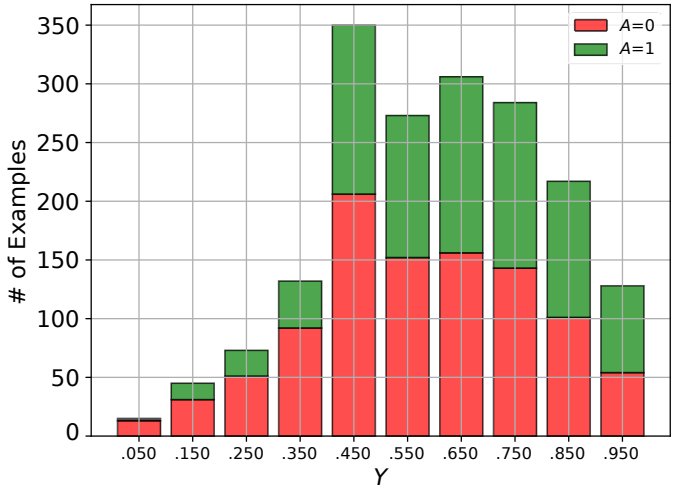


Figure 5: The data distributions of different subgroups in the Law School dataset.

6.2 Representation Learning Algorithms

To verify that the empirical estimates of the vertices in Figure 2d and Figure 3d could be used to certify the suboptimality of certain representation learning algorithms, in this section we apply four different representation learning algorithms to the Adult and Law School datasets and report their corresponding metrics to have a comparison among them.

Logit Score (Logit). The very basic baseline representation learning method we consider is logistic regression. In particular, we first fit a logistic regression model on the Adult dataset and use the logit score of the input data as its representations. Similarly, for the Law School

3. We use the edited public version of the dataset which can be downloaded here: https://github.com/algowatchpenn/GerryFair/blob/master/dataset/lawschool.csv

dataset, we consider the Ordinary Least Squares (OLS) for the very basic baseline, where we use the 1-dimensional prediction given by a trained OLS model as the corresponding representations.

Linear Representations (Linear). One common representation that has been widely used is the linear representation. In this case, we consider $Z = WX + b$, where $W \in \mathbf{R}^{dp}$, $b \in \mathbf{R}^{d}$ are the model parameters to be learned.

Multilayer Perceptrons (MLP). The representations obtained by MLP are the hidden features of a ReLU network. On the Adult dataset, the network contains one hidden layer. For the Law School dataset, we use a three-hidden-layer network.

Adversarial Representation Learning (Adv). In Adv we use the same network architecture and training hyperparameters as the ones used in MLP, except that now we use an adversarial training method with Wasserstein regularization (Chi et al., 2021) to learn invariant representations.

On the Adult dataset, for Logit, LR and MLP, we use the target prediction loss, i.e., cross-entropy in classification and mean-squared error in regression, with stochastic gradient descent to optimize the model parameters. The model parameters (hence the representations) of Adv is obtained via adversarial training, by solving a minimax optimization problem. On the Law School dataset, for OLS, LR and MLP, we use the mean-squared error as the regression loss function and optimize model parameters with stochastic gradient descent. For Adv, again, we obtain the model parameters via adversarial training.

6.3 Results and Analysis

UCI Adult: Classification In order to certify the suboptimality and the corresponding distance to the accuracy-invariance frontier of different representation learning algorithms, we need to first estimate the vertices on the information plane as shown in Figure 2d. We do this by using the plug-in estimator. In particular, we compute the conditional probability $\Pr(Y = 1 \mid A = a)$ and marginal probabilities $\Pr(Y)$, $\Pr(A)$ from the sample, and then plug them into the formula of $H(Y)$, $H(A)$ and $D_{Y|A}$, respectively.

We report the results in Table 1. It can be readily verified that the mutual information pair $(I(Y;Z), I(A;Z))$ of all the four methods are consistent with our theoretical upper and lower bounds. Furthermore, none of these algorithms locates on the frontier of the information plane, hence they are all strictly suboptimal. To better understand the strengths of different algorithms, we also visualize the results in Figure 6. From Figure 6, we can see that both MLP and Adv strictly dominate Logit and Linear, which is expected since these two models are richer than the other two baselines. Note that while MLP achieves better $I(Y;Z)$ than Adv, Adv is significantly better in terms of ensuring a small $I(A;Z)$.

Law School: Regression Again, similar to the classification task, we first estimate the vertices on the information plane as shown in Figure 3d, which includes the empirical variances and the correlation coefficient between A and Y. The results for the Law School dataset are shown in Table 2. As expected, the results are, again, consistent with our theoretical upper and lower bounds on the information plane. But different from the case in the Adult dataset, on the Law School dataset no method is strictly dominating another, and all the four methods are quite far away from the theoretical

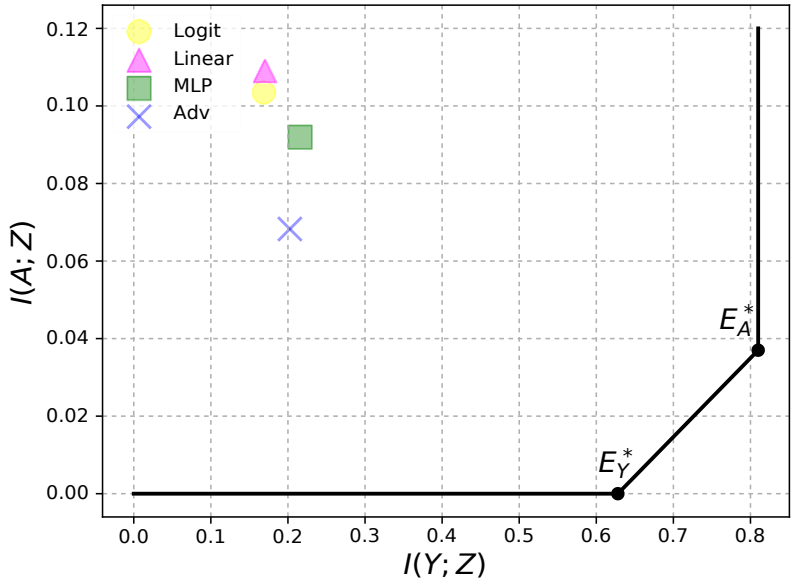


Figure 6: The information plane of Logit, Linear, MLP and Adv on the Adult dataset. On this plane, one method is strictly dominating another if it lies at the bottom-right side of the latter.

Table 1: Numerical results on the Adult dataset with Logit, Linear, MLP and Adv. For $I(Y;Z)$, the larger the better. For $I(A;Z)$, the smaller the better.

	Lower Bound	Logit	Linear	MLP	Adv	Upper Bound
$I(Y;Z)$	0	0.170	0.171	0.216	0.202	E_Y : 0.628
$I(A;Z)$	E_A : 0.037	0.103	0.109	0.092	0.068	$H(A)$: 0.909

frontier on the plane. We visualize the results in Figure 7. From Figure 7, we can conclude that both MLP and Adv are better than OLS and Linear at achieving a large Var $\mathbf{E}[Y \mid Z]$, but at the cost of increasing Var $\mathbf{E}[A \mid Z]$ simultaneously.

Table 2: Numerical results on the Law School dataset with OLS, Linear, MLP and Adv. For Var $\mathbf{E}[Y|Z]$, the larger the better. For Var $\mathbf{E}[A|Z]$, the smaller the better.

	Lower Bound	OLS	Linear	MLP	Adv	Upper Bound
Var $\mathbf{E}[Y|Z]$	0	0.014	0.013	0.022	0.018	E_Y : 0.040
Var $\mathbf{E}[A|Z]$	E_A : 0.007	0.008	0.008	0.012	0.010	Var(A) : 0.248

7. Conclusion

In this paper we provide an information plane analysis to study the general and important problem for learning invariant representations in both classification and regression settings. In both cases, we ana-

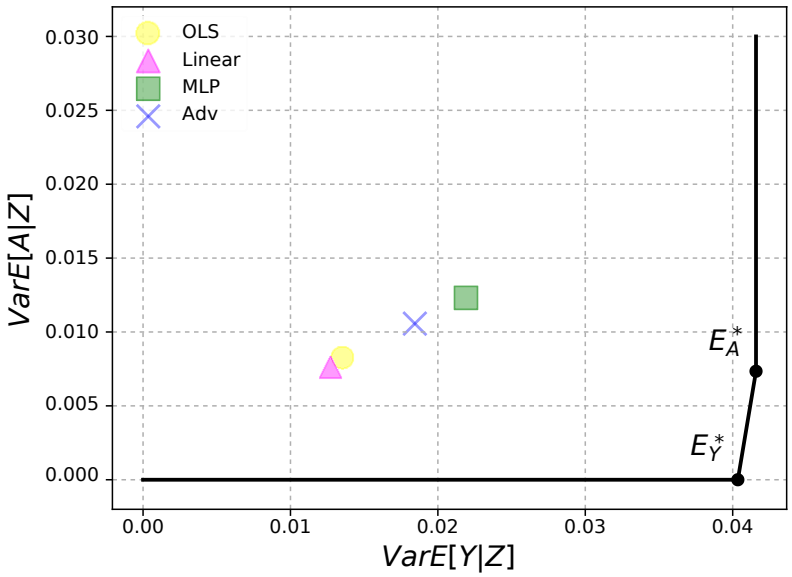


Figure 7: The information plane of OLS, Linear, MLP and Adv on the Law School dataset. On this plane, one method is strictly dominating another if it lies at the bottom-right side of the latter.

lyze the inherent tradeoffs between accuracy and invariance by providing a geometric characterization of the feasible region on the information plane, in terms of its boundedness, convexity, as well as its its extremal vertices. Furthermore, in the regression setting, we also derive a tight lower bound that for the Lagrangian form of accuracy and invariance, which leads to a complete characterization of the frontier between accuracy and invariance. Given the wide applications of invariant representations in machine learning, we believe our theoretical results could contribute to better understandings of the fundamental tradeoffs between accuracy and invariance under various settings, e.g., domain adaptation, algorithmic fairness, invariant visual representations, and privacy-preservation learning. Furthermore, our analysis on the construction of randomized features also provides a simple way to construct features that can interpolate between existing invariances and accuracies.

So far we mainly focus on the settings where the two attributes of interest are both discrete or continuous. One interesting direction for future work is to consider the mix of these two. For example, could we give similar analytic characterization on the optimal solution of the four constrained problems when one of them is discrete and the other is continuous?

Acknowledgments

HZ would like to thank Alexander Kozachinskiy for the discussion in proving Theorem 4.1. HZ and GG would like to acknowledge support from the DARPA XAI project, contract #FA87501720152 and a Nvidia GPU grant. HZ also thanks the support from a Facebook research award. CD and PR would like to acknowledge the support of NSF via IIS-1955532, IIS-1909816, OAC-1934584, and ONR via N000141812861. CD would like to thank Liu Leqi for many helpful discussions in the early stage of this project.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. arXiv preprint arXiv:1903.00089, 2019.

Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. arXiv preprint arXiv:1911.00804, 2019.

Fabio Anselmi, Joel Z Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations. Theoretical Computer Science, 633:112–121, 2016a.

Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. On invariance and selectivity in representation learning. Information and Inference: A Journal of the IMA, 5(2):134–158, 2016b.

Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In Advances in neural information processing systems, pages 137–144, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. Machine learning, 79(1-2):151–175, 2010.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020.

Jianfeng Chi, Yuan Tian, Geoffrey J Gordon, and Han Zhao. Understanding and mitigating accuracy disparity in regression. arXiv preprint arXiv:2102.12013, 2021.

Maximin Coavoux, Shashi Narayan, and Shay B Cohen. Privacy-preserving neural representations of text. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1–10, 2018.

Zhun Deng, Frances Ding, Cynthia Dwork, Rachel Hong, Giovanni Parmigiani, Prasad Patil, and Pragya Sur. Representation via representations: Domain generalization via adversarially learned invariant representations. arXiv preprint arXiv:2006.11478, 2020.

Jing Dong, Shiji Zhou, Baoxiang Wang, and Han Zhao. Algorithms and theory for supervised gradual domain adaptation. arXiv preprint arXiv:2204.11644, 2022.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R Varshney. An information-theoretic perspective on the relationship between fairness and accuracy. arXiv preprint arXiv:1910.07870, 2019.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. arXiv preprint arXiv:1511.05897, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. The Journal of Machine Learning Research, 17(1):2096–2030, 2016.

Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. In Proceedings of the 31st Neural Information Processing Systems, pages 5988–5999, 2017.

Jihun Hamm. Preserving privacy of continuous high-dimensional data with minimax filters. In Artificial Intelligence and Statistics, pages 324–332, 2015.

Jihun Hamm. Minimax filter: learning to preserve privacy from inference attacks. The Journal of Machine Learning Research, 18(1):4704–4734, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In International conference on machine learning, pages 3020–3029, 2016.

Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. arXiv preprint arXiv:2001.07426, 2020.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351, 2017.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. Advances in Neural Information Processing Systems, 33, 2020.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 624–639, 2018.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. arXiv preprint arXiv:1802.06309, 2018.

Stéphane Mallat. Group invariant scattering. Communications on Pure and Applied Mathematics, 65 (10):1331–1398, 2012.

Daniel McNamara, Cheng Soon Ong, and Robert C Williamson. Costs and benefits of fair representation learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 263–270, 2019.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In Conference on Fairness, Accountability and Transparency, pages 107–118, 2018.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In International Conference on Machine Learning, pages 10–18, 2013.

A Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with domain density transformations. Advances in Neural Information Processing Systems, 34:5264–5275, 2021.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. Nature, 435(7045):1102–1107, 2005.

Brian C Ross. Mutual information between discrete and continuous data sets. PloS one, 9(2):e87357, 2014.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 3076–3085. JMLR. org, 2017.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In 2015 IEEE Information Theory Workshop (ITW), pages 1–5. IEEE, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. arXiv preprint physics/0004057, 2000.

Leslie G Valiant. A theory of the learnable. Communications of the ACM, 27(11):1134–1142, 1984.

Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. IEEE Transactions on Information Theory, 62(6):3702–3720, 2016.

Taihong Xiao, Yi-Hsuan Tsai, Kihyuk Sohn, Manmohan Chandraker, and Ming-Hsuan Yang. Adversarial learning of privacy-preserving and task-oriented representations. arXiv preprint arXiv:1911.10143, 2019.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In International Conference on Machine Learning, pages 325–333, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pages 335–340, 2018.

Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. In Advances in neural information processing systems, pages 15675–15685, 2019.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In Advances in neural information processing systems, pages 8559–8570, 2018.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. On learning invariant representation for domain adaptation. arXiv preprint arXiv:1901.09453, 2019a.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. arXiv preprint arXiv:1910.07162, 2019b.

Han Zhao, Junjie Hu, and Andrej Risteski. On learning language-invariant representations for universal machine translation. arXiv preprint arXiv:2008.04510, 2020.

Appendix A. Proofs of Claims in Section 3

In this section we give detailed arguments to derive the objective functions of Eq. (1) and (2) respectively from the original minimax formulation. First, let us consider the classification setting.

Classification Given a fixed feature map $Z = g(X)$, due to the symmetry between Y and A, it suffices for us to consider the case of finding f that minimizes $\mathbf{E}_D[\ell(f \circ g(X), Y)]$, and analogous result follows for the case of finding the optimal f^0 that minimizes $\mathbf{E}_D[\ell(f^0 \circ g(X), A)]$ similarly. By definition of the cross-entropy loss, we have:

$$\begin{aligned}
\mathbf{E}_D[\ell(f \circ g(X), Y)] &= \mathbf{E}_D[\mathbf{I}(Y = 0)\log(1 - f(g(X))) + \mathbf{I}(Y = 1)\log(g(f(X)))] \\
&= \mathbf{E}_D[\mathbf{I}(Y = 0)\log(1 - f(Z)) + \mathbf{I}(Y = 1)\log(f(Z))] \\
&= \mathbf{E}_Z\mathbf{E}_Y[\mathbf{I}(Y = 0)\log(1 - f(Z)) + \mathbf{I}(Y = 1)\log(f(Z)) \mid Z] \\
&= \mathbf{E}_Z[\Pr(Y = 0 \mid Z)\log(1 - f(Z)) + \Pr(Y = 1 \mid Z)\log(f(Z))] \\
&= \mathbf{E}_Z[\mathrm{KL}(\Pr(Y \mid Z) \,\|\, f(Z))] + H(Y \mid Z) \\
&\quad - H(Y \mid Z).
\end{aligned}$$

It is also clear from the above proof that the minimum value of the cross-entropy loss is achieved when $f(Z)$ is a randomized classifier such that $\mathbf{E}[f(Z)] = \Pr(Y = 1 \mid Z)$. This shows that

$$\min_f \mathbf{E}_D[\ell(f \circ g(X), Y)] = H(Y \mid Z), \quad \text{and} \quad \min_{f^0} \mathbf{E}_D[\ell(f^0 \circ g(X), A)] = H(A \mid Z).$$

To see the second part of Eq. (1), simply use the identity that $H(Y \mid Z) = H(Y) - I(Y; Z)$ and $H(A \mid Z) = H(A) - I(A; Z)$ with the fact that both $H(Y)$ and $H(A)$ are constants that only depend on the joint distribution D.

Regression Again, given a fixed feature map $Z = g(X)$, because of the symmetry between Y and A let us focus on the analysis of finding f that minimizes $\mathbf{E}_D[\ell(f \circ g(X), Y)]$. In this case since $\ell(\cdot, \cdot)$ is the mean squared error, it follows that

$$\begin{aligned}
\mathbf{E}_D[\ell(f \circ g(X), Y)] &= \mathbf{E}_D(f \circ g(X) - Y)^2 = \\
&\quad \mathbf{E}_D(f(Z) - Y)^2 \\
&= \mathbf{E}_Z(f(Z) - \mathbf{E}[Y \mid Z])^2 + \mathbf{E}_Z\mathbf{E}_Y[(Y - \mathbf{E}[Y \mid Z])^2] \\
&\quad \mathbf{E}_Z\mathbf{E}_Y[(Y - \mathbf{E}[Y \mid Z])^2] \\
&= \mathbf{E}[\mathrm{Var}(Y \mid Z)],
\end{aligned}$$

where the third equality is due to the Pythagorean theorem. Furthermore, it is clear that the optimal mean-squared error is obtained by the conditional mean $f(Z) = \mathbf{E}[Y \mid Z]$. This shows that

$$\min_f \mathbf{E}_D[\ell(f \circ g(X), Y)] = \mathbf{E}[\mathrm{Var}(Y \mid Z)], \quad \text{and} \quad \min_{f^0} \mathbf{E}_D[\ell(f^0 \circ g(X), A)] = \mathbf{E}[\mathrm{Var}(A \mid Z)].$$

For the second part, use the law of total variance $\mathrm{Var}(Y) = \mathbf{E}[\mathrm{Var}(Y \mid Z)] + \mathrm{Var}(\mathbf{E}[Y \mid Z])$ and $\mathrm{Var}(A) = \mathbf{E}[\mathrm{Var}(A \mid Z)] + \mathrm{Var}(\mathbf{E}[A \mid Z])$. Realizing that both $\mathrm{Var}(Y)$ and $\mathrm{Var}(A)$ are constants that only depend on the joint distribution D, we finish the proof.

Appendix B. Missing Proofs in Classification (Section 4)

In what follows we first restate the propositions, lemmas and theorems in the main text, and then provide the corresponding proofs.

B.1 Convexity of R_{CE}

Proposition 4.1 The feasible region R_{CE} for the classification setting is convex.

Proof Let $Z_i = g_i(X)$ for $i \in \{0, 1\}$ with corresponding points $(I(Y;Z_i), I(A;Z_i)) \in R_{CE}$. Then we only need to prove that for $\forall u \in [0,1]$, $(uI(Y;Z_0) + (1-u)I(Y;Z_1), uI(A;Z_0) + (1-u)I(A;Z_1)) \in R_{CE}$ as well. For any $u \in [0,1]$, let $S \sim U(0,1)$, the uniform distribution over (0,1), such that $S \perp (Y, A)$. Consider the following randomized transformation $Z = (Z^0, S)$ where

$$Z^0 = \begin{cases} Z_0 & \text{If } S \leq u, \\ Z_1 & \text{otherwise.} \end{cases} \tag{23}$$

To compute $I(Y;Z)$, we have:

$$\begin{aligned} I(Y;Z) &= I(Y;Z^0,S) \\ &= I(S;Y) + I(Z^0;Y \mid S) \\ &= I(Z^0;Y \mid S) && \text{(S is independent of Y)} \\ &= \Pr(S \leq u)\, I(Y;Z_0) + \Pr(S > u)\, I(Y;Z_1) \\ &= uI(Y;Z_0) + (1-u)I(Y;Z_1). \end{aligned}$$

Similar argument could be used to show that $I(A;Z) = uI(A;Z_0) + (1-u)I(A;Z_1)$. So by construction we now find a randomized transformation $Z = g(X)$ such that $(uI(Y;Z_0) + (1-u)I(Y;Z_1), uI(A;Z_0) + (1-u)I(A;Z_1)) \in R_{CE}$. ■

B.2 Proof of Theorem 4.1

We proceed to provide the proof that the optimal value of (4) is the one given by Theorem 4.1.

Theorem 4.1 The optimal value E_Y of the program in Eqn. (4) i.e. the maximal mutual information under the independence constraint is given as:

$$\max_{Z : I(A;Z)=0} I(Y;Z) = H(Y) - D_{Y|A}\, H(A). \tag{6}$$

Proof For a joint distribution D over (X, A, Y) and a function $g : X \to Z$, in what follows we use $g \rfloor D$ to denote the induced distribution of D under g over (Z, A, Y). We first make the following claim: without loss of generality, for any joint distribution $g \rfloor D$ over (Z, A, Y), we could find $(Z_0, A^0, Y^0) \sim g \rfloor D$ and a deterministic function f, such that $Y^0 = f(A^0, Z_0, S)$ where $S \sim U(0,1)$, $S \perp (A^0, Z_0)$ and $I(Y^0; Z^0) \leq I(Y; Z)$ with $Z^0 = (Z_0, S)$. To see this, consider the following construction:

$$A_0, Z_0 \sim D(A, Z), \qquad S \sim U(0,1).$$

Let (a, z, s) be the sample of the above sampling process and construct

$$Y^0 = \begin{cases} 1 & \text{If } s \in \mathbf{E}[Y \mid A = a, Z = z], \\ 0 & \text{Otherwise.} \end{cases}$$

Now it is easy to verify that $(Z_0, A^0, Y^0) \sim_{TV} D$ and $\Pr(Y^0 = 1 \mid A^0 = a, Z_0 = z) = \mathbf{E}[Y \mid A = a, Z = z]$. To see the last claim, we have the following inequality hold:

$$I(Y^0; Z^0) = I(Y^0; Z_0, S) \geq I(Y; Z_0) = I(Y; Z).$$

Now to upper bound $I(Y; Z)$, we have

$$I(Y; Z) = H(Y) - H(Y \mid Z),$$

hence it suffices to lower bound $H(Y \mid Z)$. To this end, define

$$D_0 := \{z, \# \in (0,1) \mid f(0, z, \#) = 1\},$$
$$D_1 := \{z, \# \in (0,1) \mid f(1, z, \#) = 1\}.$$

Then,

$$\begin{aligned} \Pr((z, \#) \in D_0) &= \Pr(f(0, z, \#) = 1) \\ &= \Pr(f(0, z, \#) = 1 \mid A = 0) \\ &= \Pr(f(A, z, \#) = 1 \mid A = 0) \\ &= \Pr(Y = 1 \mid A = 0). \end{aligned}$$

Analogously, the following equation also holds:

$$\Pr((z, \#) \in D_1) = \Pr(Y = 1 \mid A = 1).$$

Without loss of generality, assume that $\Pr(Y = 1 \mid A = 1) \geq \Pr(Y = 1 \mid A = 0)$, then

$$\begin{aligned} &\Pr((z, \#) \in D_1 \cap D_0) \geq \Pr((z, \#) \in D_1) - \Pr((z, \#) \in D_0) \\ &\quad = |\Pr(Y = 1 \mid A = 1) - \Pr(Y = 1 \mid A = 0)|. \end{aligned}$$

But on the other hand, we know that if $(z, \#) \in D_1 \cap D_0$, then $f(1, z, \#) = 1$ and $f(0, z, \#) = 0$, and this implies that $Y = A$, hence:

$$\begin{aligned} &H(Y \mid Z) \geq H(Y \mid Z, S) \\ &\quad = \Pr((z, \#) \in D_1 \cap D_0) \cdot H(Y \mid (z, \#) \in D_1 \cap D_0) + \Pr((z, \#) \notin D_1 \cap D_0) \cdot H(Y \mid (z, \#) \notin D_1 \cap D_0) \\ &\quad \geq \Pr((z, \#) \in D_1 \cap D_0) \cdot H(Y \mid (z, \#) \in D_1 \cap D_0) \\ &\quad = \Pr((z, \#) \in D_1 \cap D_0) \cdot H(A) \\ &\quad \geq |\Pr(Y = 1 \mid A = 1) - \Pr(Y = 1 \mid A = 0)| \cdot H(A), \end{aligned}$$

which implies that

$$I(Y; Z) \leq H(Y) - |\Pr(Y = 1 \mid A = 1) - \Pr(Y = 1 \mid A = 0)| \cdot H(A) = H(Y) - D_{Y|A} \cdot H(A).$$

To see that the upper bound could be attained, let us consider the following construction. Denote $a := \Pr(Y = 1 \mid A = 0)$ and $b := \Pr(Y = 1 \mid A = 1)$. Construct a uniformly random $Z \sim U(0,1)$ and then sample A independently from Z according to the corresponding marginal distribution of A in D. Next, define:

$$Y = \begin{cases} 1 & \text{if } Z \leq a \wedge A = 0 \text{ or } Z \leq b \wedge A = 1, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that $Z \perp A$ by construction. Furthermore, by the construction of Y, we also have $(A, Y) \sim D(A, Y)$ hold. Since $I(Y;Z) = H(Y) - H(Y \mid Z)$, we only need to verify $H(Y \mid Z) = D_{Y|A}(H(A))$ in this case. Assume without loss of generality $a \leq b$, there are three different cases depending on the value of Z:

- $Z \leq a$: In this case no matter what the value of A, we always have $Y = 1$.
- $Z > b$: In this case no matter what the value of A, we always have $Y = 0$.
- $a < Z \leq b$: In this case $Y = A$, hence the conditional distribution of Y given $Z \in (a, b]$ is equal to the conditional distribution of A given $Z \in (a, b]$. But by our construction, A is independent of Z, which means that in this case the conditional distribution of A given $Z \in (a, b]$ is just the distribution of A.

Combine all the above three cases, we have:

$$\begin{aligned} H(Y \mid Z) &= \Pr(Z \leq a)\, H(Y \mid Z \leq a) + \Pr(Z > b)\, H(Y \mid Z > b) + \Pr(a < Z \leq b)\, H(Y \mid a < Z \leq b) = 0 + 0 \\ &\quad + |b - a|\, H(A \mid a < Z \leq b) \\ &= |\Pr(Y = 1 \mid A = 1) - \Pr(Y = 1 \mid A = 0)|\, H(A) \\ &= D_{Y|A}(H(A)), \end{aligned}$$

which completes the proof. ■

B.3 Proof of Theorem 4.2

Theorem 4.2 The optimal value E_A of the program in Eqn. (7) i.e. the Minimum Mutual Information under the Sufficient Statistics Constraint is

$$\min_{Z : I(Y;Z) = H(Y)} I(A;Z) = I(A;Y). \tag{8}$$

Proof First, realize that $H(Z) - I(Y;Z) = H(Y)$ by our constraint. Furthermore, we also know that $0 \leq H(Y \mid Z, A) \leq H(Y \mid Z) = H(Y) - I(Y;Z) = 0$, which means $H(Y \mid Z, A) = 0$. With these two observations, we have:

$$\begin{aligned} I(A;Z) &= H(Z) - H(Z \mid A) \\ &\geq H(Y) - H(Z \mid A) \\ &\geq H(Y) - H(Y, Z \mid A) \\ &= H(Y) - H(Y \mid A) - H(Y \mid Z, A) \\ &= H(Y) - H(Y \mid A) \\ &= I(A;Y). \end{aligned}$$

To attain the equality, simply set $Z = f_Y(X) = Y$. Specifically, this implies that 1-bit is sufficient to encode all the information for the optimal solution, which completes the proof. ■

B.4 Missing Proofs in Section 4.3 and Section 4.4

In what follows we provide proofs for Corollary 4.1 and Corollary 4.2.

Corollary 4.1 Let $Z = g(X)$ be any representation. If $I(Z;A) = 0$ then

$$\inf_h \mathbb{E}[\ell(Y, h(Z))] \geq D_{Y\|A} - H(A).$$

If $I(Y;Z) = H(Y)$, then

$$\inf_h \mathbb{E}[\ell(A, h(Z))] \geq H(A \mid Y).$$

Proof For the first inequality, let $\ell(\cdot,\cdot)$ be the cross-entropy loss. If $I(A;Z) = 0$, then by Theorem 4.1

$$\begin{aligned}\inf_h \mathbb{E}[\ell(Y, h(Z))] &= H(Y \mid Z) = H(Y) - I(Y;Z) \\ &= H(Y) - (H(Y) - D_{Y\|A} - H(A)) \\ &= D_{Y\|A} - H(A).\end{aligned}$$

Similarly, if the feature Z preserves all the information w.r.t. Y, e.g., $I(Y;Z) = H(Y)$, Theorem 4.2 immediately implies the following bound,

$$\begin{aligned}\inf_h \mathbb{E}[\ell(A, h(Z))] &= H(A \mid Z) = H(A) - I(A;Z) \\ &= H(A) - I(A;Y) = H(A \mid Y).\end{aligned}$$

■

Corollary 4.2 Suppose that the representation Z satisfies

$$\frac{I(A;Z)}{I(A;Y)} + \frac{H(Y \mid Z)}{D_{Y\|A} - H(A)} > 1. \tag{9}$$

Then Z is suboptimal.

Proof From Fig. 2d, the line connecting the points E_Y and E_A is given by:

$$I(A;Z) = \frac{I(A;Y)}{D_{Y\|A} - H(A)} I(Y;Z) - (H(Y) - D_{Y\|A} - H(A)). \tag{24}$$

Hence if

$$\frac{I(A;Z)}{I(A;Y)} > 1 - \frac{H(Y \mid Z)}{D_{Y\|A} - H(A)},$$

then the point $(I(Y;Z), I(A;Z))$ lies strictly on the upper-left side of the line connecting E_Y and E_A, hence it is suboptimal. ■

In what follows we provide a proof sketch for Proposition 4.2.

Proposition 4.2 When Z is discrete with alphabet size at most k, with the plug-in estimator (Wu and Yang, 2016) for mutual information applied to the left-hand side of (9) on a sample of size n from the joint distribution, if

$$\frac{\hat{I}(A;Z)}{\hat{I}(A;Y)} + \frac{\hat{H}(Y|Z)}{\hat{D}_{Y|A}\hat{H}(A)} > 1 + \#, \tag{10}$$

then the representation Z is certifiably suboptimal with probability at least $1 - \exp(-\Omega(n\#^2))$.

Proof It suffices to provide an estimator for

$$\frac{I(A;Z)}{I(A;Y)} + \frac{H(Y|Z)}{D_{Y|A}\, H(A)} \tag{25}$$

Since $D_{Y|A} = |\Pr_D(Y=1|A=0) - \Pr_D(Y=1|A=1)|$ is the difference of two conditional probabilities, for which we can simply estimate with a counting estimator that converges at rate $O(1/\sqrt{n})$.

Using the entropy estimators of (Wu and Yang, 2016), we can estimate all the quantities $H(Y)$, $H(A)$, $H(Z)$, $H(A,Z)$, and $H(A,Y)$, $H(Y,Z)$ up to an $O(1/\sqrt{n})$ additive error. Notice that:

$$\begin{aligned} I(A;Z) &= H(A) + H(Z) - H(A,Z) \\ I(A;Y) &= H(A) + H(Y) - H(A,Y) \\ H(Y|Z) &= H(Y,Z) - H(Y) \end{aligned}$$

Therefore, $I(A;Z)$, $I(A;Y)$, $H(Y|Z)$ can also be estimated with an $O(1/\sqrt{n})$ additive error. By our assumption, the denominators $I(A;Y)$, $H(A)$ are bound away from zero, therefore simply plug-in their estimator for $H(Y)$, $H(A)$, $H(Z)$, $H(A,Z)$, $H(A,Y)$, $H(Y,Z)$ provides the desired estimator. ■

Appendix C. Missing Proofs in Regression (Section 5)

C.1 Convexity of R_{LS}

Analogous to the classification setting, here we first show that the feasible region R_{LS} is convex:

Lemma 5.1 R_{LS} is convex.

Proof Let $Z_i = g_i(X)$ for $i \in \{0,1\}$ with corresponding points $(\mathrm{Var}_{\mathbf{E}}[Y|Z_i], \mathrm{Var}_{\mathbf{E}}[A|Z_i]) \in R_{LS}$. Then it suffices if we could show for $\forall u \in [0,1]$, $(u\,\mathrm{Var}_{\mathbf{E}}[Y|Z_0] + (1-u)\,\mathrm{Var}_{\mathbf{E}}[Y|Z_1], u\,\mathrm{Var}_{\mathbf{E}}[A|Z_0]) + (1-u)\,\mathrm{Var}_{\mathbf{E}}[A|Z_1]) \in R_{LS}$ as well.

We give a constructive proof. Due to the symmetry between A and Y, we will only prove the result for Y, and the same analysis could be directly applied to A as well. For any $u \in [0,1]$, let $S \sim U(0,1)$, the uniform distribution over $(0,1)$, such that $S \perp (Y, A)$. Consider the following randomized transformation $Z = (Z^0, S)$, where:

$$Z^0 = \begin{cases} Z_0 & \text{If } S \leq u, \\ Z_1 & \text{otherwise.} \end{cases} \tag{26}$$

To compute $\mathrm{Var}\,\mathbf{E}[Y \mid Z]$, define $K := \mathbf{E}[Y \mid Z]$, then by the law of total variance, we have:

$$\mathrm{Var}\,\mathbf{E}[Y \mid Z] = \mathrm{Var}(K) = \mathbf{E}[\mathrm{Var}(K \mid S)] + \mathrm{Var}\,\mathbf{E}[K \mid S].$$

We first compute $\mathrm{Var}\,\mathbf{E}[K \mid S]$:

$$\begin{aligned} \mathrm{Var}\,\mathbf{E}[K \mid S] &= \mathrm{Var}\,\mathbf{E}[\mathbf{E}[Y \mid Z] \mid S] \\ &= \mathrm{Var}\,\mathbf{E}[Y \mid S] && \text{(The law of total expectation)} \\ &= \mathrm{Var}\,\mathbf{E}[Y] && (Y \perp S) \\ &= 0. \end{aligned}$$

On the other hand, for $\mathbf{E}[\mathrm{Var}(K \mid S)]$, we have:

$$\begin{aligned} \mathbf{E}[\mathrm{Var}(K \mid S)] &= \Pr(S = 0)\,\mathrm{Var}(K \mid S = 0) + \Pr(S = 1)\,\mathrm{Var}(K \mid S = 1) \\ &= u\,\mathrm{Var}(K \mid S = 0) + (1 - u)\,\mathrm{Var}(K \mid S = 1) \\ &= u\,\mathrm{Var}\,\mathbf{E}[Y \mid Z_0] + (1 - u)\,\mathrm{Var}\,\mathbf{E}[Y \mid Z_1]. \end{aligned}$$

Combining both equations above yields:

$$\mathrm{Var}\,\mathbf{E}[Y \mid Z] = u\,\mathrm{Var}\,\mathbf{E}[Y \mid Z_0] + (1 - u)\,\mathrm{Var}\,\mathbf{E}[Y \mid Z_1].$$

Similar argument could be used to show that $\mathrm{Var}\,\mathbf{E}[A \mid Z] = u\,\mathrm{Var}\,\mathbf{E}[A \mid Z_0] + (1 - u)\,\mathrm{Var}\,\mathbf{E}[A \mid Z_1]$. So by construction we now find a randomized transformation $Z = g(X)$ such that $(u\,\mathrm{Var}\,\mathbf{E}[Y \mid Z_0] + (1 - u)\,\mathrm{Var}\,\mathbf{E}[Y \mid Z_1], u\,\mathrm{Var}\,\mathbf{E}[A \mid Z_0] + (1 - u)\,\mathrm{Var}\,\mathbf{E}[A \mid Z_1]) \in R_{LS}$, which completes the proof. ■

C.2 Proof of Theorem 5.1

In this section, we will prove Theorem 5.1. We will provide a proof in a generalized noisy setting, i.e., we no longer assume the noiseless condition (Assumption 5.1) so that the corresponding theorems in the noiseless setting follow as a special case. To this end, we first re-define

$$f_Y(X) := \mathbf{E}[Y \mid X] \tag{27}$$

$$f_A(X) := \mathbf{E}[A \mid X] \tag{28}$$

and $f_Y, f_A \in \mathbf{H}$. We reuse the notations a, y to denote

$$f_Y(X) = \mathbf{E}[Y|X] = \langle y, \phi(X) \rangle \tag{29}$$

$$f_A(X) = \mathbf{E}[A|X] = \langle a, \phi(X) \rangle. \tag{30}$$

It is easy to see that the noiseless setting is indeed a special case where $Y = \mathbf{E}[Y|X]$, $A = \mathbf{E}[A|X]$ almost surely.

For readers' convenience, we restate the Theorem 5.1 below:

Theorem 5.1 The optimal solution of optimization problem (12) is upper bounded by

$$\max_{Z, \mathrm{Var}\,\mathbf{E}[A|Z]=0} \mathrm{Var}\,\mathbf{E}[Y|Z] \le \mathrm{Var}(Y) - \frac{\mathrm{Cov}(\cdot, \cdot)^2}{\mathrm{Var}(\cdot)\mathrm{Var}(A)} = \mathrm{Var}(Y)(1 - r_{YA}^2),$$

where r_{YA} is the correlation coefficient of Y and A.

The following theorem is the generalized version of Theorem 5.1 in noisy setting:

Theorem C.1 The optimal solution of optimization problem (12) is upper bounded by

$$\max_{Z, \mathrm{Var}\,\mathbf{E}[A|Z]=0} \mathrm{Var}\,\mathbf{E}[Y|Z] \le \mathrm{Var}\,\mathbf{E}[Y|X] - C\frac{\mathrm{Cov}^2(\mathbf{E}[\cdot], \mathbf{E}[A|\cdot])}{\mathrm{Var}(\cdot)\mathrm{Var}\,\mathbf{E}[A|X]}. \tag{31}$$

It is easy to see Theorem 5.1 is an immediate corollary of this result: under the noiseless assumption, we have $\mathrm{Var}\,\mathbf{E}[Y|X] = \mathrm{Var}(Y)$ and $\mathrm{Var}\,\mathbf{E}[A|X] = \mathrm{Var}(A)$.

Proof

Using the law of total expectation,

$$\mathbf{E}[Y|Z] = \mathbf{E}[\mathbf{E}[Y|X]|Z] = \int_{\mathcal{X}} \mathbf{E}[Y|X,Z]\, p(X|Z)\, dX.$$

Since $Z = g(X)$ is a function of X, we have $Z \perp Y | X$, so $\mathbf{E}[Y|X,Z] = \mathbf{E}[Y|X] = f_Y(X)$. Therefore,

$$\begin{aligned}\mathbf{E}[Y|Z] &= \int_{\mathcal{X}} \mathbf{E}[Y|X,Z]\, p(X|Z)\, dX \\ &= \int_{\mathcal{X}} f_Y(X)\, p(X|Z)\, dX \\ &= \mathbf{E}[f_Y(X)|Z].\end{aligned}$$

Hence,

$$\mathrm{Var}\,\mathbf{E}[Y|Z] = \mathrm{Var}\,\mathbf{E}[f_Y(X)|Z]. \tag{32}$$

Therefore,

$$\begin{aligned}\mathrm{Var}\,\mathbf{E}[Y|Z] &= \mathrm{Var}\,\mathbf{E}[f_Y(X)|Z] \\ &= \mathrm{Var}\,\mathbf{E}[\langle y, j(X)\rangle | Z] \\ &= \mathrm{Var}\langle y, \mathbf{E}[j(X)|Z]\rangle \qquad \text{(Linearity of Expectation)} \\ &= \langle y, \mathrm{Var}\,\mathbf{E}[j(X)|Z] y\rangle.\end{aligned}$$

Similarly, for $A = \langle a, j(X) \rangle$, we have:

$$\mathrm{Var}\,\mathbf{E}[A \mid Z] = \langle a, \mathrm{Var}\,\mathbf{E}[j(X) \mid Z]\, a \rangle.$$

To simplify the notation, define $V := \mathrm{Var}\,\mathbf{E}[j(X) \mid Z]$. Then again, by the law of total variance, it is easy to verify that $0 \preceq V \preceq S = \mathrm{Var}\, j(X)$. Hence the original maximization problem could be relaxed as follows:

$$\underset{V}{\text{maximize}} \quad \langle y, V y \rangle, \qquad \text{subject to} \quad 0 \preceq V \preceq S, \qquad \langle a, V a \rangle = 0.$$

We now apply the transformation $V' := S^{-1/2} V S^{-1/2}$, $y_0 := S^{1/2} y$ and $a_0 := S^{1/2} a$ to further simplify the above optimization formulation:

$$\underset{V'}{\text{maximize}} \quad \langle y', V' y' \rangle, \qquad \text{subject to} \quad 0 \preceq V' \preceq I, \qquad \langle a', V_0 a' \rangle = 0.$$

To proceed, we first decompose y_0 orthogonally w.r.t. a':

$$y' = y'^{\perp a'} + y'^{\parallel a'},$$

where $y'^{\perp a'}$ is the component of y' that is perpendicular to a_0 and $y'^{\parallel a'}$ is the parallel component of y_0 to a'. Using this orthogonal decomposition, we have $\forall V' \preceq I$:

$$\begin{aligned} \langle y', V' y' \rangle &= \langle (y'^{\perp a'} + y'^{\parallel a'}), V' (y'^{\perp a'} + y'^{\parallel a'}) \rangle \qquad (33) \\ &= \langle y'^{\perp a'}, V' y'^{\perp a'} \rangle + \langle y'^{\parallel a'}, V' y'^{\parallel a'} \rangle + 2 \langle y'^{\perp a'}, V_0 y'^{\parallel a'} \rangle \\ &= \langle y'^{\perp a'}, V' y'^{\perp a'} \rangle \qquad (V'^{1/2} y'^{\parallel a'} = 0 \text{ since } V'^{1/2} a' = 0) \\ &\leq \langle y'^{\perp a'}, y'^{\perp a'} \rangle, \qquad (V' \preceq I) \end{aligned}$$

Next, realize that the vector $y'^{\perp a'}$ could be constructed as follows:

$$y'^{\perp a'} = y' - y'^{\parallel a'} = y' - \frac{\langle y', a' \rangle}{\langle a', a' \rangle} a',$$

We then plug the above expression of $y'^{\perp a'}$ to (33):

$$\begin{aligned} \langle y'^{\perp a'}, y'^{\perp a'} \rangle &= \left\langle y' - \frac{\langle y', a' \rangle}{\langle a', a' \rangle} a', y' - \frac{\langle y', a' \rangle}{\langle a', a' \rangle} a' \right\rangle \\ &= \langle y', y' \rangle - 2 \frac{\langle y', a' \rangle}{\langle a', a' \rangle} \langle y', a' \rangle + \frac{\langle y', a' \rangle^2}{\langle a', a' \rangle^2} \langle a', a' \rangle \\ &= \langle y', y' \rangle - \frac{\langle y', a' \rangle^2}{\langle a', a' \rangle} \\ &= \langle y, S y \rangle - \frac{\langle y, S a \rangle^2}{\langle a, S a \rangle} \\ &= \mathrm{Var}\,\mathbf{E}[Y \mid X] - \frac{\mathrm{Cov}^2(\mathbf{E}[Y \mid X], \mathbf{E}[A \mid X])}{\mathrm{Var}\,\mathbf{E}[A \mid X]} \\ &= \mathrm{Var}\,\mathbf{E}[Y \mid X] \left(1 - \frac{\mathrm{Cov}^2(\mathbf{E}[Y \mid X], \mathbf{E}[A \mid X])}{\mathrm{Var}\,\mathbf{E}[Y \mid X]\, \mathrm{Var}\,\mathbf{E}[A \mid X]} \right) \\ &= \mathrm{Var}\,\mathbf{E}[Y \mid X] \left(1 - r^2_{YA} \right), \end{aligned}$$

by using the fact that $\mathrm{Var}\,\mathbb{E}[Y \mid X] = \langle y, \Sigma y\rangle$, $\mathrm{Var}\,\mathbb{E}[A \mid X] = \langle a, \Sigma a\rangle$ and $\mathrm{Cov}(\mathbb{E}[Y \mid X], \mathbb{E}[A \mid X]) = \langle y, \Sigma a\rangle$.

To complete the proof, we only need to verify that the inequality in (33) can indeed be attained. We do so by construction, let

$$V^0 := I - \frac{a^0 (a^0)^\top}{\|a^0\|^2},$$

where $a_0^0 := a^0/\|a^0\|$ is the unit vector of a^0. Clearly, $0 \preceq V^0 \preceq I$. Furthermore,

$$\begin{aligned}\langle y^{0\perp a^0}, V^0 y^{\perp a^0}\rangle - \langle y^{0\perp a^0}, y^{0\perp a^0}\rangle &= \langle y^{0\perp a^0}, (a_0^0 (a_0^0)^\top - I) y^{\perp a^0}\rangle = \langle y^{0\perp a^0}, a^0 (a^0)^\top y^{\perp a^0}\rangle \\ &= \langle y^{0\perp a^0}, a^0\rangle^2 / \|a^0\|^2 \\ &= 0,\end{aligned}$$

completing the proof. ∎

C.3 Proof of Theorem 5.2

Again, we will provide a proof of Theorem 5.2 in a generalized noisy setting, i.e., we no longer assume the noiseless condition (Assumption 5.1) so that the corresponding theorems in the noiseless setting follow as a special case. We first restate Theorem 5.2 below:

Theorem 5.2 The optimal solution of optimization problem (13) is lower bounded by

$$\min_{Z, \mathrm{Var}\,\mathbb{E}[Y\mid Z] = \mathrm{Var}(Y)} \mathrm{Var}\,\mathbb{E}[A \mid Z] \geq \mathrm{Var}(A)\left(1 - \frac{\langle a, \Sigma y\rangle^2}{\|a\|^2 \|y\|^2}\right) = \mathrm{Var}(A)\,(1 - r_{YA}^2), \tag{14}$$

where r_{YA} is the correlation coefficient of Y and A.

The following theorem is the generalized version of Theorem 5.2 in noisy setting:

Theorem C.2 The optimal solution of optimization problem (13) is lower bounded by

$$\min_{Z, \mathrm{Var}\,\mathbb{E}[Y\mid Z] = \mathrm{Var}\,\mathbb{E}[Y\mid X]} \mathrm{Var}\,\mathbb{E}[A \mid Z] \geq \mathrm{Var}\,\mathbb{E}[A\mid X] - \frac{\mathrm{Var}\,\mathbb{E}[Y\mid X]\,\langle a, y\rangle^2}{\|y\|^2}. \tag{34}$$

It is easy to see Theorem 5.2 is an immediate corollary of this result: under the noiseless assumption, we have $\mathrm{Var}\,\mathbb{E}[Y\mid X] = \mathrm{Var}(Y)$ and $\mathrm{Var}\,\mathbb{E}[A\mid X] = \mathrm{Var}(A)$.

Proof As in the proof of Theorem 5.1, we have the following identities hold:

$$\begin{aligned}\mathrm{Var}\,\mathbb{E}[A \mid Z] &= \langle a, \mathrm{Var}\,\mathbb{E}[\phi(X) \mid Z]\, a\rangle, \\ \mathrm{Var}\,\mathbb{E}[Y \mid Z] &= \langle y, \mathrm{Var}\,\mathbb{E}[\phi(X) \mid Z]\, y\rangle.\end{aligned}$$

Again, let $V := \mathrm{Var}\,\mathbb{E}[\phi(X) \mid Z]$ so that we can relax the optimization problem as follows:

$$\begin{aligned}\underset{V}{\text{minimize}} \quad & \langle a, V a\rangle, \\ \text{subject to} \quad & 0 \preceq V \preceq \Sigma, \\ & \langle y, V y\rangle = \mathrm{Var}\,\mathbb{E}[Y\mid X] = \langle y, \Sigma y\rangle.\end{aligned}$$

Again, we apply the transformation $V^0 := S^{-1/2} V S^{-1/2}$, $y_0 := S^{1/2} y$ and $a_0 := S^{1/2} a$ to further simplify the above optimization formulation:

$$\begin{aligned} \underset{V \succeq 0}{\text{minimize}} \quad & \langle a^0, V^0 a^0 \rangle, \\ \text{subject to} \quad & 0 \preceq V^0 \preceq I, \langle y^0, V^0 y^0 \rangle = \\ & \langle y^0, y^0 \rangle. \end{aligned}$$

Note that since by the constraint $\langle y^0, V^0 y^0 \rangle = \langle y^0, y^0 \rangle$, which means $\langle y^0, (V^0 - I) y^0 \rangle = 0$, we also have $(V^0 - I)^{1/2} y^0 = 0$. To proceed, we first decompose a_0 orthogonally w.r.t. y^0:

$$a^0 = a^{0 \perp y^0} + a^{0 \parallel y^0},$$

where $a^{0 \perp y^0}$ is the component of a^0 that is perpendicular to y_0 and $a^{0 \parallel y^0}$ is the parallel component of a_0 to y^0. Using this orthogonal decomposition, we have $\forall V^0 \succeq 0$:

$$\begin{aligned} \langle a^0, V_o a^0 \rangle &= \langle a^{0 \perp y^0} + a^{0 \parallel y^0}, V^0 (a^{0 \perp y^0} + a^{0 \parallel y^0}) \rangle \qquad (35) \\ &= \langle a^{0 \perp y^0}, V^0 a^{0 \perp y^0} \rangle + \langle a^{0 \parallel y^0}, V^0 a^{0 \parallel y^0} \rangle + 2 \langle a^{0 \perp y^0}, V_o a^{0 \parallel y^0} \rangle \\ &= \langle a^{0 \perp y^0}, V^0 a^{0 \perp y^0} \rangle + \langle a^{0 \parallel y^0}, V^0 a^{0 \parallel y^0} \rangle + 2 \langle a^{0 \perp y^0}, V_o a^{0 \parallel y^0} \rangle - 2 \langle a^{0 \perp y^0}, a^{0 \parallel y^0} \rangle \\ &= \langle a^{0 \perp y^0}, V^0 a^{0 \perp y^0} \rangle + \langle a^{0 \parallel y^0}, V^0 a^{0 \parallel y^0} \rangle + 2 \langle a^{0 \perp y^0}, (V^0 - I) a^{0 \parallel y^0} \rangle \\ &= \langle a^{0 \perp y^0}, V^0 a^{0 \perp y^0} \rangle + \langle a^{0 \parallel y^0}, V^0 a^{0 \parallel y^0} \rangle \quad ((V^0 - I)^{1/2} a^{0 \parallel y^0} = 0 \text{ since } (V^0 - I)^{1/2} y^0 = 0) \\ & \langle a^{0 \parallel y^0}, V_o a^{0 \parallel y^0} \rangle \qquad (V^0 \succeq 0) \end{aligned}$$

On the other hand, from linear algebra, it is clear that

$$a^{0 \parallel y^0} = \frac{\langle a^0, y^0 \rangle}{\langle y^0, y^0 \rangle} y^0.$$

Plug the above formula to (35), yielding:

$$\begin{aligned} \langle a^{0 \parallel y^0}, V^0 a^{0 \parallel y^0} \rangle &= \frac{\langle a^0, y^0 \rangle}{\langle y^0, y^0 \rangle} y^0, V^0 \frac{\langle a^0, y^0 \rangle}{\langle y^0, y^0 \rangle} y^0 \rangle \\ &= \frac{\langle a^0, y^0 \rangle^2}{\langle y^0, y^0 \rangle^2} \langle y^0, V^0 y^0 \rangle \\ &= \frac{\langle a^0, y^0 \rangle^2}{\langle y^0, y^0 \rangle^2} \langle y^0, y^0 \rangle \qquad (\langle y^0, V^0 y^0 \rangle = \langle y^0, y^0 \rangle) \\ &= \frac{\langle a, S y \rangle^2}{\langle y, V y \rangle} \\ &= \frac{\mathrm{Cov}^2(\mathbb{E}[Y \mid X], \mathbb{E}[A \mid X])}{\mathrm{Var}\, \mathbb{E}[Y \mid X]} \\ &= \mathrm{Var}\, \mathbb{E}[A \mid X] \, r_{YA}^2, \end{aligned}$$

where we use the fact that $\mathrm{Var}\, \mathbb{E}[Y \mid X] = \langle y, S y \rangle$, $\mathrm{Var}\, \mathbb{E}[A \mid X] = \langle a, S a \rangle$ and $\mathrm{Cov}(\mathbb{E}[Y \mid X], \mathbb{E}[A \mid X]) = \langle y, S a \rangle$.

To complete the proof, we only need to verify that the inequality in (35) can indeed be attained. We do so by construction, let

$$V^0 := y^0_0 \quad \frac{\langle \cdot, y_0 \rangle}{},$$

where $y_0 := y^0/\|y^0\|$ is the unit vector of y^0. Clearly, $0 \quad V^0 \quad I$ and $\langle y^0, y^0 \rangle = \langle y^0, V^0 y^0 \rangle$. Furthermore,

$$\begin{aligned}\langle a^{0?y^0}, V^0 a^{0?y^0} \rangle &= \langle a^{0?y^0}, (y^0_0 \quad y^0) a^{0?y^0} \rangle = \langle a^{0?y^0}, y^0 \rangle^2 \\ &= \langle a^{0?y^0}, y^0 \rangle^2 / \|y^0\|^2 \\ &= 0,\end{aligned}$$

completing the proof. ■

Next, we provide the proof for Corollary 5.1:

Corollary 5.1 Let $Z = g(X)$ be any representation. If $\mathrm{Var}\,\mathbb{E}[A \mid Z] = 0$ then

$$\inf_h \mathbb{E}[\ell(Y, h(Z))] \quad \mathrm{Var}(Y)\ r_{YA}^2. \tag{15}$$

Furthermore, if $\mathrm{Var}\,\mathbb{E}[Y \mid Z] = \mathrm{Var}\, Y$, then

$$\inf_h \mathbb{E}[\ell(A, h(Z))] \quad \mathrm{Var}(A)(1 \quad r_{YA}^2). \tag{16}$$

Proof For the first inequality, let $\ell(\cdot, \cdot)$ be the squared loss. If $\mathrm{Var}\,\mathbb{E}[A \mid Z] = 0$, then by Theorem 5.1

$$\begin{aligned}\inf_h \mathbb{E}[\ell(Y, h(Z))] &= \mathbb{E}\,\mathrm{Var}(Y \mid Z) \\ &= \mathrm{Var}(Y) \quad \mathrm{Var}\,\mathbb{E}[Y \mid Z] \\ &\quad \mathrm{Var}(Y) \quad \mathrm{Var}(Y)\ (1 \quad r_{YA}^2) \\ &= \mathrm{Var}(Y)\ r_{YA}^2.\end{aligned}$$

Similarly, if the feature Z preserves all the information w.r.t. Y, e.g., $\mathrm{Var}\,\mathbb{E}[Y \mid Z] = \mathrm{Var}(Y)$, then Theorem 5.2 immediately implies the following bound,

$$\begin{aligned}\inf_h \mathbb{E}[\ell(A, h(Z))] &= \mathbb{E}\,\mathrm{Var}(A \mid Z) \\ &= \mathrm{Var}(A) \quad \mathrm{Var}\,\mathbb{E}[A \mid Z] \\ &\quad \mathrm{Var}(A) \quad \mathrm{Var}(A)\ r_{YA}^2 \\ &= \mathrm{Var}(A)\ (1 \quad r_{YA}^2).\end{aligned}$$

■

C.4 Proof of Theorem 5.4

We will prove a generalized version of Theorem 5.4 without noiseless assumption, stated below:

Theorem C.3 The optimal solution of the Lagrangian has the following lower bound:

$$\mathrm{OPT}(\lambda) \geq \frac{1}{2}\Big[\lambda \operatorname{Var}\mathbf{E}[A|X] + \operatorname{Var}\mathbf{E}[Y|X] - \sqrt{(\lambda \operatorname{Var}\mathbf{E}[A|X] + \operatorname{Var}\mathbf{E}[Y|X])^2 - 4\lambda \operatorname{Cov}^2(\mathbf{E}[Y|X], \mathbf{E}[A|X])}\Big].$$

To have a sanity check, note that when the noiseless assumption holds, we have $\operatorname{Var}\mathbf{E}[A|X] = \operatorname{Var}(A)$ and $\operatorname{Var}\mathbf{E}[Y|X] = \operatorname{Var}(Y)$, hence the bound above simplifies to:

$$\frac{1}{2}\Big[\lambda \operatorname{Var}(A) + \operatorname{Var}(Y) - \sqrt{(\lambda \operatorname{Var}(A) + \operatorname{Var}(Y))^2 - 4\lambda(1-\rho^2)\operatorname{Var}(A)\operatorname{Var}(Y)}\Big],$$

which reduces to the lower bound in Theorem 5.4.

Proof [Proof of Theorem 5.4]

Recall that our objective is:

$$\underset{Z=g(X)}{\text{minimize}} \quad \lambda \operatorname{Var}\mathbf{E}[A \mid Z] - \operatorname{Var}\mathbf{E}[Y \mid Z] \tag{36}$$

As in the proof of Theorem 5.1, we have the following identities hold:

$$\operatorname{Var}\mathbf{E}[A \mid Z] = \langle a, \operatorname{Var}\mathbf{E}[\phi(X) \mid Z]a\rangle,$$
$$\operatorname{Var}\mathbf{E}[Y \mid Z] = \langle y, \operatorname{Var}\mathbf{E}[\phi(X) \mid Z]y\rangle.$$

Again, let $V := \operatorname{Var}\mathbf{E}[\phi(X) \mid Z]$ so that we can relax the optimization problem as follows:

$$\begin{aligned} \underset{V}{\text{minimize}} \quad & \langle \lambda a, Va\rangle - \langle y, Vy\rangle = \operatorname{tr}(V(\lambda aa^T - yy^T)), \\ \text{subject to} \quad & 0 \preceq V \preceq S. \end{aligned} \tag{37}$$

It is worth pointing out that the constraint in the optimization problem (37) is a superset (hence a relaxation) of the one in (36), since there may not exist representation $Z = g(X)$ that attain a given $V = \operatorname{Var}\mathbf{E}[\phi(X) \mid Z]$. As a result, the optimal solution of (37) is a lower bound of that of (36).

Similar to the proofs of Theorem 5.1 and Theorem 5.2, we apply the transformation $V' := S^{-1/2}VS^{-1/2}$, $y_0 := S^{1/2}y$ and $a' := S^{1/2}a$ to further simplify the above optimization formulation:

$$\begin{aligned} \underset{V' \succeq 0}{\text{minimize}} \quad & \langle \lambda a', V'a'\rangle - \langle y', V'y'\rangle = \operatorname{tr}(V'(\lambda a'a'^T - y'y'^T)), \\ \text{subject to} \quad & 0 \preceq V' \preceq I. \end{aligned} \tag{38}$$

One key observation to the optimization problem (38) is that the matrix $R := \lambda a'a'^T - y'y'^T$ has rank at most 2. Furthermore, thanks to Lemma C.1, we can fully characterize the spectrum of this matrix: Let $s_i(R)$ be the i-th largest eigenvalue of R, then by Lemma C.1, the spectrum of R admits the following form:

$$\begin{aligned} s_1(R) &= \frac{1}{2}\Big[\lambda\langle a', a'\rangle - \langle y, y\rangle + \sqrt{(\lambda\langle a', a'\rangle + \langle y, y\rangle)^2 - 4\lambda\langle a', y\rangle^2}\Big], \\ s_d(R) &= \frac{1}{2}\Big[\lambda\langle a', a'\rangle - \langle y, y\rangle - \sqrt{(\lambda\langle a', a'\rangle + \langle y, y\rangle)^2 - 4\lambda\langle a', y\rangle^2}\Big], \\ s_2(R) &= \cdots = s_{d-1}(R) = 0. \end{aligned}$$

Now by the spectral theorem, R could be decomposed as follows:

$$R = s_1(R)u_1u_1^T + s_d(R)u_du_d^T,$$

where u_i is the corresponding orthogonal eigenvectors. Hence, due to the Von-Neumann's trace inequality, we have:

$$\begin{aligned}
\mathrm{tr}(V^0(\lambda a^0 a^{0T} - y^0 y^{0T})) &= \mathrm{tr}(V^0 R) \\
&\le \sum_{i=1}^{d} s_i(R)s_{d-i+1}(V^0) && \text{(Von-Neumann's trace inequality)} \\
&= s_1(R)s_d(V^0) + s_d(R)s_1(V^0) && (s_2(R) = \ldots = s_{d-1}(R) = 0) \\
&\le s_d(R)s_1(V^0) && (s_1(R) > 0 \text{ and } s_d(V^0) \ge 0) \\
&\le s_d(R). && (s_d(R) < 0 \text{ and } s_1(V^0) \le 1)
\end{aligned}$$

Now it suffices to verify that the lower bound above is attainable. To see this, let $V^0 = u_du_d^T$, i.e., the projection matrix of the eigenvector corresponding to $s_d(R)$. The last step is to plug $s_d(R)$ into the original optimization objective (36), yielding:

$$\begin{aligned}
\mathrm{OPT}(\lambda) &= \min_Z \lambda \mathrm{Var}[\mathbb{E}[A|Z]] - \mathrm{Var}[\mathbb{E}[Y|Z]] \\
&= s_d(R) \\
&= \frac{1}{2}\Big[\lambda\langle a, \Sigma a\rangle - \langle y, \Sigma y\rangle - \sqrt{(\lambda\langle a, \Sigma a\rangle + \langle y, \Sigma y\rangle)^2 - 4\lambda\langle a, \Sigma y\rangle^2}\Big] \\
&= \frac{1}{2}\Big[\lambda \mathrm{Var}[\mathbb{E}[A|X]] - \mathrm{Var}[\mathbb{E}[Y|X]] - \sqrt{(\lambda \mathrm{Var}[\mathbb{E}[A|X]] + \mathrm{Var}[\mathbb{E}[Y|X]])^2 - 4\lambda \mathrm{Cov}^2(\mathbb{E}[Y|X], \mathbb{E}[A|X])}\Big]
\end{aligned}$$

Hence we have completed the proof. ■

C.5 Explicit formula for eigenvalues

The following lemma is used the in the proof of Theorem 5.4 to characterize the spectrum of the matrix $R = \lambda a^0 a^{0T} - y^0 y^{0T}$.

Lemma C.1 Let $R := \lambda a^0 a^{0T} - y^0 y^{0T}$, where $y^0 := \Sigma^{1/2} y$ and $a_0 := \Sigma^{1/2} a$. Then the eigenvalues of R are

$$\begin{aligned}
s_1(R) &= \frac{1}{2}\Big[\lambda\langle a, \Sigma a\rangle - \langle y, \Sigma y\rangle + \sqrt{(\lambda\langle a, \Sigma a\rangle + \langle y, \Sigma y\rangle)^2 - 4\lambda\langle a, \Sigma y\rangle^2}\Big], \\
s_d(R) &= \frac{1}{2}\Big[\lambda\langle a, \Sigma a\rangle - \langle y, \Sigma y\rangle - \sqrt{(\lambda\langle a, \Sigma a\rangle + \langle y, \Sigma y\rangle)^2 - 4\lambda\langle a, \Sigma y\rangle^2}\Big] \\
s_2(R) &= \cdots = s_{d-1}(R) = 0
\end{aligned}$$

Proof Since $\mathrm{rank}(R) \le \mathrm{rank}(\lambda a^0 a^{0T}) + \mathrm{rank}(y^0 y^{0T}) \le 2$, R has at most two non-zero eigenvalues $s_1(R)$ and $s_d(R)$. Notice that

$$\mathrm{tr}(R) = \sum_{i=1}^{d} s_i(R) = s_1(R) + s_d(R),$$
$$\mathrm{tr}(R^2) = \sum_{i=1}^{d} s_i^2(R) = s_1^2(R) + s_d^2(R)$$

We can write $\mathrm{tr}(R)$ and $\mathrm{tr}(R^2)$ explicitly:

$$\mathrm{tr}(R) = \lambda \mathrm{tr}(a^0 a^{0T}) + \mathrm{tr}(y^0 y^{0T}) = \lambda \langle a, Sa \rangle + \langle y, Sy \rangle$$

$$\begin{aligned}
\mathrm{tr}(R^2) &= \mathrm{tr}((\lambda a^0 a^{0T} + y^0 y^{0T})(\lambda a^0 a^{0T} + y^0 y^{0T})) \\
&= \lambda^2 (a^{0T} a^0)\, \mathrm{tr}(a^0 a^{0}_{\tau}) + \lambda (a^0_{\tau} y^0)\, \mathrm{tr}(a^0 y^{0T}) + \lambda (a_{0\tau} y^0)\, \mathrm{tr}(y^0 a^{0T}) + (y^{0T} y^0)\, \mathrm{tr}(y^0 y^{0T}) \\
&= \lambda^2 (a^{0T} a^0)^2 + 2\lambda (a^{0T} y^0)^2 + (y^{0T} y^0)^2 \\
&= \lambda^2 (a^T S a)^2 + 2\lambda (a^T S y)^2 + (y^T S y)^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
s_1(R) + s_d(R) &= \lambda \langle a, Sa \rangle + \langle y, Sy \rangle \\
s_1(R) s_d(R) &= \frac{1}{2} \left((s_1(R) + s_d(R))^2 - (s_1^2(R) + s_d^2(R)) \right) \\
&= \lambda \langle a, Sy \rangle^2 + \lambda \langle a, Sa \rangle \langle y, Sy \rangle
\end{aligned}$$

Thus $s_1(R)$ and $s_d(R)$ are the roots of the quadratic equation in terms of x:

$$x^2 - (\lambda \langle a, Sa \rangle + \langle y, Sy \rangle) x + \lambda \langle a, Sy \rangle^2 - \lambda \langle a, Sa \rangle \langle y, Sy \rangle = 0$$

We complete the proof by solving this quadratic equation. ■

C.6 Proof of Theorem 5.3

We consider the constrained problem:

$$\begin{aligned}
&\underset{Z}{\text{maximize}} && \mathrm{Var}\left[\mathbb{E}[Y \mid Z]\right] \\
&\text{subject to} && \mathrm{Var}\left[\mathbb{E}[A \mid Z]\right] \le c.
\end{aligned} \tag{39}$$

Again, in what follows we provide proof for a generalized theorem without the noiseless assumption. One can readily verify that $\mathrm{Var}\left[\mathbb{E}[Y|X]\right] = \mathrm{Var}(Y)$ and $\mathrm{Var}\left[\mathbb{E}[A|X]\right] = \mathrm{Var}(A)$ under the noiseless setting. Hence the theorem below reduces to Theorem 5.3 when the noiseless assumption holds.

Theorem C.4 The optimal solution of the constrained problem (17) has the following upper bound: for any $c \in [0, r_{YA}^2\, \mathrm{Var}\left[\mathbb{E}[A \mid X]\right]]$, we have

$$\mathrm{Var}\left[\mathbb{E}[Y \mid Z]\right] - \mathrm{Var}\left[\mathbb{E}[Y|X]\right] \le 2 r_{YA} \sqrt{(1 - r_{YA}^2)\alpha(1-\alpha)} + 1 - \alpha - r_{YA}^2 + 2\alpha r_{YA}^2 \tag{40}$$

where $\alpha := c / \mathrm{Var}\left[\mathbb{E}[A \mid X]\right]$.

Proof Recall that

$$\mathrm{Var}\,\mathbf{E}[Y|Z] \qquad \sup_{\beta \geq 0} \{\mathrm{OPT}(\beta) \quad \beta c\} \tag{41}$$

Therefore, it is crucial to find the supremum on the RHS, for which we provide the following lemma:

Lemma C.2 Define

$$y(\beta;\, a,b,c,t) := \beta c \qquad \sqrt{a^2 + \beta^2 b^2 + 2\beta t a b} \tag{42}$$

where $a, b \geq 0, |c| \leq b, |t| \leq 1$ Then,

$$\sup_{\beta \geq 0} y(\beta;\, a,b,c,t) = \begin{cases} \frac{a}{b}\sqrt{(1 \quad t^2)(b^2 \quad c^2)} + ct & \text{if } c \leq t \leq b, \\ a & \text{else .} \end{cases} \tag{43}$$

Proof To avoid notational clutter, we use $y(\beta)$ as a shorthand for $y(\beta;\, a,b,c,t)$ when the meaning is clear from context. Direct calculation gives the derivative of y:

$$\frac{\partial y}{\partial \beta} = c \quad \frac{b(\beta b + at)}{\sqrt{a^2 + \beta^2 b^2 + 2\beta t a b}} \tag{44}$$

When $c = b$, the derivative $\frac{\partial y}{\partial \beta}$ is always positive. Therefore, the supremum is achieved at $y(+\infty) = at$, which coincides with the first case in equation (43).

Solving the equation $\frac{\partial y}{\partial \beta} = 0$ gives two real roots:

$$\beta = \frac{at}{b} \quad \frac{ac}{b}\sqrt{\frac{1 \quad t^2}{b^2 \quad c^2}} \tag{45}$$

When $b > c \leq b$, one of the roots $\beta = \frac{at}{b} + \frac{ac}{b}\sqrt{\frac{1 \quad t^2}{b^2 \quad c^2}}$ is positive. Hence, the supremum is achieved at $y(0)$, $y(\beta)$ or $y(+\infty)$. We can verify that $y(\beta) = \frac{a}{b}\sqrt{(1 \quad t^2)(b^2 \quad c^2)} + ct$ is indeed the maximum among them.

When $c \leq b$, there is no stationary point in $[0, +\infty)$, therefore the supremum is achieved at $y(0)$ or $y(+\infty)$. We can verify that $y(0) = a$ is the larger one.

Combining the three cases, we have completed the proof. ∎

Using the definition of $y(\beta;\, a,b,c,t)$, it is now clear that

$$\sup_{\beta \geq 0}\left\{\mathrm{OPT}(\beta) \quad \beta c\right\} = \frac{1}{2}\sup_{\beta \geq 0} y(\beta;\, a',b',c',t') + \frac{1}{2}\mathrm{Var}\,\mathbf{E}[Y|X]$$

where

$$a' = \mathrm{Var}\,\mathbf{E}[Y|X],\ b' = \mathrm{Var}\,\mathbf{E}[A|X],\ c' = \mathrm{Var}\,\mathbf{E}[A|X] \quad 2c,\ t' = 1 \quad 2r_{YA}^2 \tag{46}$$

To simplify the equations, we introduce the notation

$$\alpha = \frac{c}{\mathrm{Var}\,\mathbf{E}[A|X]} \tag{47}$$

Notice that $c_0 \quad t^0b^0 = 2(r_{YA}^2 \ \mathrm{Var}\,\mathbb{E}[AjX] \quad c) = 2\,\mathrm{Var}\,\mathbb{E}[AjX](r_{YA}^2 \quad a)$, by Lemma C.2, we have

(1) When $a > r_{YA}^2$: we have $c_0 \quad t^0b^0 < 0$, therefore,

$$\sup_{l\,0} y(l; a^0, b^0, c^0, t^0) = a^0 = \mathrm{Var}\,\mathbb{E}[YjX]$$

hence, in this case, we have

$$\mathrm{Var}\,\mathbb{E}[YjZ] \quad \sup_{l\,0}\left\{ OPT(l) \quad lc \right\} = \frac{1}{2}\mathrm{Var}\,\mathbb{E}[YjX] + \frac{1}{2}\mathrm{Var}\,\mathbb{E}[YjX] = \mathrm{Var}\,\mathbb{E}[YjX] \quad (48)$$

which is a trivial upper bound. Indeed, in this regime, $\mathrm{Var}\,\mathbb{E}[AjZ]$ is allowed to be larger than the lower bound in Theorem C.2, where one achieves no trade-off on the utility $\mathrm{Var}\,\mathbb{E}[YjZ]$.

(2) When $a \quad r^2_{AY}$: we have $c_0 \quad t^0b^0 \ 0$, therefore,

$$\begin{aligned}\sup_{l\,0} y(l; a^0, b^0, c^0, t^0) &= \frac{a^0}{b^0}\left(1 \quad t^{02}\right)\left(b^{02} \quad c^{02}\right) + c^0t^0 \\ &= \mathrm{Var}\,\mathbb{E}[YjX] \quad 4r_{YA} \quad (1 \quad r_{YA}^2)a(1 \quad a) + (1 \quad 2a)(1 \quad 2r_{YA}^2)\end{aligned}$$

hence, in this case, we have

$$\begin{aligned}\mathrm{Var}\,\mathbb{E}[YjZ] &\quad \sup_{l\,0}\left\{ OPT(l) \quad lc \right\} \\ &= \frac{1}{2}\sup_{l\,0} y(l; a^0, b^0, c^0, t^0) + \frac{1}{2}\mathrm{Var}\,\mathbb{E}[YjX] \\ &= \mathrm{Var}\,\mathbb{E}[YjX] \quad 2r_{YA} \quad (1 \quad r_{YA}^2)a(1 \quad a) + 1 \quad a \quad r_{YA}^2 + 2ar_{YA}^2\end{aligned}$$

Specifically, when $c = 0$, i.e. $a = 0$, this upper bound simplifies to:

$$\mathrm{Var}\,\mathbb{E}[YjZ] \ \mathrm{Var}\,\mathbb{E}[YjX](1 \quad r_{YA}^2), \quad (49)$$

which is exactly Theorem C.1. ■

To finish this section, it is also worth pointing out that the lower bound in (22) admits a geometric interpretation, as shown in Figure 8. Essentially, the lower bound of $OPT(l)$ in Theorem 5.4 corresponds to the half of $jOAj \quad (jOYj + jAYj)$ in the triangle of Figure 8. Hence due to the triangle inequality, this term is always nonpositive. In particular, if $c = 0$, then $\sup_{l\,0} OPT(l)$ is attained at $l = 0$.

C.7 Tightness of the Bounds

In the main text we mention that a sufficient condition on the regularity of (X, j) is that $j(X)$ follows a Gaussian distribution. To prove this theorem, we first prove the following lemma, which shows that for rank-1 matrix M with a special structure, it suffices for us to choose $g(X)$ as a deterministic linear transform.

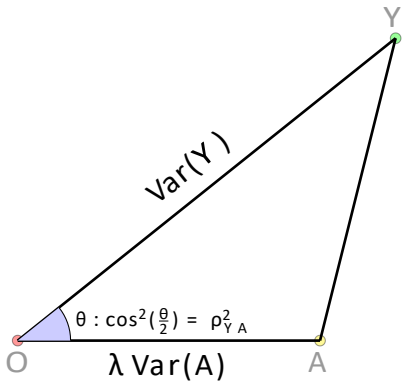


Figure 8: A geometric interpretation of OPT(λ). By the cosine theorem, $\|AY\|^2 = \mathrm{Var}^2(Y) + \lambda^2 \mathrm{Var}^2(A) - 2\lambda \mathrm{Var}(A)\mathrm{Var}(Y)(2r^2_{YA} - 1)$. Hence the lower bound of OPT(λ) in Theorem 5.4 corresponds to the half of $\|OA\| - (\|OY\| + \|AY\|)$ in the above triangle. Hence due to the triangle inequality, this term is always nonpositive.

Lemma C.3 Let $\Sigma = \sum_{i=1}^{d} s_i u_i u_i^T$ be the spectral decomposition of Σ. If $f(X)$ is Gaussian and $M = s_j u_j u_j^T$ for some $j \in [d]$, then there exists $L = M\Sigma^{-1/2}$ such that for $Z = Lf(X)$, we have

$$\mathrm{Var}\,\mathbf{E}[f(X) \mid Z] = M.$$

Proof Note that when $f(X)$ is Gaussian, $(f(X), Lf(X))$ is jointly Gaussian for any $L \in \mathbf{R}^{dd}$, since $(f(X), Lf(X))$ is a linear transformation of $f(X)$. Let $Z = Lf(X)$. It can be readily verified that the covariance $\mathrm{Cov}(f(X), Z)$ is given by

$$\begin{aligned}\mathrm{Cov}(f(X), Z) &= \mathbf{E}[\langle f(X) - \mathbf{E}[f(X)], Z - \mathbf{E}[Z]\rangle] \\ &= \mathbf{E}[\langle f(X) - \mathbf{E}[f(X)], Lf(X) - L\mathbf{E}[f(X)]\rangle] \\ &= \Sigma L^T.\end{aligned}$$

Furthermore, we know that the conditional distribution $f(X) \mid Z$ is Gaussian as well, with mean and covariance given by:

$$\begin{aligned}\mathbf{E}[f(X) \mid Z] &= \mathbf{E}[f(X)] + \mathrm{Cov}(f(X), Z)\,\mathrm{Var}(Z)^{-1}(Z - \mathbf{E}[Z]) \\ &= \mathbf{E}[f(X)] + \Sigma L^T (L\Sigma L^T)^{-1}(Z - L\mathbf{E}[f(X)]), \\ \mathrm{Var}(f(X) \mid Z) &= \mathrm{Var}(f(X)) - \mathrm{Cov}(f(X), Z)\,\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z, f(X)) \\ &= \Sigma - \Sigma L^T (L\Sigma L^T)^{-1} L\Sigma.\end{aligned}$$

Hence,

$$\mathbf{E}[\mathrm{Var}(f(X) \mid Z)] = \Sigma - \Sigma L^T (L\Sigma L^T)^{-1} L\Sigma,$$

and by the law of total variance, we have

$$\begin{aligned}\mathrm{Var}\,\mathbf{E}[j(X)\mid Z] &= \mathrm{Var}(j(X)) \quad \mathbf{E}[\mathrm{Var}(j(X)\mid Z)] \\ &= S \quad S \quad SL^T(LSL^T)^{-1}LS \\ &= SL^T(LSL^T)^{-1}LS.\end{aligned}$$

Now it suffices that for $M = s_j u_j u_j^T$, we could find the corresponding $L \in \mathbf{R}^{dd}$ such that for $Z = Lj(X)$,

$$M = \mathrm{Var}\,\mathbf{E}[j(X)\mid Z] = SL^T(LSL^T)^{-1}LS.$$

To proceed, define $P := LS^{1/2}$. In order to find the linear transform L, we need to solve the following equation in terms of L:

$$\begin{aligned}M = SL^T(LSL^T)^{\dagger}LS \quad &() \quad S^{-1/2}MS^{-1/2} = S^{1/2}L^T(LSL^T)^{\dagger}LS^{1/2} \\ &() \quad S^{-1/2}MS^{-1/2} = P^T(PP^T)^{\dagger}P \\ &() \quad u_j u_j^T = P^T(PP^T)^{\dagger}P \\ &() \quad M(M^TM)^{\dagger}M^T = P^T(PP^T)^{\dagger}P.\end{aligned}$$

It is then straightforward to read the solution from the above equation, giving us $M^T = P$, which implies

$$M = M^T = LS^{1/2} \;=)\; L = MS^{-1/2},$$

completing the proof. ■

With the help of Lemma C.3, we can now prove Theorem 5.5. The high-level idea of the proof, again, is by constructing randomized feature transformations.

Theorem 5.5 (X, j) is regular if $j(X)$ follows a Gaussian distribution.

Proof Recall that by definition of regularity, it suffices for us to prove the following: for any positive semidefinite matrix M: $S \quad M \quad 0$, there exists (randomized) $Z = g(X)$, such that

$$\mathrm{Var}\,\mathbf{E}[j(X)\mid Z] = M.$$

Let $S = \sum_{i=1}^{d} s_i u_i u_i^T$ be the spectral decomposition of S. Since $0 \quad M \quad S$, there exists $0 \quad m_i \quad s_i, \forall i \in [d]$, such that

$$M = \sum_{i=1}^{d} m_i u_i u_i^T.$$

Define $s = \sum_{i=1}^{d} m_i / s_i$, then M could be equivalently expressed as

$$M = \sum_{i=1}^{d} m_i u_i u_i^T = \sum_{i=1}^{d} \frac{m_i}{s s_i}\, s s_i u_i u_i^T =: \sum_{i=1}^{d} \frac{m_i}{s s_i}\, s M_i.$$

In other words, M is a convex combination of $\{sM_i\}_{i\in[d]}$. Following Lemma C.3, for each $M_i = s_i u_i u_i^T$, there exists $L_i := M_i S^{-1/2}$ and we can construct $Z_i := sL_i j(X)$, such that

$$\mathrm{Var}\,\mathbf{E}[j(X)\mid Z_i] = sM_i.$$

Now consider the following randomized feature Z. First, let $S \sim U(0,1)$ be a uniformly random variable over $(0,1)$. Construct

$$Z = \begin{cases} z_1 & \text{If } 0 \le S < \frac{m_1}{s s_1}, \\ z_2 & \text{If } \frac{m_1}{s s_1} \le S < \frac{m_1}{s s_1} + \frac{m_2}{s s_2}, \\ \vdots & \\ z_d & \text{If } \sum_{i=1}^{d-1} \frac{m_i}{s s_i} \le S < 1. \end{cases} \tag{50}$$

To compute $\mathrm{Var}_{\mathbf{E}}[j(X) \mid Z]$, define $K := \mathbf{E}[j(X) \mid Z]$, then by the law of total variance, we have:

$$\mathrm{Var}_{\mathbf{E}}[j(X) \mid Z] = \mathrm{Var}(K) = \mathbf{E}[\mathrm{Var}(K \mid S)] + \mathrm{Var}_{\mathbf{E}}[K \mid S].$$

We first compute $\mathrm{Var}_{\mathbf{E}}[K \mid S]$:

$$\begin{aligned} \mathrm{Var}_{\mathbf{E}}[K \mid S] &= \mathrm{Var}_{\mathbf{E}}[\mathbf{E}[j(X) \mid Z] \mid S] \\ &= \mathrm{Var}_{\mathbf{E}}[j(X) \mid S] && \text{(The law of total expectation)} \\ &= \mathrm{Var}_{\mathbf{E}}[j(X)] && (j(X) \perp S) \\ &= 0. \end{aligned}$$

On the other hand, for $\mathbf{E}[\mathrm{Var}(K \mid S)]$, we have:

$$\begin{aligned} \mathbf{E}[\mathrm{Var}(K \mid S)] &= \sum_{i=1}^{d} \Pr(S = i)\, \mathrm{Var}(K \mid S = i) \\ &= \sum_{i=1}^{d} \frac{m_i}{s s_i}\, \mathrm{Var}(K \mid S = i) \\ &= \sum_{i=1}^{d} \frac{m_i}{s s_i}\, \mathrm{Var}_{\mathbf{E}}[j(X) \mid Z, S = i] \\ &= \sum_{i=1}^{d} \frac{m_i}{s s_i}\, \mathrm{Var}_{\mathbf{E}}[j(X) \mid Z = z_i] \\ &= \sum_{i=1}^{d} \frac{m_i}{s s_i}\, s M_i \\ &= M, \end{aligned}$$

completing the proof. ■

To prove Theorem 5.6, in what follows we prove that under the regularity condition, all the bounds in Theorem 5.1, Theorem 5.2 and Theorem 5.4 are tight.

C.7.1 TIGHTNESS OF THE BOUND IN THEOREM 5.1

In the proof of Theorem 5.1, we showed an upper bound on the optimal solution via an SDP relaxation. Therefore, the upper bound is achievable whenever the SDP relaxation is tight.

Corollary C.1 When (X, j) is regular, the upper bound in Theorem 5.1 is achievable.

Proof This immediately follows from the proof of Theorem 5.1 and Theorem 5.5: if (X, j) is regular, then there exists $Z = g(X)$ such that $\mathrm{Var}[j(X) \mid Z] = V$, where V is the optimal solution of the SDP relaxation in the proof of Theorem 5.1. ■

C.7.2 TIGHTNESS OF THE BOUND IN THEOREM 5.2

In the proof of Theorem 5.2, we showed a lower bound on the optimal solution via an SDP relaxation. Therefore, the lower bound is achievable whenever the SDP relaxation is tight.

Corollary C.2 When (X, j) is regular, the lower bound in Theorem 5.2 is achievable.

Proof This immediately follows from the proof of Theorem 5.2 and Theorem 5.5: if (X, j) is regular, then there exists $Z = g(X)$ such that $\mathrm{Var}[j(X) \mid Z] = V$, where V is the optimal solution of the SDP relaxation in the proof of Theorem 5.2. ■

C.7.3 TIGHTNESS OF THE BOUND IN THEOREM 5.4

In the proof of Theorem 5.4, we showed a lower bound on the tradeoff via an SDP relaxation. Therefore, the lower bound is achievable whenever the SDP relaxation is tight.

Proof From the proof of Theorem 5.4, we can see that if there exists $Z = g(X)$, such that

$$\mathrm{Var}[j(X) \mid Z] = S^{1/2} u_d u_d^T S^{1/2},$$

where u_d is the unit eigenvector of R with eigenvalue $s_d(R)$, then the equality is achievable. It is easy to see that

$$S - S^{1/2} u_d u^T S^{1/2} \succeq 0.$$

Therefore, choosing $M = S^{1/2} u_d u_d^T S^{1/2}$ in the definition of regularity guarantees the existence of Z. Hence we have completed the proof. ■

C.8 Proof without RKHS assumption

The following lemma is crucial.

Lemma C.4 For any f, g,

$$\frac{1}{2}\left\{\mathrm{Var}[f(X)] + \mathrm{Var}[g(X)] - \sqrt{(\mathrm{Var}[f(X)] + \mathrm{Var}[g(X)])^2 - 4\,\mathrm{Cov}(f(X), g(X))^2}\right\} \quad (51)$$

and

$$\frac{1}{2}\left\{\mathrm{Var}[f(X)] + \mathrm{Var}[g(X)] + \sqrt{(\mathrm{Var}[f(X)] + \mathrm{Var}[g(X)])^2 - 4\,\mathrm{Cov}(f(X), g(X))^2}\right\} \quad (52)$$

Proof After re-organizing the terms and multiplying a factor of 2 on both sides, the inequality (51) is equivalent to the following:

$$\sqrt{(\mathrm{Var}[f(X)] + \mathrm{Var}[g(X)])^2 \quad 4(\mathrm{Cov}(f(X), g(X))^2}$$

$$\mathrm{Var}[f(X)] \quad \mathrm{Var}[g(X)] \quad 2\,\mathrm{Var}\,\mathbb{E}[f(X)|Z] + 2\,\mathrm{Var}\,\mathbb{E}[g(X)|Z]$$

$$= \quad \mathrm{Var}[f(X)] + \mathrm{Var}[g(X)] + 2\mathbb{E}\,\mathrm{Var}(f(X)|Z) \quad 2\mathbb{E}\,\mathrm{Var}(g(X)|Z)$$

In the last step, we used the law of total variance. Similarly, the inequality (52) is equivalent to:

$$\sqrt{(\mathrm{Var}[f(X)] + \mathrm{Var}[g(X)])^2 \quad 4(\mathrm{Cov}(f(X), g(X))^2}$$

$$\mathrm{Var}[f(X)] \quad \mathrm{Var}[g(X)] \quad 2\mathbb{E}\,\mathrm{Var}(f(X)|Z) + 2\mathbb{E}\,\mathrm{Var}(g(X)|Z)$$

Therefore, it suffices to prove that

$$\sqrt{(\mathrm{Var}[f(X)] + \mathrm{Var}[g(X)])^2 \quad 4(\mathrm{Cov}(f(X), g(X))^2}$$

$$\mathrm{Var}[f(X)] \quad \mathrm{Var}[g(X)] \quad 2\mathbb{E}\,\mathrm{Var}(f(X)|Z) + 2\mathbb{E}\,\mathrm{Var}(g(X)|Z) \tag{53}$$

Let $p(X) = f(X) + g(X)$ and $q(X) = f(X) \quad g(X)$. Notice that

$$\mathrm{Var}[f(X)] \quad \mathrm{Var}[g(X)] = \mathrm{Cov}(p(X), q(X))$$
$$\mathrm{Var}(f(X)|Z) \quad \mathrm{Var}(g(X)|Z) = \mathrm{Cov}(p(X), q(X)|Z)$$

and

$$(\mathrm{Var}[f(X)] + \mathrm{Var}[g(X)])^2 \quad 4\,\mathrm{Cov}(f(X), g(X))^2 = \mathrm{Var}[p(X)]\,\mathrm{Var}[q(X)].$$

The inequality (53) can be further simplified to

$$\sqrt{\mathrm{Var}[p(X)]\,\mathrm{Var}[q(X)] \quad \mathrm{Cov}(p(X), q(X)) \quad 2\mathbb{E}\,\mathrm{Cov}(p(X), q(X)|Z).}$$

When $p(X) = c$ or $q(X) = c$ for some constant c a.s., both sides equal to zero, hence the inequality obviously holds.

For the other cases, we use the following simple observation: $B \quad \sqrt{AC}$, $A > 0, C > 0$, $\inf_{t \in \mathbb{R}} At^2 + 2Bt + C \quad 0$. Therefore, we only need to prove that for any $t \in \mathbb{R}$,

$$t^2\mathrm{Var}[p(X)] + 2t(\mathrm{Cov}(p(X), q(X)) \quad 2\mathbb{E}\,\mathrm{Cov}(p(X), q(X)|Z)) + \mathrm{Var}[q(X)] \quad 0$$

In fact, the inequality we wanted to prove can be re-written as:

$$\mathrm{Var}[t\,p(X) + q(X)] \quad 4\mathbb{E}\,\mathrm{Cov}(t\,p(X), q(X)|Z) \quad 0. \tag{54}$$

Notice that

$$4\,\mathrm{Cov}(t\,p(X), q(X)|Z) = \mathrm{Var}[t\,p(X) + q(X)|Z] \quad \mathrm{Var}[t\,p(X) \quad q(X)|Z] \tag{55}$$

We have

$$\begin{aligned}
&\mathrm{Var}[t\,p(X)+q(X)] \quad 4\mathbf{E}\,\mathrm{Cov}(t\,p(X),q(X)|Z)\\
&=(\mathrm{Var}[t\,p(X)+q(X)] \quad \mathbf{E}\,\mathrm{Var}[t\,p(X)+q(X)|Z])+\mathbf{E}\,\mathrm{Var}[t\,p(X) \quad q(X)|Z]\\
&=\mathrm{Var}(\mathbf{E}[t\,p(X)+q(X)|Z])+\mathbf{E}\,\mathrm{Var}[t\,p(X) \quad q(X)|Z]\\
&0,
\end{aligned}$$

where in the second to last step we used the law of total variance. Therefore we have completed the proof. ■

Here are two corollaries of the lemma that will be useful for proving Theorem C.1 and C.2.

Corollary C.3 Suppose the random variable Z satisfies $\mathrm{Var}(\mathbf{E}[g(x)|Z]) = 0$, then we have the following upper bound:

$$\mathrm{Var}\,\mathbf{E}[f(X)|Z] \quad \mathrm{Var}[f(X)] \quad \frac{\mathrm{Cov}(f(\ ),g\ )^2}{\mathrm{Var}[g(X)]}. \tag{56}$$

Proof By applying the second inequality in Lemma C.4 to functions f and $^p / g$, for any $/ \quad 0$,

$$\frac{1}{2}\left[\mathrm{Var}\,\mathbf{E}[f(X)|Z] \quad / \,\mathrm{Var}\,\mathbf{E}[g(X)|Z] \quad \sqrt{(\mathrm{Var}[f(X)]+/\,\mathrm{Var}[g(X)])^2 \quad 4/\,\mathrm{Cov}(f(X),g(X))^2}\right] \tag{57}$$

Since $\mathrm{Var}\,\mathbf{E}[g(X)|Z] = 0$, Equation (57) is equivalent to:

$$\begin{aligned}
&\frac{1}{2}\left[\mathrm{Var}\,\mathbf{E}[f(X)|Z] \quad / \,\mathrm{Var}[g(X)] + \sqrt{(\mathrm{Var}[f(X)]+/\,\mathrm{Var}[g(X)])^2 \quad 4/\,\mathrm{Cov}(f(X),g(X))^2}\right]\\
&=\frac{2/\,\mathrm{Var}[f(X)]\,\mathrm{Var}[g(X)] \quad 2/\,\mathrm{Cov}(f(X),g(X))^2}{\mathrm{Var}[f(X)]+/\,\mathrm{Var}[g(X)]+\sqrt{(\mathrm{Var}[f(X)]+/\,\mathrm{Var}[g(X)])^2 \quad 4/\,\mathrm{Cov}(f(X),g(X))^2}},
\end{aligned} \tag{58}$$

Taking the limit $/ \ ! \ ¥$ completes the proof. ■

Corollary C.4 Suppose the random variable Z satisfies $\mathrm{Var}(\mathbf{E}[g(x)|Z]) = \mathrm{Var}(g(x))$, then we have the following lower bound:

$$\mathrm{Var}\,\mathbf{E}[f(X)|Z] \quad \frac{\mathrm{Cov}(f(\ ),g\ ))^2}{\mathrm{Var}[g(X)]} \tag{59}$$

Proof By applying the first inequality in Lemma C.4 to functions f and $^p / g$, for any $/ \quad 0$,

$$\frac{1}{2}\left[\mathrm{Var}\,\mathbf{E}[f(X)|Z] \quad / \,\mathrm{Var}\,\mathbf{E}[g(X)|Z] \quad \sqrt{(\mathrm{Var}[f(X)]+/\,\mathrm{Var}[g(X)])^2 \quad 4/\,\mathrm{Cov}(f(X),g(X))^2}\right] \tag{60}$$

Since $\mathrm{Var}(\mathbf{E}[g(X)|Z]) = \mathrm{Var}[g(X)]$, Equation (60) is equivalent to:

$$\mathbf{E}\,\mathrm{Var}(f(X)|Z) \le \frac{1}{2}\left\{\mathrm{Var}[f(X)] + \beta\,\mathrm{Var}[g(X)] - \sqrt{(\mathrm{Var}[f(X)] + \beta\,\mathrm{Var}[g(X)])^2 - 4\beta\,\mathrm{Cov}(f(X), g(X))^2}\right\} = \mathrm{Var}[f(X)] + \beta\,\mathrm{Var}[g(X)] + \sqrt{(\mathrm{Var}[f(X)] + \beta\,\mathrm{Var}[g(X)])^2 - 4\beta\,\mathrm{Cov}(f(X), g(X))^2}}{2}, \tag{61}$$

Taking the limit $\beta \to \infty$ completes the proof. ■

Now we are ready to prove Theorem C.1, C.2 and C.3.

Theorem C.1 The optimal solution of optimization problem (12) is upper bounded by

$$\max_{Z,\,\mathrm{Var}\mathbf{E}[A|Z]=0} \mathrm{Var}\,\mathbf{E}[Y|Z] \le \mathrm{Var}\,\mathbf{E}[Y|X] - C\frac{\mathrm{Cov}^2(\mathbf{E}[\quad], \mathbf{E}[A|\quad])}{\mathrm{Var}\mathbf{E}[A|X]\,\mathrm{Var}\mathbf{E}[Y|X]}. \tag{31}$$

Proof [Proof of Theorem C.1] We apply corollary C.3 by choosing $f := f_Y$ and $g := f_A$, this gives us the following lower bound when $\mathrm{Var}(\mathbf{E}[f_A(X)|Z]) = 0$

$$\mathrm{Var}(\mathbf{E}[f_Y(X)|Z]) \le \mathrm{Var}[f_Y(X)] - \frac{\mathrm{cov}(f_Y(X), f_A(X))^2}{\mathrm{Var}[f_A(X)]} = \mathrm{Var}[f_Y(X)](1 - r_{YA}^2).$$

Therefore we have completed the proof. ■

Theorem C.2 The optimal solution of optimization problem (13) is lower bounded by

$$\min_{Z,\,\mathrm{Var}\mathbf{E}[Y|Z]=\mathrm{Var}\mathbf{E}[Y|X]} \mathrm{Var}\,\mathbf{E}[A|Z] \ge \frac{\mathrm{Var}\,\mathbf{E}[Y|X]\,\langle a, y\rangle^2}{\langle y, y\rangle^2}. \tag{34}$$

Proof [Proof of Theorem C.2] We apply corollary C.3 by choosing $f := f_A$ and $g := f_Y$, this gives us the following lower bound when $\mathrm{Var}(\mathbf{E}[f_Y(X)|Z]) = \mathrm{Var}[f_Y(X)]$

$$\mathrm{Var}(\mathbf{E}[f_A(X)|Z]) \le \frac{\mathrm{Cov}(f(X), g(X))^2}{\mathrm{Var}[g(X)]} = \mathrm{Var}[f_Y(X)]r_{YA}^2.$$

Therefore we have completed the proof. ■

Theorem C.3 The optimal solution of the Lagrangian has the following lower bound:

$$\mathrm{OPT}(\beta) \ge \frac{1}{2}\left\{-\beta\,\mathrm{Var}\,\mathbf{E}[A|X] - \mathrm{Var}\,\mathbf{E}[Y|X] - \sqrt{(\beta\,\mathrm{Var}\,\mathbf{E}[A|X] + \mathrm{Var}\,\mathbf{E}[Y|X])^2 - 4\beta\,\mathrm{Cov}^2(\mathbf{E}[Y|X], \mathbf{E}[A|X])}\right\}.$$

Proof [Proof of Theorem C.3]

The desired inequality follows directly from applying Lemma C.4 above to $f := f_Y$ and $g := \sqrt{\beta} f_A$. ■