

Contrastive learning, multi-view redundancy, and linear models

Christopher Tosh

*Data Science Institute
Columbia University
New York, NY 10027*

C.TOSH@COLUMBIA.EDU

Akshay Krishnamurthy

*Microsoft Research
New York, NY 10011*

AKSHAYKR@MICROSOFT.COM

Daniel Hsu

*Department of Computer Science and Data Science Institute
Columbia University
New York, NY 10027*

DJHSU@CS.COLUMBIA.EDU

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

Self-supervised learning is an empirically successful approach to unsupervised learning based on creating artificial supervised learning problems. A popular self-supervised approach to representation learning is contrastive learning, which leverages naturally occurring pairs of similar and dissimilar data points, or multiple views of the same data. This work provides a theoretical analysis of contrastive learning in the multi-view setting, where two views of each datum are available. The main result is that linear functions of the learned representations are nearly optimal on downstream prediction tasks whenever the two views provide redundant information about the label.

Keywords: semi-supervised learning, contrastive learning, representation learning

1. Introduction

Self-supervised learning has emerged as a popular and empirically successful class of methods for learning useful representations from unlabeled data. Broadly speaking, self-supervised learning refers to techniques that take advantage of naturally occurring structure in unlabeled data to create artificial supervised learning problems and then solves these problems using machine learning methods such as deep learning. The hope is that in solving these problems, a learning algorithm will also create internal representations for data that are useful for other downstream learning tasks. Self-supervised techniques include de-noising autoencoders (Vincent et al., 2008), image inpainting (Pathak et al., 2016), and the focus of this work, contrastive learning (Hadsell et al., 2006; Oord et al., 2018; Logeswaran and Lee, 2018; Hjelm et al., 2018; Arora et al., 2019; Bachman et al., 2019; Tian et al., 2019; Tosh et al., 2020; Chen et al., 2020).

A common theme among many of these self-supervised representation learning works is the exploitation of naturally occurring similar points, or multiple views of the same data points. To train a de-noising autoencoder, one first creates an alternate “view” of data points by corrupting them with added noise and then trains the autoencoder to reconstruct the original. Image inpainting removes patches of images and trains models to reconstruct the original image. Contrastive learning trains

models to distinguish naturally occurring similar pairs of points, such as neighboring sentences (Logeswaran and Lee, 2018) or randomly cropped and blurred versions of the same image (Chen et al., 2020), from random pairs of points.

However, exploiting multiple views of data for representation learning is not a new technique. Canonical correlation analysis (CCA) (Hotelling, 1936) is a classical (unsupervised) technique that finds the linear transformation that aligns two views of data so that the resulting coordinates are uncorrelated. A fascinating line of work (Ando and Zhang, 2005; Kakade and Foster, 2007; Ando and Zhang, 2007; Foster et al., 2009) investigated the quality of representations produced by CCA (and related linear methods) for downstream regression problems. Most relevant to the current work, Kakade and Foster (2007) and Foster et al. (2009) demonstrated that linear regression with the CCA representation in a low-dimensional space will have near-optimal performance relative to the best linear function of the original representation when there is some redundancy among the two views, i.e., whenever the best linear prediction of the label on each individual view is nearly as good as the best linear prediction of the label when both views are used together.

In this work, we examine contrastive learning from the perspective of multi-view redundancy, analogously to the CCA analysis of Kakade and Foster (2007) and Foster et al. (2009). We show that when there is some redundancy between the two views on the label, contrastive learning leads to representations such that *linear* functions of these representations are competitive with the (possibly *non-linear*) Bayes optimal predictor of the label. Our analysis is rather general, and we give bounds on the dimensionality of the representations that is sufficient to lead to good performance in the downstream prediction task. We consider two specific representations based on contrastive learning. The first of these is a general-purpose construction that uses the “landmark embedding” technique of Tosh et al. (2020). The second representation is formed by solving a particular bivariate optimization problem. In both cases, we show that we can use low-dimensional representations and still achieve near-optimal downstream performance with linear methods. We instantiate our results in some simple latent variable models for illustration.

1.1. Overview of results

In the multi-view setting, data points are represented as triples of random variables (X, Z, Y) where X and Z represent two views of the data and Y is some label or regression value to be predicted. “Views” should be interpreted liberally here. For example, they could correspond to the first and second halves of a document or to two different distortions of the same image. However, the main property that we will require of our views is that they share redundant information with respect to predicting Y . That is, predicting Y from X or Z individually should be nearly as accurate as predicting Y from X and Z together.

When X and Z do satisfy this redundancy property, we show that there is a surprisingly effective prediction strategy: given X , first try to infer Z and then predict Y based only on the inferred Z . Specifically, we prove the following lemma.

Lemma 1 [Restated] *If X, Z, Y are random variables, then*

$$\mathbb{E} \left[(\mathbb{E}[\mathbb{E}[Y | Z] | X] - \mathbb{E}[Y | X, Z])^2 \right] \leq \varepsilon_X + 2\sqrt{\varepsilon_X \varepsilon_Z} + \varepsilon_Z$$

where $\varepsilon_W = \mathbb{E} \left[(\mathbb{E}[Y | W] - \mathbb{E}[Y | X, Z])^2 \right]$ for each $W \in \{X, Z\}$.

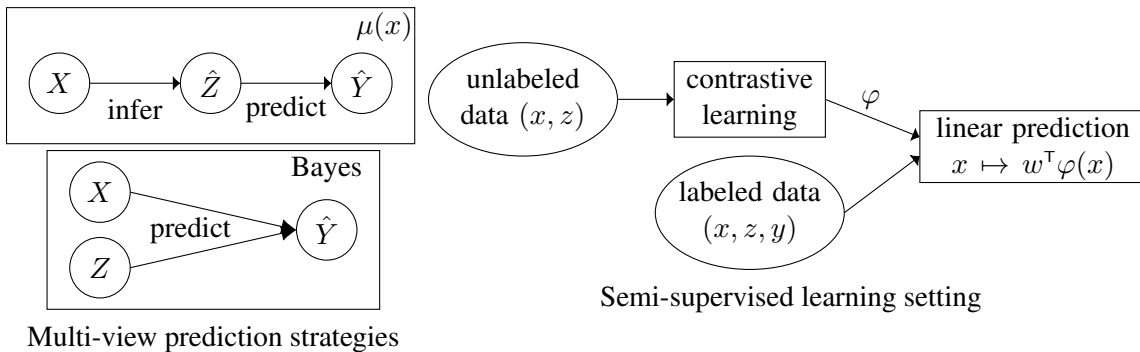


Figure 1: Left: Two prediction strategies in the multi-view setup. When the views are redundant, $\mu(x)$ (top) is competitive with the Bayes optimal predictor (bottom). Right: our semi-supervised setting. The representation φ is trained in an unsupervised fashion using the two views (x, z) . Then a linear predictor of features φ is trained using labeled data.

The strategy of Lemma 1, illustrated in Figure 1 (left), is reminiscent of the information bottleneck method (Tishby et al., 1999), in which predicting Y from X is done by first compressing X to a “smaller” representation \hat{X} and then predicting Y using \hat{X} . In our case, the separate view Z acts as a natural intermediate target instead of \hat{X} . As it will turn out, Lemma 1 is the basis of all of the following results.

It is also worth mentioning that the quantities ε_X and ε_Z in Lemma 1 can be bounded as functions of the conditional mutual information of X , Y , and Z (Tao, 2006; Ahlswede, 2007; Wu and Verdú, 2011). Specifically, Wu and Verdú (2011) demonstrated that when $Y \in [-1, 1]$, we have

$$\mathbb{E} \left[\left(\mathbb{E}[Y | X] - \mathbb{E}[Y | X, Z] \right)^2 \right] \leq \frac{1}{2} I(Y; Z | X)$$

where $I(Y; Z | X)$ is the mutual information of Y and Z conditioned on X . A symmetric inequality also holds when X and Z are swapped. Combining this result with Lemma 1, the performance of the strategy that infers Z from X and then predicts Y based only on this inferred Z can also be bounded as a function of $I(Y; Z | X)$ and $I(Y; X | Z)$.

We are primarily concerned with the setting where we have lots of unlabeled data, i.e., (X, Z) pairs, and rather less labeled data (X, Z, Y) . In such situations, one natural strategy is to use the unlabeled data to learn a representation of X (or Z or (X, Z)), and then use the small collection of labeled data to learn a simple function, like a linear predictor, on top of this new representation. The setting is displayed in Figure 1 (right). In this work, we will look at the specific representation learning algorithm posed by Tosh et al. (2020), which is a type of contrastive learning algorithm.

The approach of Tosh et al. (2020) is to learn a function f that distinguishes between true data points (X, Z) , and fabricated data points (X, \tilde{Z}) , where X and \tilde{Z} come from independently sampled data points (X, Z) and (\tilde{X}, \tilde{Z}) . The idea is that such a function f will learn enough about the relationship between X and Z to allow us to predict the label Y from X through Z as in Lemma 1.

In Section 3, we show that one can extract an embedding of a view X from the learned function f such that linear functions on top of this embedding are competitive with the best predictor of Y from X . Moreover, this embedding will be the same one proposed by Tosh et al. (2020) which uses landmark views Z_1, \dots, Z_m that are i.i.d. copies of Z and embeds a point x with the prediction values $f(x, Z_1), \dots, f(x, Z_m)$.

Theorem 3 [Restated] Given a solution $f^*: \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}$ to the contrastive learning problem and embedding points Z_1, \dots, Z_m sampled i.i.d. from the marginal distribution of Z , the landmark embedding $\varphi^*: \mathcal{X} \rightarrow \mathbb{R}$ based on f^* and Z_1, \dots, Z_m defined in Section 3 satisfies, with high probability,

$$\min_{w \in \mathbb{R}^m} \mathbb{E} \left[(w^\top \varphi^*(X) - \mathbb{E}[Y | X, Z])^2 \right] \leq \varepsilon_X + 2\sqrt{\varepsilon_X \varepsilon_Z} + \varepsilon_Z + O_m(1/m).$$

To avoid clutter, we have used big- O_m in the above statement to suppress all factors that do not depend on m , including logarithmic factors in the failure probability and other quantities that depend on the distribution of X and Z . The full statement is provided in Section 3.

While the results of Section 3 demonstrate that contrastive learning can lead to useful representations in the presence of redundancy, the landmark embedding technique is not reflective of what is done in practice. In practice, self-supervised representation learning algorithms typically optimize embedding functions directly (Hadsell et al., 2006; Chen et al., 2020). To address this, in Section 4, we investigate the strategy of trying to learn the embedding functions directly. That is, we look at the *bivariate architecture* setting where we learn \mathbb{R}^m -valued functions η, ψ such that $\eta(x)^\top \psi(z)$ distinguishes between the real and fake data points. As we will see, the benefit of this approach is that when there exist pairs of accurate functions, η (and also ψ) allow us to do useful linear predictions. However, it is unclear *a priori* how large the output dimension of η and ψ needs to be to achieve this.

As a first step towards understanding this dimensionality question, we consider the setting where there is some hidden variable H that renders the two views X and Z conditionally independent. Note that there is always a trivial random variable to achieve this, namely $H = (X, Z)$. However, we show that when the hidden variable obeys a nicer structure, the dimensionality of the embedding can be drastically improved. Specifically, we show the following.

- (a) When the hidden variable H takes values in a finite set, the cardinality of this set is an upper bound on the dimensionality needed for an *exact* embedding.
- (b) In the general setting, there exist approximate embeddings where the approximation factor decreases at a rate of $O_m(1/m)$, where the big- O_m notation suppresses dependence on a particular variance quantity of the hidden variable structure.

Importantly, there is no assumption that the hidden variable structure is known. Rather, our results imply that solving the bivariate contrastive learning problem automatically recovers embeddings whose performance can be bounded by factors that depend on the underlying hidden variables.

Finally, in Section 5, we analyze how errors in optimizing the contrastive objectives propagate to the performance of these representations on downstream linear prediction tasks. We investigate this error propagation for both the landmark and direct embeddings, and we show that the downstream prediction risk has a smooth relationship with the excess contrastive loss.

Along the way, we illustrate these results with simplified running examples of a topic model (Blei et al., 2003) and a Gaussian latent variable model. However, our results are applicable to other multi-view settings, including co-training, certain mixture models, hidden Markov models, and phylogenetic tree models (e.g., Blum and Mitchell, 1998; Dasgupta et al., 2002; Mossel and Roch, 2005; Chaudhuri et al., 2009; Allman et al., 2009; Anandkumar et al., 2012).

Additionally, in the appendix we consider the transfer learning scenario, where the unlabeled distribution used for representation learning is not the same as the test distribution used for downstream prediction. We show that in certain settings, the landmark embeddings can be “fine-tuned” to work under the test distribution.

1.2. Related work

A number of recent works have sought to theoretically explain the success of contrastive learning specifically, and self-supervised learning more generally. [Arora et al. \(2019\)](#) presented a theoretical treatment of contrastive learning that considered the specific setting of trying to minimize the loss $L(\phi) = \mathbb{E}[\ell(\phi(X)^\top(\phi(X_+) - \phi(X_-)))]$, where (X, X_+) is a random “positive” pair, (X, X_-) is a random “negative” pair, and ℓ is a binary classification loss such hinge or logistic loss. They showed that if there is an underlying collection of latent classes and positive examples are generated by draws from the same class, then minimizing the contrastive loss over embedding functions ϕ yields good representations for distinguishing latent classes with linear models.

In work concurrent with the present paper, [Lee et al. \(2020\)](#) considered a self-supervised scheme in which two views (X, Z) are available for each data point, and the representation learning objective is a reconstruction error of Z based on a function of X : $L(\phi) = \mathbb{E}\|Z - \phi(X)\|^2$. (They assume that Z takes values in a suitable normed space.) They showed that if the two views are approximately independent conditioned on the label, then linear functions of the learned representation are capable of predicting the label. This approach resembles the representation learning methods of [Ando and Zhang \(2005, 2007\)](#), as well as methods for learning predictive state representations dynamical systems ([Littman and Sutton, 2002](#); [Hsu et al., 2009](#); [Langford et al., 2009](#); [Song et al., 2010](#)). The self-supervised problem we study is instead a classification problem rather than a (possibly multidimensional output) regression problem.

Most relevant to the current work, [Tosh et al. \(2020\)](#) also considered the problem of contrastive learning under certain generative assumptions. Specifically, they showed that when the two views of the data point correspond to random partitions of a document, contrastive learning recovers information related to the underlying topics that generated the document. The contrastive learning problem they study is also a classification problem rather than a regression problem.

Also related is the use of self-supervised learning for exploration in a model for reinforcement learning called Block MDPs ([Du et al., 2019](#); [Misra et al., 2020](#)). In these settings, self-supervised learning is used to derive decoders of unobserved latent state from observations. The analyses in these works apply to cases where exact decoding of the state is possible. In particular, the method studied by [Misra et al. \(2020\)](#) uses a contrastive learning objective similar to the one we analyze. In this paper, we study an example that resembles the Block MDP, but our analysis applies more broadly to scenarios where latent variables cannot be perfectly decoded from the observations.

The contrastive estimation technique we study, now known as “Noise Contrastive Estimation” (NCE; [Gutmann and Hyvärinen, 2010](#)), was also theoretically analyzed in other contexts, including density level set estimation ([Steinwart et al., 2005](#); [Abe et al., 2006](#)), parametric estimation ([Gutmann and Hyvärinen, 2010](#); [Ma and Collins, 2018](#)), and nonlinear ICA ([Hyvarinen and Morioka, 2016, 2017](#); [Hyvarinen et al., 2019](#)). Although the setups in these works do not consider the use of a learned representation in a downstream task, NCE has inspired many empirical works that use the technique in this way. The primary motivation for NCE given in these works is the relationship between NCE and maximizing mutual information (e.g., [Oord et al., 2018](#); [Hjelm et al.,](#)

2018; Bachman et al., 2019; Tian et al., 2019), and the usefulness of the learned representation is attributed to this connection. Although this connection also makes an appearance in our work, it is subordinate to multi-view redundancy in our analysis. McAllester and Stratos (2020) highlight some limitations on measuring mutual information in these contexts, which raises some doubt that this mutual information perspective can solely explain the success of NCE. Other doubts about the mutual information perspective are raised by Tschannen et al. (2019).

2. Contrastive learning

In this section, we formalize contrastive learning in the multi-view setting and introduce the redundancy assumption that is key to our analysis.

2.1. Multi-view data distribution and notation

We consider the multi-view setting, in which data points take the form $(x, z, y) \in \mathcal{X} \times \mathcal{Z} \times \mathbb{R}$, for some pair of data spaces $(\mathcal{X}, \mathcal{Z})$. Here x and z refer to the separate views of the data point, and y refers to its label or regression value. We assume that there is some distribution over (x, z, y) triples, and we denote the corresponding random variables with capital letters (X, Z, Y) . For simplicity, we assume these random variables have either (joint) probability mass functions or probability density functions, and denote them by their corresponding letters, e.g. $p_{X,Z,Y}, p_X$, etc. At various points, we will introduce a hidden variable H , and there we shall use $p_{H|X}$ to denote the conditional distribution of H given X . For instance, $p_{H|X}(h | x) = \Pr(H = h | X = x)$ when H is discrete. Finally, we use $p \otimes q$ to denote the product distribution with marginals p and q .

Our main interest is in the semi-supervised learning setting, in which we have both *unlabeled data* from $\mathcal{X} \times \mathcal{Z}$, typically modeled as i.i.d. copies of (X, Z) , as well as *labeled data* from $\mathcal{X} \times \mathcal{Z} \times \mathbb{R}$, modeled as i.i.d. copies of (X, Z, Y) . In many cases, the unlabeled data are plentiful, whereas the labeled data are very few due to the cost of obtaining labels. In this setting, we will use contrastive learning on the unlabeled data to learn a representation that ultimately simplifies the downstream supervised learning task which uses the labeled data. (In particular, the downstream supervised learning will be accomplished using just *linear* predictors that, we prove, are competitive even with non-linear predictors.)

2.2. Contrastive distribution

Following Tosh et al. (2020), define the *contrastive distribution* $\mathcal{D}_{\text{contrast}}$ via the following process:

- Let (\tilde{X}, \tilde{Z}) be an independent copy of (X, Z) , so $(X, Z), (\tilde{X}, \tilde{Z}) \stackrel{\text{iid}}{\sim} p_{X,Z}$.
- Independently toss a fair coin; if heads, output $(X, Z, 1)$; otherwise, output $(X, \tilde{Z}, -1)$.

We let $(X_c, Z_c, Y_c) \sim \mathcal{D}_{\text{contrast}}$. Note that Y_c has nothing to do with the random variable Y ; it is simply the outcome of the fair coin in the generative process above. Therefore, sampling from $\mathcal{D}_{\text{contrast}}$ can be accomplished using the process described above as long as one can sample from $p_{X,Z}$ —the distribution of unlabeled data. In practice, this process provides a way to create a self-supervised data set of (x_c, z_c, y_c) triples using only unlabeled data.

2.3. Contrastive learning problem

The goal of the contrastive learning problem is to find a predictor of Y_c from (X_c, Z_c) ; that is, predict whether the two views (X_c, Z_c) are from the same data point (i.e., $(X_c, Z_c) \sim p_{X,Z}$), or from two independent data points (i.e., $(X_c, Z_c) \sim p_X \otimes p_Z$). This is a binary classification problem, and a standard approach for doing this is to minimize the expected logistic loss between the label Y_c and our prediction $f(X_c, Z_c)$. Formally, we solve the following optimization problem:

$$f^* \in \operatorname{argmin}_{f: \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}} L_{\text{lm}}(f), \quad L_{\text{lm}}(f) := \mathbb{E} \left[\log \left(1 + \exp \left(-Y_c f(X_c, Z_c) \right) \right) \right]. \quad (1)$$

(The lm subscript is for ‘‘landmark,’’ which will be discussed later.) Note that the optimal solution f^* to (1) (over all functions from $\mathcal{X} \times \mathcal{Z}$ to \mathbb{R}) predicts the pointwise mutual information between two views:

$$f^*(x, z) = \log \frac{p_{X,Z}(x, z)}{p_X(x)p_Z(z)}.$$

Given f^* , it is easy to compute the density ratio of the joint distribution of X and Z and the product of their marginals:

$$g^*(x, z) := \exp(f^*(x, z)) = \frac{p_{X,Z}(x, z)}{p_X(x)p_Z(z)}.$$

Of course, in practice, we typically cannot minimize (1) over all (measurable) functions directly. Instead, we may use an empirical approximation to the objective based on a finite sample, and (attempt to) minimize this approximation over a particular class of functions (e.g., neural networks). The issues that arise from such discrepancies are important, but a detailed study is beyond the scope of the present work. In Section 5, we analyze how errors introduced due to these concerns can affect performance in the downstream supervised learning task.

2.4. Redundancy

We will assume that for the task of predicting Y , there is a certain amount of redundancy between X and Z . That is, the quantities

$$\varepsilon_X := \mathbb{E} \left[\left(\mathbb{E}[Y | X] - \mathbb{E}[Y | X, Z] \right)^2 \right] \quad \text{and} \quad \varepsilon_Z := \mathbb{E} \left[\left(\mathbb{E}[Y | Z] - \mathbb{E}[Y | X, Z] \right)^2 \right]$$

are both small. As our results hold for any values of ε_X and ε_Z , we opt not to formalize the redundancy assumption at a particular scale. However, our results are most meaningful in those cases where ε_X and ε_Z are small. We also stress that these predictors ($x \mapsto \mathbb{E}[Y | X = x]$, $z \mapsto \mathbb{E}[Y | Z = z]$, $(x, z) \mapsto \mathbb{E}[Y | (X, Z) = (x, z)]$) are *not* assumed to be linear.

Intuitively, when the redundancy assumption holds, one should be able to get a good prediction on Y by first predicting Z from X , and then using the resulting information to predict Y . This strategy is formalized by the following function (see the left panel of Figure 1):

$$\mu(x) := \mathbb{E}[\mathbb{E}[Y | Z] | X = x].$$

The following lemma tells us that this strategy does indeed work.

Lemma 1 $\mathbb{E}[(\mu(X) - \mathbb{E}[Y | X, Z])^2] \leq \varepsilon_X + 2\sqrt{\varepsilon_X \varepsilon_Z} + \varepsilon_Z =: \varepsilon_\mu.$

Due to space constraints, all proofs are deferred to the appendix.

The function μ is related to our contrastive function g^* in the following way:

$$\mu(x) = \int \mathbb{E}[Y | Z = z] \frac{p_{X,Z}(x, z)}{p_X(x)} dz = \int \mathbb{E}[Y | Z = z] g^*(x, z) p_Z(z) dz.$$

Thus, g^* provides the change-of-measure from the marginal distribution of Z to the conditional distribution of Z given $X = x$. Note that g^* depends only on (the distributions of) X and Z ; it does not depend on Y at all. Therefore g^* is useful for all prediction targets Y for which the redundancy assumption holds.

3. Landmark embedding representations

How should one use g^* to produce a finite dimensional embedding of $x \in \mathcal{X}$? When g^* is implemented using a neural network, a common approach is construct a mapping defined by some internal hidden units (e.g., removing the top layer or two of the network). How well the resulting embedding fares in downstream tasks may depend on details of the implementation, such as the specific architecture and connection weights.

A different, and generic, approach to constructing an embedding from g^* , proposed by [Tosh et al. \(2020\)](#), is to use the external behavior of g^* on a random sample Z_1, \dots, Z_m of views, called *landmarks*, and embed according to

$$\varphi^*(x) := (g^*(x, Z_1), \dots, g^*(x, Z_m)).$$

We assume Z_1, \dots, Z_m are taken from i.i.d. copies $(X_1, Z_1, Y_1), \dots, (X_m, Z_m, Y_m)$ of (X, Z, Y) . In practice, these landmarks can be taken from a random sample of m unlabeled data points. (One can, of course, construct an embedding of $z \in \mathcal{Z}$ symmetrically.)

To gain intuition on why this approach is sound, note that if $w \in \mathbb{R}^m$ satisfies $w_i = \frac{1}{m} \mathbb{E}[Y_i | Z_i]$, then as $m \rightarrow \infty$ we have

$$w^\top \varphi^*(x) = \frac{1}{m} \sum_{i=1}^m \mathbb{E}[Y_i | Z_i] g^*(x, Z_i) \rightarrow \int \mathbb{E}[Y | Z = z] g^*(x, z) p_Z(z) dz = \mu(x).$$

Thus, in the limit, $\varphi^*(x)$ provides a useful representation for downstream *linear* prediction under the redundancy assumption. In this section, we analyze the approximation error that arises when we are restricted to a finite number of landmarks.

3.1. Landmark embedding error

The following lemma quantifies the error from using finite-dimensional landmark embeddings.

Lemma 2 *Let $(X_1, Z_1, Y_1), \dots, (X_m, Z_m, Y_m), (X, Z, Y)$ be i.i.d., and define φ^* using the landmarks Z_1, \dots, Z_m . With probability $1 - \delta$, there exists a weight vector $w \in \mathbb{R}^m$ such that*

$$\mathbb{E}[(w^\top \varphi^*(X) - \mu(X))^2 | Z_1, \dots, Z_m] \leq \varepsilon_{\text{lm}}, \quad \text{where} \quad \varepsilon_{\text{lm}} := \frac{2 \text{var}(\mathbb{E}[Y_1 | Z_1] g^*(X, Z_1))}{\lfloor m / \log_2(1/\delta) \rfloor}.$$

The upshot of Lemma 2 is that to get an embedding that can linearly approximate μ within squared loss error of ϵ , it suffices to embed using no more than

$$O\left(\frac{\text{var}(\mathbb{E}[Y_1 | Z_1]g^*(X, Z_1))}{\epsilon}\right)$$

landmarks. Thus, we have the following immediate consequence of Lemma 1 and Lemma 2 (and the AM/GM inequality).

Theorem 3 *Let $(X_1, Z_1, Y_1), \dots, (X_m, Z_m, Y_m), (X, Z, Y)$ be i.i.d., and define φ^* using the landmarks Z_1, \dots, Z_m . With probability $1 - \delta$, there exists a weight vector $w \in \mathbb{R}^m$ such that*

$$\mathbb{E}[(w^\top \varphi^*(X) - \mathbb{E}[Y | X, Z])^2 | Z_1, \dots, Z_m] \leq \epsilon_\mu + 2\sqrt{\epsilon_\mu \epsilon_{\text{lm}}} + \epsilon_{\text{lm}}$$

where ϵ_μ is defined in Lemma 1 and ϵ_{lm} is defined in Lemma 2.

3.2. Topic model

We now turn to a simple topic modeling example to illustrate the error bound.

Let P_1, \dots, P_K be distributions over a finite vocabulary \mathcal{V} . Each distribution corresponds to a topic, and we assume that their supports are disjoint; this is similar to the setting studied by [Papadimitriou et al. \(2000\)](#). For simplicity, suppose each document is exactly two tokens long, so that the two views (X, Z) are individual tokens (and $\mathcal{X} = \mathcal{Z} = \mathcal{V}$). We assume these tokens are drawn from a random mixture of P_1, \dots, P_K , where the mixing weights are drawn from a symmetric Dirichlet distribution with parameter $\alpha > 0$ (following the LDA model of [Blei et al., 2003](#)). Thus the generative model for a single document is:

- Draw $\Theta = (\Theta_1, \dots, \Theta_K) \sim \text{Dirichlet}(\alpha)$.
- Given Θ , draw X and Z independently from the mixture distribution $\sum_{k=1}^K \Theta_k P_k$.

Proposition 4 *Assume that Y takes values in $[-1, 1]$. In the topic model setting,*

$$\epsilon_{\text{lm}} \leq \begin{cases} O(\log(1/\delta)/m) & \text{if } \alpha = \Theta(1) \text{ as } K \rightarrow \infty; \\ O(K^2 \log(1/\delta)/m) & \text{if } \alpha \leq 1/K. \end{cases}$$

The $\alpha = \Theta(1)$ and $\alpha \leq 1/K$ correspond to the “non-sparse” and “sparse” regimes of LDA; here, sparsity is considered in an approximate sense ([Telgarsky, 2013](#)).

3.3. Gaussian model

As another example, we consider a simple multi-view Gaussian latent variable model

$$H \sim \mathcal{N}(0, \sigma^2), \quad X | H \sim \mathcal{N}(H, 1), \quad Z | H \sim \mathcal{N}(H, 1),$$

and we assume $X \perp\!\!\!\perp Z | H$.

Proposition 5 *Assume Y takes values in $[-1, 1]$. In the Gaussian model setting, for any $\sigma^2 > 0$,*

$$\varepsilon_{\text{lm}} = O\left(\frac{\log(1/\delta)}{m}\right) \cdot \left(1 + \frac{\sigma^4}{1 + 2\sigma^2}\right).$$

The analysis shows that the variance of H captures the difficulty of obtaining a low-dimensional representation in this model. This is intuitive, as the density ratio $g^*(x, z)$ between the joint distribution and product of marginals scales with the variance of H . We will revisit this example when we consider direct embeddings in Section 4.

4. Direct embeddings under hidden variable structure

The results of Section 3 demonstrate that contrastive learning, coupled with the landmark embedding method, produces representations that are useful for linear prediction. However, in practice, contrastive learning is not coupled with some landmark embedding. Rather, as an alternative to learning the bivariate function $f^*(\cdot, \cdot)$, practitioners directly optimize \mathbb{R}^m -valued embedding functions and use these for downstream learning tasks. A typical approach is to solve the following optimization problem:

$$(\eta^*, \psi^*) \in \underset{\eta, \psi}{\operatorname{argmin}} L_{\text{direct}}((x, z) \mapsto \eta(x)^\top \psi(z)), \quad L_{\text{direct}}(f) := \mathbb{E} \left[\log \left(1 + \frac{1}{f(X_c, Z_c)^{Y_c}} \right) \right]. \quad (2)$$

Here, the minimization is over functions $\eta: \mathcal{X} \rightarrow \mathbb{R}^m$ and $\psi: \mathcal{Z} \rightarrow \mathbb{R}^m$ for some embedding dimension m . Note that this is the same as minimizing the cross-entropy loss when our prediction on $(x, z) \in \mathcal{X} \times \mathcal{Z}$ is $\log(\eta(x)^\top \psi(z))$. The loss in (2) bears a resemblance to the contrastive losses proposed by [Hadsell et al. \(2006\)](#), but differing in the fact that (a) the embedding functions are allowed to differ in the two views and (b) inner products are used as opposed to distances.

Of crucial importance here is the size of the embedding dimension m needed to guarantee good performance on downstream linear predictions. In this section, we tackle this problem in the setting where there is some hidden variable H that renders the two views X and Z conditionally independent. Note that, in general, there is always such a random variable (by taking $H = (X, Z)$); however there may be a much more succinct hidden variable structure, and when this is the case, we show that relatively low-dimensional embeddings can achieve good predictive performance.

4.1. Discrete hidden variables

When H is a discrete random variable, taking values in some finite set S , [Tosh et al. \(2020\)](#) observed that we may write

$$g^*(x, z) = \eta^*(x)^\top \psi^*(z)$$

where $\eta^*: \mathcal{X} \rightarrow \mathbb{R}^{|S|}$ and $\psi^*: \mathcal{Z} \rightarrow \mathbb{R}^{|S|}$ satisfy

$$\eta^*(x) = (\Pr(H = h \mid X = x))_{h \in S} \quad \text{and} \quad \psi^*(z) = \frac{1}{p_Z(z)} (p_{Z|H}(z \mid h))_{h \in S}.$$

It is not too hard to show that there is a linear function of η^* that reproduces μ . Namely, we may take $w \in \mathbb{R}^{|S|}$ to be

$$w = \mathbb{E}[\psi^*(Z)\mathbb{E}[Y | Z]] = \int \psi^*(z)\mathbb{E}[Y | Z = z]p_Z(z) dz.$$

For this choice of w , we have

$$w^\top \eta^*(x) = \int \eta^*(x)^\top \psi^*(z)\mathbb{E}[Y | Z = z]p_Z(z) dz = \int \mathbb{E}[Y | Z = z]g^*(x, z)p_Z(z) dz = \mu(x).$$

Models with discrete hidden variables include multi-view mixture models (Chaudhuri et al., 2009; Anandkumar et al., 2012), as well as models with richer hidden variable structure, such as hidden Markov models and phylogenetic trees (Mossel and Roch, 2005; Allman et al., 2009).

4.2. General hidden variables

In general, there may not be a discrete random variable that makes X and Z conditionally independent. However, there is always some random variable that does satisfy this. How much such a random variable buys us will naturally depend on its structure and relationship with the views X and Z . We will establish the existence of a low-dimensional embedding via a probabilistic construction. However, before we do this, we first verify that approximating g^* in a certain sense is sufficient to get good predictions.

Lemma 6 *For every $\eta: \mathcal{X} \rightarrow \mathbb{R}^m$ and $\psi: \mathcal{Z} \rightarrow \mathbb{R}^m$, there exists a $w \in \mathbb{R}^m$ such that*

$$\mathbb{E}[(w^\top \eta(X) - \mu(X))^2] \leq \mathbb{E}[Y^2] \cdot \varepsilon_{\text{direct}}(\eta, \psi)$$

where

$$\varepsilon_{\text{direct}}(\eta, \psi) := \mathbb{E} \left[\left(\eta(X)^\top \psi(\tilde{Z}) - g^*(X, \tilde{Z}) \right)^2 \right]$$

and $(X, \tilde{Z}) \sim p_X \otimes p_Z$.

Thus, it suffices to find embeddings whose inner product only approximates g^* under $p_X \otimes p_Z$.

4.2.1. A PROBABILISTIC CONSTRUCTION

We next show the existence of embeddings η, ψ that are low-dimensional and that approximate g^* in the sense of Lemma 6.

Lemma 7 *Let H denote a random variable that renders X and Z conditionally independent. For any $m > 0$, there exist $\eta: \mathcal{X} \rightarrow \mathbb{R}^m$ and $\psi: \mathcal{Z} \rightarrow \mathbb{R}^m$ such that*

$$\mathbb{E} \left[\left(\eta(X)^\top \psi(\tilde{Z}) - g^*(X, \tilde{Z}) \right)^2 \right] \leq \frac{1}{m} \text{var} \left(\frac{p_{X|H}(X | \bar{H})p_{Z|H}(\tilde{Z} | \bar{H})}{p_X(X)p_Z(\tilde{Z})} \right)$$

where $(X, \tilde{Z}, \bar{H}) \sim p_X \otimes p_Z \otimes p_H$.

Note that while Lemmas 6 and 7 guarantee the existence of low-dimensional embeddings that are useful for downstream linear prediction, they do not guarantee that there exist useful minimizers of (2) for any given dimension. In Section 5 we discuss this issue further.

4.2.2. EXAMPLES

We revisit the topic model and Gaussian model examples from Section 3.2 and Section 3.3.

Proposition 8 *Assume that Y takes values in $[-1, 1]$. In the topic model setting, there exists $\eta: \mathcal{X} \rightarrow \mathbb{R}^m$ and $\psi: \mathcal{Z} \rightarrow \mathbb{R}^m$ such that*

$$\varepsilon_{\text{direct}}(\eta, \psi) \leq \begin{cases} O(1/m) & \text{if } \alpha = \Theta(1) \text{ as } K \rightarrow \infty; \\ O(K^2/m) & \text{if } \alpha \leq 1/K. \end{cases}$$

This is essentially the same bound as what was obtained in Proposition 4 for the landmark embedding. We see that even though the hidden variable structure is not discrete, we still obtain bounds that are polynomial in the dimension of the hidden variable.

Proposition 9 *Assume that Y takes values in $[-1, 1]$. In the Gaussian model setting, if $\sigma^2 < 1/2$, then there exist $\eta: \mathcal{X} \rightarrow \mathbb{R}^m$ and $\psi: \mathcal{Z} \rightarrow \mathbb{R}^m$ such that*

$$\varepsilon_{\text{direct}}(\eta, \psi) \leq \frac{1}{m} \cdot \frac{(1 + \sigma^2)^2}{\sqrt{1 - 4\sigma^4}}.$$

Here, we note that the existence argument for η and ψ requires a stronger condition than what was required for the landmark embedding to work. This is reflected in the condition $\sigma^2 < 1/2$.

5. Error analysis

We now turn to the problem of bounding the error in the representation incurred by lack of data, imprecise optimization, or restricted function classes. Specifically, we will be interested in the following measure of risk for an embedding function $\phi: \mathcal{X} \rightarrow \mathbb{R}^m$:

$$R(\phi) := \inf_{w \in \mathbb{R}^m} \mathbb{E} \left[(w^\top \phi(X) - \mu(X))^2 \right]. \quad (3)$$

To obtain a guarantee on the mean squared error in approximating $\mathbb{E}[Y \mid X, Z]$, we can simply appeal to Lemma 1.

Our goal is to relate the risk of an embedding to the excess loss in terms of either L_{lm} from (1) or L_{direct} from (2). Note that these two loss functions operate on different scales: the minimizer of L_{lm} is the log-odds ratio function $\log \circ g^*$, and the minimizer of L_{direct} is the odds ratio function g^* .

We first state the error guarantee for the landmark embedding method. The analysis is a modification of the argument from Tosh et al. (2020).

Theorem 10 *Assume Y takes values in $[-1, 1]$, and let $g_{\max} := \sup_{(x,z) \in \text{supp } p_X \otimes p_Z} g^*(x, z)$. Pick any $f: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that $g := \exp \circ f$ satisfies $\sup_{(x,z) \in \text{supp } p_X \otimes p_Z} g(x, z) \leq g_{\max}$. Let $\varphi: \mathcal{X} \rightarrow \mathbb{R}^m$ be the embedding function given by*

$$\varphi(x) = (g(x, Z_1), \dots, g(x, Z_m))$$

where Z_1, \dots, Z_m are i.i.d. copies of Z . With probability at least $1 - \delta$ (over the realization of Z_1, \dots, Z_m),

$$R(\varphi) \leq 2\varepsilon_{\text{lm}} + 4(1 + g_{\max})^2 \sqrt{2\varepsilon_{\text{opt,lm}}\varepsilon_{\text{lm}}} + 16(1 + g_{\max})^4 \varepsilon_{\text{opt,lm}}$$

where ε_{lm} is defined in Lemma 2 and $\varepsilon_{\text{opt,lm}}$ is the excess L_{lm} -loss:

$$\varepsilon_{\text{opt,lm}} := L_{\text{lm}}(f) - L_{\text{lm}}(\log \circ g^*).$$

We now turn to bounding the risk associated with direct embeddings.

Theorem 11 *Let $g_{\text{max}} := \sup_{(x,z) \in \text{supp } p_X \otimes p_Z} g^*(x, z)$. Pick any embedding functions $\eta: \mathcal{X} \rightarrow \mathbb{R}^m$ and $\psi: \mathcal{Z} \rightarrow \mathbb{R}^m$ such that $\sup_{(x,z) \in \text{supp } p_X \otimes p_Z} \eta(x)^\top \psi(z) \leq g_{\text{max}}$. We have*

$$R(\eta) \leq \mathbb{E}[Y^2] (1 + g_{\text{max}})^4 \varepsilon_{\text{opt,direct}}$$

where $\varepsilon_{\text{opt,direct}}$ is the excess L_{direct} -loss:

$$\varepsilon_{\text{opt,direct}} := L_{\text{direct}}((x, z) \mapsto \eta(x)^\top \psi(z)) - L_{\text{direct}}(g^*).$$

We point out that $\varepsilon_{\text{opt,direct}}$ is the excess loss relative to the odds ratio g^* , and not necessarily the best m -dimensional representation. Thus, when there are no perfect m -dimensional embedding functions η^*, ψ^* satisfying $g^*(x, z) = \eta^*(x)^\top \psi^*(z)$, the quantity $\varepsilon_{\text{opt,direct}}$ accounts for both the error due to optimization and the error due to representational non-realizability. Moreover, as discussed after the statement of Lemma 7, we do not actually have a handle on the relationship between embedding dimension and the extent to which bivariate architectures can provide approximations of (2). It therefore remains an interesting future direction to provide a more complete picture of the general setting in which the error due to optimization is teased apart from the representational issues arising from the use of finite dimensional embedding functions.

Acknowledgments

We acknowledge support from NSF grants CCF-1740833, DMREF-1534910, IIS-1563785; a Bloomberg Data Science Research Grant; a JP Morgan Faculty Award; and a Sloan Research Fellowship. We also thank the reviewers whose helpful comments led to improvements in this paper.

References

- N. Abe, B. Zadrozny, and J. Langford. Outlier detection by active learning. In *International Conference on Knowledge Discovery and Data Mining*, 2006.
- R. Ahlswede. The final form of Tao’s inequality relating conditional expectation and conditional mutual information. *Advances in Mathematics of Communications*, 2007.
- E. S. Allman, C. Matias, J. A. Rhodes, et al. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 2009.
- A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory*, 2012.
- R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 2005.

- R. K. Ando and T. Zhang. Two-view feature generation model for semi-supervised learning. In *International Conference on Machine Learning*, 2007.
- S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, 2019.
- P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, 2019.
- D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003.
- A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Conference on Computational Learning Theory*, 1998.
- K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *International Conference on Machine Learning*, 2009.
- T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.
- S. Dasgupta, M. L. Littman, and D. A. McAllester. PAC generalization bounds for co-training. In *Advances in Neural Information Processing Systems*, 2002.
- S. Du, A. Krishnamurthy, N. Jiang, A. Agarwal, M. Dudik, and J. Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, 2019.
- D. P. Foster, R. Johnson, S. M. Kakade, and T. Zhang. Multi-view dimensionality reduction via canonical correlation analysis. Technical Report TR-2009-5, TTI-Chicago, 2009.
- M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*, 2010.
- R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Conference on Computer Vision and Pattern Recognition*, 2006.
- R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.
- H. Hotelling. Relations between two sets of variates. *Biometrika*, 1936.
- D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden Markov models. In *Conference on Learning Theory*, 2009.
- A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems*, 2016.

- A. Hyvarinen and H. Morioka. Nonlinear ica of temporally dependent stationary sources. In *International Conference on Artificial Intelligence and Statistics*, 2017.
- A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.
- S. M. Kakade and D. P. Foster. Multi-view regression via canonical correlation analysis. In *International Conference on Computational Learning Theory*, 2007.
- J. Langford, R. Salakhutdinov, and T. Zhang. Learning nonlinear dynamic models. In *International Conference on Machine Learning*, 2009.
- J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting what you already know helps: provable self-supervised learning. *arXiv:2008.01064*, 2020.
- M. L. Littman and R. S. Sutton. Predictive representations of state. In *Advances in Neural Information Processing Systems*, 2002.
- L. Logeswaran and H. Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018.
- Z. Ma and M. Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *Empirical Methods in Natural Language Processing*, 2018.
- D. McAllester and K. Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, 2020.
- D. Misra, M. Henaff, A. Krishnamurthy, and J. Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International Conference on Machine Learning*, 2020.
- E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden Markov models. In *Symposium on Theory of Computing*, 2005.
- A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.
- C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 2000.
- D. Pathak, P. Krahenbuhl, J. Donahue, D. T. and A. A. Efros. Context encoders: Feature learning by inpainting. In *Conference on Computer Vision and Pattern Recognition*, 2016.
- L. Song, B. Boots, S. M. Siddiqi, G. Gordon, and A. Smola. Hilbert space embeddings of hidden Markov models. In *International Conference on Machine Learning*, 2010.
- I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 2005.
- T. Tao. Szemerédi’s regularity lemma revisited. *Contributions to Discrete Mathematics*, 2006.

- M. Telgarsky. Dirichlet draws are sparse with high probability. *arXiv:1301.4917*, 2013.
- Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. *arXiv:1906.05849*, 2019.
- N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Allerton Conference on Communication, Control and Computing*, 1999.
- C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive estimation reveals topic posterior information to linear models. *arXiv:2003.02234*, 2020.
- M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2019.
- P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine learning*, 2008.
- Y. Wu and S. Verdú. Functional properties of minimum mean-square error and mutual information. *IEEE Transactions on Information Theory*, 2011.

Appendix A. Omitted proofs

A.1. Proof of Lemma 1

By the law of total expectation and Jensen's inequality, we have

$$\begin{aligned}
 \mathbb{E}[(\mu(X) - \mathbb{E}[Y | X])^2] &= \mathbb{E}[(\mathbb{E}[\mathbb{E}[Y | Z] | X] - \mathbb{E}[Y | X])^2] \\
 &= \mathbb{E}[(\mathbb{E}[\mathbb{E}[Y | Z] - \mathbb{E}[Y | X, Z] | X])^2] \\
 &\leq \mathbb{E}[\mathbb{E}[(\mathbb{E}[Y | Z] - \mathbb{E}[Y | X, Z])^2 | X]] \quad (\text{Jensen's inequality}) \\
 &= \mathbb{E}[(\mathbb{E}[Y | Z] - \mathbb{E}[Y | X, Z])^2] \\
 &= \varepsilon_Z.
 \end{aligned}$$

Using the AM/GM inequality, for any $\lambda > 0$,

$$\begin{aligned}
 &\mathbb{E}[(\mu(X) - \mathbb{E}[Y | X, Z])^2] \\
 &= \mathbb{E}[(\mathbb{E}[Y | X] - \mathbb{E}[Y | X, Z] + \mu(X) - \mathbb{E}[Y | X])^2] \\
 &\leq (1 + 1/\lambda)\mathbb{E}[(\mathbb{E}[Y | X] - \mathbb{E}[Y | X, Z])^2] + (1 + \lambda)\mathbb{E}[(\mu(X) - \mathbb{E}[Y | X])^2] \\
 &\leq (1 + 1/\lambda)\varepsilon_X + (1 + \lambda)\varepsilon_Z.
 \end{aligned}$$

Optimizing the bound with respect to λ gives

$$\mathbb{E}[(\mu(X) - \mathbb{E}[Y | X, Z])^2] \leq \varepsilon_X + 2\sqrt{\varepsilon_X\varepsilon_Z} + \varepsilon_Z. \quad \blacksquare$$

A.2. Proof of Lemma 2

We partition the m coordinates of the embedding into blocks of $n := \lfloor m/\log_2(1/\delta) \rfloor$ coordinates per block. We first consider the part of the embedding corresponding to the first block, say, $\varphi_{1:n}^*: \mathcal{X} \rightarrow \mathbb{R}^n$. Define the weight vector $v \in \mathbb{R}^n$ by

$$v = v(Z_1, \dots, Z_n) := \frac{1}{n}(\mathbb{E}[Y_1 | Z_1], \dots, \mathbb{E}[Y_n | Z_n]).$$

Define $A_i(x) = \mathbb{E}[Y_i | Z_i]g^*(x, Z_i) - \mu(x)$ for all $x \in \mathcal{X}$, so $A_1(x), \dots, A_n(x)$ are i.i.d. mean-zero random variables. This implies

$$\mathbb{E}[(v^\top \varphi_{1:n}^*(x) - \mu(x))^2] = \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^n A_i(x)\right)^2\right] = \frac{\mathbb{E}[A_1(x)^2]}{n}.$$

Now replacing x with X and taking expectations gives

$$\mathbb{E}[(v^\top \varphi_{1:n}^*(X) - \mu(X))^2] = \frac{\mathbb{E}[A_1(X)^2]}{n} = \frac{\text{var}(\mathbb{E}[Y_1 | Z_1]g^*(X, Z_1))}{n}.$$

By Markov's inequality, the event

$$\mathbb{E}[(v^\top \varphi_{1:n}^*(X) - \mu(X))^2 | Z_1, \dots, Z_m] \leq \frac{2}{n} \text{var}(\mathbb{E}[Y_1 | Z_1]g^*(X, Z_1)).$$

has probability at least $1/2$. We can analogously define such a ‘‘good’’ event for each block of coordinates. With probability at least $1 - \delta$, at least one of these good events occurs; in this event, we can pick any such ‘‘good’’ block, set the corresponding weights in w according to the construction above, and set the remaining weights in w to zero. This produces the desired guarantee. \blacksquare

A.3. Proof of Proposition 4

We first make a simple observation about the relationship between terms in the definitions of ε_{lm} and $\varepsilon_{\text{direct}}$.

Proposition 12 *If $X \perp\!\!\!\perp Z \mid H$, then for any $(x, z) \in \mathcal{X} \times \mathcal{Z}$,*

$$g^*(x, z)^2 \leq \mathbb{E} \left[\left(\frac{p_{X|H}(x | H)p_{Z|H}(z | H)}{p_X(x)p_Z(z)} \right)^2 \right]$$

where $H \sim p_H$.

Proof Fix any $(x, z) \in \mathcal{X} \times \mathcal{Z}$. Then

$$\begin{aligned} g^*(x, z)^2 &= \left(\frac{p_{X,Z}(x, z)}{p_X(x)p_Z(z)} \right)^2 \\ &= \frac{1}{(p_X(x)p_Z(z))^2} \left(\int p_{X,Z,H}(x, z, h) \, dh \right)^2 \\ &= \frac{1}{(p_X(x)p_Z(z))^2} \left(\int p_{X|H}(x | h)p_{Z|H}(z | h)p_H(h) \, dh \right)^2 \\ &\leq \frac{1}{(p_X(x)p_Z(z))^2} \int \left(p_{X|H}(x | h)p_{Z|H}(z | h) \right)^2 p_H(h) \, dh \\ &= \mathbb{E} \left[\left(\frac{p_{X|H}(x | H)p_{Z|H}(z | H)}{p_X(x)p_Z(z)} \right)^2 \right], \end{aligned}$$

where the inequality follows from Jensen's inequality. \blacksquare

Now we return to the proof of Proposition 4. Since Y takes values in $[-1, 1]$, it suffices to bound $\mathbb{E}[g^*(X, Z_1)^2]$. By Proposition 12, for any (x, z_1) we have

$$g^*(x, z_1) \leq \mathbb{E} \left[\left(\frac{p_{X|\Theta}(x | \Theta)p_{Z|\Theta}(z_1 | \Theta)}{p_X(x)p_Z(z_1)} \right)^2 \right].$$

Now replacing (x, z_1) with (X, Z_1) and taking expectations on both sides, the result follows by Proposition 13. \blacksquare

We note that a direct analysis of $\mathbb{E}[g^*(X, Z_1)^2]$ is also straightforward, and would ultimately involve only second-moments of $\bar{\Theta}$ (as opposed to fourth-moments, considered in Proposition 13). However, the final bound is the same.

A.4. Proof of Proposition 5

The marginal distribution of (X, Z) is $\mathcal{N}(0, \Sigma)$, where

$$\Sigma := \begin{pmatrix} 1 + \sigma^2 & \sigma^2 \\ \sigma^2 & 1 + \sigma^2 \end{pmatrix}.$$

Thus, letting $v := (x, z)$, we have

$$p_{X,Z}(v) = \frac{1}{2\pi\sqrt{1+2\sigma^2}} \exp\left(-\frac{1}{2}v^\top \Sigma^{-1}v\right),$$

$$p_X(\cdot) = p_Z(\cdot) = \frac{1}{\sqrt{2\pi(1+\sigma^2)}} \exp\left(-\frac{(\cdot)^2}{2(1+\sigma^2)}\right)$$

Therefore, for any fixed $(x, z) \in \mathbb{R}^2$,

$$g^*(x, z)^2 = \frac{p_{X,Z}(x, z)^2}{p_X(x)^2 p_Z(z)^2} = \frac{(1+\sigma^2)^2}{1+2\sigma^2} \exp\left(-v^\top \Sigma^{-1}v + \frac{x^2+z^2}{1+\sigma^2}\right) = \frac{(1+\sigma^2)^2}{1+2\sigma^2} \exp(v^\top A v)$$

where

$$A := -\frac{\sigma^2}{1+2\sigma^2} \begin{pmatrix} \frac{\sigma^2}{1+\sigma^2} & -1 \\ -1 & \frac{\sigma^2}{1+\sigma^2} \end{pmatrix} = \lambda_1 u_1 u_1^\top + \lambda_2 u_2 u_2^\top$$

has eigenvalues $\lambda_1 = \frac{\sigma^2}{(1+\sigma^2)(1+2\sigma^2)}$ and $\lambda_2 = -\frac{\sigma^2}{1+\sigma^2}$ corresponding to some orthonormal eigenvectors u_1 and u_2 . Replacing (x, z) with $V := (X, \tilde{Z}) \sim p_X \otimes p_Z$ and taking expectation gives

$$\begin{aligned} \mathbb{E}[g^*(X, \tilde{Z})^2] &= \frac{(1+\sigma^2)^2}{1+2\sigma^2} \mathbb{E}\left[\exp\left(\lambda_1(u_1^\top V)^2 + \lambda_2(u_2^\top V)^2\right)\right] \\ &= \frac{(1+\sigma^2)^2}{1+2\sigma^2} \mathbb{E}\left[\exp\left(\lambda_1(u_1^\top V)^2\right)\right] \mathbb{E}\left[\exp\left(\lambda_2(u_2^\top V)^2\right)\right] \\ &= \frac{(1+\sigma^2)^2}{1+2\sigma^2} \cdot \frac{1}{\sqrt{1-2\lambda_1(1+\sigma^2)}} \cdot \frac{1}{\sqrt{1-2\lambda_2(1+\sigma^2)}} \\ &= \frac{(1+\sigma^2)^2}{1+2\sigma^2} \cdot \frac{1}{\sqrt{1-\frac{2\sigma^2}{1+2\sigma^2}}} \cdot \frac{1}{\sqrt{1+2\sigma^2}} \\ &= \frac{(1+\sigma^2)^2}{1+2\sigma^2}. \end{aligned}$$

Above, we use the fact that $u_1^\top V$ and $u_2^\top V$ are independent $\mathcal{N}(0, 1+\sigma^2)$ random variables so that the expectations in the second line are the chi-squared moment generating function. \blacksquare

A.5. Proof of Lemma 6

We take w to be

$$w := \mathbb{E}\left[\mathbb{E}[Y | Z]\psi(Z)\right].$$

Let $(\tilde{X}, \tilde{Z}, \tilde{Y})$ be an independent copy of (X, Z, Y) , and observe that

$$\mu(X) = \mathbb{E}\left[\mathbb{E}[Y | Z] | X\right] = \mathbb{E}\left[\mathbb{E}[\tilde{Y} | \tilde{Z}]g^*(X, \tilde{Z}) | X\right].$$

Therefore

$$\begin{aligned}
 \mathbb{E}[(w^\top \eta(X) - \mu(X))^2] &= \mathbb{E} \left[\left(\mathbb{E} \left[\mathbb{E}[\tilde{Y} \mid \tilde{Z}] \psi(\tilde{Z}) \right]^\top \eta(X) - \mu(X) \right)^2 \right] \\
 &= \mathbb{E} \left[\left(\mathbb{E} \left[\mathbb{E}[\tilde{Y} \mid \tilde{Z}] \psi(\tilde{Z}) \right]^\top \eta(X) - \mathbb{E}[\tilde{Y} \mid \tilde{Z}] g^*(X, \tilde{Z}) \mid X \right)^2 \right] \\
 &\leq \mathbb{E} \left[\mathbb{E} \left[\mathbb{E}[\tilde{Y} \mid \tilde{Z}]^2 \mid X \right] \cdot \mathbb{E} \left[\left(\psi(\tilde{Z})^\top \eta(X) - g^*(X, \tilde{Z}) \right)^2 \mid X \right] \right] \\
 &= \mathbb{E} \left[\mathbb{E} \left[\mathbb{E}[\tilde{Y} \mid \tilde{Z}]^2 \right] \cdot \mathbb{E} \left[\left(\psi(\tilde{Z})^\top \eta(X) - g^*(X, \tilde{Z}) \right)^2 \mid X \right] \right] \\
 &\leq \mathbb{E} \left[\mathbb{E}[\tilde{Y}^2] \cdot \mathbb{E} \left[\left(\psi(\tilde{Z})^\top \eta(X) - g^*(X, \tilde{Z}) \right)^2 \mid X \right] \right] \\
 &= \mathbb{E}[Y^2] \cdot \mathbb{E} \left[\left(\psi(\tilde{Z})^\top \eta(X) - g^*(X, \tilde{Z}) \right)^2 \right].
 \end{aligned}$$

Above, the first inequality follows from Cauchy-Schwarz, the subsequent equality uses the independence of (\tilde{Z}, \tilde{Y}) and X , and the second the inequality follows from Jensen's inequality and the law of total expectation. \blacksquare

A.6. Proof of Lemma 7

We will prove this using the probabilistic method, constructing a random embedding of dimension m that satisfies the lemma in expectation. This will suffice to show that there exists such an embedding.

Let H_1, \dots, H_m be i.i.d. copies of H . Define, for each $(x, z) \in \mathcal{X} \times \mathcal{Z}$,

$$\begin{aligned}
 \eta(x) &= \frac{1}{p_X(x) \sqrt{m}} (p_{X|H}(x \mid H_i))_{i=1}^m \\
 \psi(z) &= \frac{1}{p_Z(z) \sqrt{m}} (p_{Z|H}(z \mid H_i))_{i=1}^m \\
 B_i(x, z) &= \frac{p_{X|H}(x \mid H_i) p_{Z|H}(z \mid H_i)}{p_X(x) p_Z(z)} - g^*(x, z).
 \end{aligned}$$

Observe that $B_1(x, z), \dots, B_m(x, z)$ are i.i.d. mean-zero random variables, and

$$\eta(x)^\top \psi(z) - g^*(x, z) = \frac{1}{m} \sum_{i=1}^m B_i(x, z),$$

and

$$\mathbb{E}[(\eta(x)^\top \psi(z) - g^*(x, z))^2] = \frac{\mathbb{E}[B_1(x, z)^2]}{m}.$$

Now replacing (x, z) with $(X, \tilde{Z}) \sim p_X \otimes p_Z$ and taking expectations gives

$$\mathbb{E} \left[(\eta(X)^\top \psi(\tilde{Z}) - g^*(X, \tilde{Z}))^2 \right] = \frac{\mathbb{E}[B_1(X, \tilde{Z})^2]}{m} = \frac{1}{m} \text{var} \left(\frac{p_{X|H}(X \mid \bar{H}) p_{Z|H}(\tilde{Z} \mid \bar{H})}{p_X(X) p_Z(\tilde{Z})} \right).$$

\blacksquare

A.7. Proof of Proposition 8

By Lemma 7, it suffices to bound

$$\text{var} \left(\frac{p_{X|\Theta}(X | \bar{\Theta})p_{Z|\Theta}(\tilde{Z} | \bar{\Theta})}{p_X(X)p_Z(\tilde{Z})} \right),$$

for $(X, \tilde{Z}, \bar{\Theta}) \sim p_X \otimes p_Z \otimes p_\Theta$. This, in turn, is bounded above by

$$\mathbb{E} \left[\left(\frac{p_{X|\Theta}(X | \bar{\Theta})p_{Z|\Theta}(\tilde{Z} | \bar{\Theta})}{p_X(X)p_Z(\tilde{Z})} \right)^2 \right].$$

So Proposition 8 follows immediately from the following result.

Proposition 13 *In the topic modeling setting,*

$$\mathbb{E} \left[\left(\frac{p_{X|\Theta}(X | \bar{\Theta})p_{Z|\Theta}(\tilde{Z} | \bar{\Theta})}{p_X(X)p_Z(\tilde{Z})} \right)^2 \right] = \begin{cases} \Theta(1) & \text{if } \alpha = \Theta(1) \text{ as } K \rightarrow \infty; \\ \Theta(K^2) & \text{if } \alpha \leq 1/K. \end{cases}$$

Proof For any word $v \in \mathcal{V}$, let $k(v) \in [K]$ denote the unique topic for which $P_k(v) > 0$. For any $v \in \mathcal{V}$ and $\theta \in \Delta^{K-1}$, we have

$$\begin{aligned} p_X(v) &= \frac{1}{K} P_{k(v)}(v), \\ p_{X|\Theta}(v | \theta) &= \theta_{k(v)} P_{k(v)}(v). \end{aligned}$$

Therefore, we have for any $\theta \in \Delta^{K-1}$, $x \in \mathcal{V}$, and $z \in \mathcal{V}$,

$$\left(\frac{p_{X|\Theta}(x | \theta)p_{Z|\Theta}(z | \theta)}{p_X(x)p_Z(z)} \right)^2 = K^4 \cdot \theta_{k(x)}^2 \theta_{k(z)}^2.$$

Replacing (x, z, θ) with $(X, \tilde{Z}, \bar{\Theta})$ and taking expectations gives

$$\begin{aligned} \mathbb{E} \left[\left(\frac{p_{X|\Theta}(X | \bar{\Theta})p_{Z|\Theta}(\tilde{Z} | \bar{\Theta})}{p_X(X)p_Z(\tilde{Z})} \right)^2 \right] &= \mathbb{E} \left[K^4 \cdot \bar{\Theta}_{k(X)}^2 \bar{\Theta}_{k(\tilde{Z})}^2 \right] \\ &= K^4 \cdot \sum_{k=1}^K \sum_{k'=1}^K \Pr(k(X) = k) \Pr(k(\tilde{Z}) = k') \mathbb{E} \left[\bar{\Theta}_k^2 \bar{\Theta}_{k'}^2 \right] \\ &= K^2 \cdot \sum_{k=1}^K \sum_{k'=1}^K \mathbb{E} \left[\bar{\Theta}_k^2 \bar{\Theta}_{k'}^2 \right] \\ &= K^3 \cdot \left(\mathbb{E}[\bar{\Theta}_1^4] + (K-1) \mathbb{E}[\bar{\Theta}_1^2 \bar{\Theta}_2^2] \right) \end{aligned}$$

where the fourth and fifth steps follow by symmetry. The fourth-moments in the final expression are:

$$\begin{aligned}\mathbb{E}[\bar{\Theta}_1^2 \bar{\Theta}_2^2] &= \frac{\Gamma(K\alpha)}{\Gamma(K\alpha + 4)} \cdot \left(\frac{\Gamma(\alpha + 2)}{\Gamma(\alpha)} \right)^2 \\ \mathbb{E}[\bar{\Theta}_1^4] &= \frac{\Gamma(K\alpha)}{\Gamma(K\alpha + 4)} \cdot \frac{\Gamma(\alpha + 4)}{\Gamma(\alpha)}.\end{aligned}$$

Therefore, we have the following:

1. For $\alpha = \Theta(1)$ and $K \rightarrow \infty$,

$$\mathbb{E} \left[\left(\frac{p_{X|\Theta}(X | \bar{\Theta}) p_{Z|\Theta}(\tilde{Z} | \bar{\Theta})}{p_X(X) p_Z(\tilde{Z})} \right)^2 \right] = K^3 \left(\Theta \left(\frac{1}{K^4} \right) + \Theta \left(\frac{1}{K^3} \right) \right) = \Theta(1).$$

2. For $\alpha \leq 1/K$,

$$\mathbb{E} \left[\left(\frac{p_{X|\Theta}(X | \bar{\Theta}) p_{Z|\Theta}(\tilde{Z} | \bar{\Theta})}{p_X(X) p_Z(\tilde{Z})} \right)^2 \right] = K^3 \left(\Theta \left(\frac{1}{K} \right) + \Theta(\alpha) \right) = \Theta(K^2).$$

■

A.8. Proof of Proposition 9

The proof is similar to that of Proposition 5. Using similar computations, we obtain for any fixed $(x, z) \in \mathbb{R}^2$,

$$\begin{aligned}\mathbb{E} \left[\left(\frac{p_{X|H}(x | H) p_{Z|H}(z | H)}{p_X(x) p_Z(z)} \right)^2 \right] &= \frac{(1 + \sigma^2)^2}{\sqrt{1 + 4\sigma^2}} \exp \left(\frac{2\sigma^2(x + z)^2}{1 + 4\sigma^2} - \frac{\sigma^2(x^2 + z^2)}{1 + \sigma^2} \right) \\ &= \frac{(1 + \sigma^2)^2}{\sqrt{4\sigma^2 + 1}} \exp(v^\top A v)\end{aligned}$$

where

$$A := \frac{\sigma^2}{1 + 4\sigma^2} \begin{pmatrix} \frac{1-2\sigma^2}{1+\sigma^2} & 2 \\ 2 & \frac{1-2\sigma^2}{1+\sigma^2} \end{pmatrix} = \lambda_1 u_1 u_1^\top + \lambda_2 u_2 u_2^\top$$

has eigenvalues $\lambda_1 = \frac{3\sigma^2}{(1+\sigma^2)(1+4\sigma^2)}$ and $\lambda_2 = -\frac{\sigma^2}{1+\sigma^2}$ corresponding to some orthonormal eigenvectors u_1 and u_2 . Now replacing (x, z) with $V := (X, \tilde{Z}) \sim p_X \otimes p_Z$ and taking expectation gives, for $(X, \tilde{Z}, \bar{H}) \sim p_X \otimes p_Z \otimes p_H$,

$$\mathbb{E} \left[\left(\frac{p_{X|H}(X | \bar{H}) p_{Z|H}(\tilde{Z} | \bar{H})}{p_X(X) p_Z(\tilde{Z})} \right)^2 \right] = \frac{(1 + \sigma^2)^2}{\sqrt{4\sigma^2 + 1}} \mathbb{E} \left[\exp \left(\frac{3\sigma^2(u_1^\top V)^2}{(1 + \sigma^2)(1 + 4\sigma^2)} - \frac{\sigma^2(u_2^\top V)^2}{1 + \sigma^2} \right) \right].$$

Since $u_1^\top V$ and $u_2^\top V$ are independent $\mathcal{N}(0, 1 + \sigma^2)$ random variables, this expression simplifies to

$$\mathbb{E} \left[\left(\frac{p_{X|H}(X | \bar{H}) p_{Z|H}(\tilde{Z} | \bar{H})}{p_X(X) p_Z(\tilde{Z})} \right)^2 \right] = \frac{(1 + \sigma^2)^2}{\sqrt{4\sigma^2 + 1}} \cdot \frac{1}{\sqrt{1 - \frac{6\sigma^2}{1+4\sigma^2}}} \cdot \frac{1}{\sqrt{1 + 2\sigma^2}} = \frac{(1 + \sigma^2)^2}{\sqrt{1 - 4\sigma^4}}.$$

The condition $\sigma^2 < 1/2$ is used to ensure that the expectation in the last equation display is finite. \blacksquare

A.9. Proof of Theorem 10

Let $(X_1, Z_1, Y_1), \dots, (X_m, Z_m, Y_m), (X, Z, Y)$ be i.i.d., and let $\varphi^*(x)_i := g^*(x, Z_i)$ for all $i = 1, \dots, m$. Let $n := \lfloor m / \log_2(1/\delta) \rfloor$. We shall adopt the same block repetition argument as in the proof of Lemma 2, where the m coordinates are partitioned into groups of n coordinates each.

We first analyze what happens in the first block of coordinates. From the arguments in Lemma 2, we know with probability $\geq 3/4$,

$$\mathbb{E} \left[\left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i | Z_i] g^*(X, Z_i) - \mu(X) \right)^2 \mid Z_1, \dots, Z_m \right] \leq \frac{4 \text{var}(\mathbb{E}[Y_1 | Z_1] g^*(X, Z_1))}{n} = 2\varepsilon_{\text{lm}}. \quad (4)$$

We also claim that, with probability $\geq 3/4$,

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (g(X, Z_i) - g^*(X, Z_i))^2 \mid Z_1, \dots, Z_m \right] \leq 4(1 + g_{\max})^4 \varepsilon_{\text{opt,lm}}. \quad (5)$$

To see this, we make the following definitions

$$p^*(x, z) := \frac{g^*(x, z)}{1 + g^*(x, z)}, \quad \text{and} \quad p(x, z) := \frac{g(x, z)}{1 + g(x, z)}.$$

Recall that $(X_c, Z_c, Y_c) \sim \mathcal{D}_{\text{contrast}}$. Now we have

$$\begin{aligned} \varepsilon_{\text{opt,lm}} &= L_{\text{lm}}(f) - \inf_{f^*: \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}} L_{\text{lm}}(\log \circ g^*) \\ &= \mathbb{E} \left[Y_c \log \left(\frac{p^*(X_c, Z_c)}{p(X_c, Z_c)} \right) + (1 - Y_c) \log \left(\frac{1 - p^*(X_c, Z_c)}{1 - p(X_c, Z_c)} \right) \right] \\ &= \mathbb{E} \left[p^*(X_c, Z_c) \log \left(\frac{p^*(X_c, Z_c)}{p(X_c, Z_c)} \right) + (1 - p^*(X_c, Z_c)) \log \left(\frac{1 - p^*(X_c, Z_c)}{1 - p(X_c, Z_c)} \right) \right] \\ &= \mathbb{E} [\text{KL}(p^*(X_c, Z_c), p(X_c, Z_c))] \end{aligned}$$

where the second-to-last line follows from the fact that g^* is the odds ratio for the contrastive learning problem and $\text{KL}(p, q)$ denotes the KL divergence between two Bernoulli random variables. Pinsker's inequality tells us that, for any $(x, z) \in \mathcal{X} \times \mathcal{Z}$,

$$\text{KL}(p^*(x, z), p(x, z)) \geq 2(p^*(x, z) - p(x, z))^2 \geq \frac{2}{(1 + g_{\max})^4} (g^*(x, z) - g(x, z))^2$$

Since $(X_c, Z_c) \sim \frac{1}{2}p_{X,Z} + \frac{1}{2}p_X \otimes p_Z$,

$$\begin{aligned} \mathbb{E} \left[(g^*(X_c, Z_c) - g(X_c, Z_c))^2 \right] &= \frac{1}{2} \mathbb{E} \left[(g^*(X, Z) - g(X, Z))^2 \right] + \frac{1}{2} \mathbb{E} \left[(g^*(X, Z_1) - g(X, Z_1))^2 \right] \\ &\geq \frac{1}{2} \mathbb{E} \left[(g^*(X, Z_1) - g(X, Z_1))^2 \right]. \end{aligned}$$

Therefore, we conclude that

$$\mathbb{E} \left[(g^*(X, Z_1) - g(X, Z_1))^2 \right] \leq (1 + g_{\max})^4 \varepsilon_{\text{opt,lm}},$$

and hence also

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (g^*(X, Z_i) - g(X, Z_i))^2 \right] \leq (1 + g_{\max})^4 \varepsilon_{\text{opt,lm}}.$$

By Markov's inequality, (5) holds with probability $3/4$. A union bound grants that (4) and (5) hold simultaneously with probability $\geq 1/2$. Call this the ‘‘good’’ event for this first block of landmarks.

Now considering all blocks, with probability $1 - \delta$, the good event holds for at least one group of coordinates. As in the proof of Lemma 2, we will set $w_i = \frac{1}{n} \mathbb{E}[Y_i | Z_i]$ for the coordinates in the good group and we set $w_i = 0$ for all other coordinates. Thus, with probability $1 - \delta$ we can conclude two facts. First, that φ^* satisfies

$$\mathbb{E} \left[(w^\top \varphi^*(X) - \mu(X) | Z_1, \dots, Z_m)^2 \right] \leq 2\varepsilon_{\text{lm}}.$$

Second, there is some block of n coordinates (which we take to be $\{1, \dots, n\}$ without loss of generality) such that

$$\begin{aligned} \mathbb{E} \left[(w^\top \varphi(X) - w^\top \varphi^*(X))^2 | Z_1, \dots, Z_m \right] &\leq \|w\|_2^2 \cdot \mathbb{E} \left[\|\varphi(X) - \varphi^*(X)\|_2^2 | Z_1, \dots, Z_m \right] \\ &\leq \frac{1}{n} \mathbb{E} \left[\|\varphi(X) - \varphi^*(X)\|_2^2 | Z_1, \dots, Z_m \right] \\ &= \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (g(X, Z_i) - g^*(X, Z_i))^2 | Z_1, \dots, Z_m \right] \\ &\leq 4(1 + g_{\max})^4 \varepsilon_{\text{opt,lm}} \end{aligned}$$

where the first inequality follows from Cauchy-Schwarz, the second inequality comes from the boundedness of Y , and the third inequality is (5). Putting it all together with the AM/GM inequality gives us the theorem statement. \blacksquare

A.10. Proof of Theorem 11

From Lemma 6,

$$R(\eta) = \inf_{w \in \mathbb{R}^m} \mathbb{E} \left[(w^\top \eta(X) - \mu(X))^2 \right] \leq \mathbb{E}[Y^2] \mathbb{E} \left[(\eta(X)^\top \psi(Z) - g^*(X, Z))^2 \right]$$

Therefore, we focus on bounding the second factor on the right-hand side. For the most part, the proof uses similar arguments as in that of Theorem 10.

Using the definitions

$$p^*(x, z) := \frac{g^*(x, z)}{1 + g^*(x, z)} \quad \text{and} \quad p(x, z) := \frac{\eta(x)^\top \psi(z)}{1 + \eta(x)^\top \psi(z)}$$

we have

$$\begin{aligned} \varepsilon_{\text{opt,direct}} &= L_{\text{direct}}((x, z) \mapsto \eta(x)^\top \psi(z)) - L_{\text{direct}}(g^*) \\ &= \mathbb{E} \left[Y_c \log \left(\frac{p^*(X_c, Z_c)}{p(X_c, Z_c)} \right) + (1 - Y_c) \log \left(\frac{1 - p^*(X_c, Z_c)}{1 - p(X_c, Z_c)} \right) \right] \\ &= \mathbb{E} \left[p^*(X_c, Z_c) \log \left(\frac{p^*(X_c, Z_c)}{p(X_c, Z_c)} \right) + (1 - p^*(X_c, Z_c)) \log \left(\frac{1 - p^*(X_c, Z_c)}{1 - p(X_c, Z_c)} \right) \right] \\ &= \mathbb{E} [\text{KL}(p^*(X_c, Z_c), p(X_c, Z_c))]. \end{aligned}$$

By Pinsker's inequality, for any $(x, z) \in \mathcal{X} \times \mathcal{Z}$,

$$\text{KL}(p^*(x, z), p(x, z)) \geq 2(p^*(x, z) - p(x, z))^2 \geq \frac{2}{(1 + g_{\max})^4} (g^*(x, z) - \eta(x)^\top \psi(z))^2.$$

Finally, since $(X_c, Z_c) \sim \frac{1}{2}p_{X,Z} + \frac{1}{2}p_X \otimes p_Z$,

$$\mathbb{E} \left[(g^*(X_c, Z_c) - \eta(X_c)^\top \psi(Z_c))^2 \right] \geq \frac{1}{2} \mathbb{E} \left[(g^*(X, Z) - \eta(X)^\top \psi(Z))^2 \right].$$

Putting it all together gives us the theorem statement. ■

Appendix B. Transfer learning

Finally, we consider the setting where there is a shift from the distribution $p_{X,Z,Y}$ on which we learned our representations to some new test distribution $q_{X,Z,Y}$. Under what conditions can we guarantee that our representations will transfer gracefully?

We will consider the scenario where the marginal distributions of X and Z are allowed to change, but the conditional distribution of Z given X remains the same. That is, we will impose the condition

$$p_{Z|X}(z | x) = q_{Z|X}(z | x)$$

for all x and z on p and q . To help keep our notation straight, we will use $\mathbb{E}_p[\cdot]$ to denote expectations taken with respect to $p_{X,Z,Y}$ and $\mathbb{E}_q[\cdot]$ to denote expectations taken with respect to $q_{X,Z,Y}$.

Note that in this setting, the natural analogue of μ is given by

$$\mu_q(x) = \mathbb{E}_q [\mathbb{E}_q[Y | Z] | X = x].$$

Under redundancy, μ_q enjoys a similar guarantee as μ .

Lemma 14 *Make the definitions*

$$\begin{aligned}\varepsilon_X^{(q)} &:= \mathbb{E}_q \left[\left(\mathbb{E}_q[Y | X] - \mathbb{E}_q[Y | X, Z] \right)^2 \right] \text{ and} \\ \varepsilon_Z^{(q)} &:= \mathbb{E}_q \left[\left(\mathbb{E}_q[Y | Z] - \mathbb{E}_q[Y | X, Z] \right)^2 \right].\end{aligned}$$

Then we have $\mathbb{E}_q[(\mu_q(X) - \mathbb{E}_q[Y | X, Z])^2] \leq \varepsilon_X^{(q)} + 2\sqrt{\varepsilon_X^{(q)}\varepsilon_Z^{(q)}} + \varepsilon_Z^{(q)} =: \varepsilon_\mu^{(q)}$.

As the proof of Lemma 14 is identical to that of Lemma 1, we omit it.

Lemma 14 tells us that μ_q is a natural function to approximate. However, given that we solved the contrastive optimization problem on p , it is unclear whether or not our representations will transfer gracefully over to approximating μ_q .

Our approach is to ‘fine-tune’ our landmark representation by choosing the landmarks Z_1, \dots, Z_m according to the target distribution $q_Z(\cdot)$, and embed according to

$$\varphi_q^*(x) := (g^*(x, Z_1), \dots, g^*(x, Z_m)).$$

The following lemma, analogous to Lemma 2, shows that this does indeed work.

Lemma 15 *Let $(X_1, Z_1, Y_1), \dots, (X_m, Z_m, Y_m), (X, Z, Y)$ be i.i.d. draws from $q_{X,Z,Y}$ and suppose the landmarks used to define φ_q^* are Z_1, \dots, Z_m . With probability $1 - \delta$, there exists a weight vector $w \in \mathbb{R}^m$ such that*

$$\mathbb{E}_q[(w^\top \varphi_q^*(X) - \mu_q(X))^2 | Z_1, \dots, Z_m] \leq \frac{2}{\lfloor m / \log_2(1/\delta) \rfloor} \text{var} \left(\frac{p_Z(Z_1)}{q_Z(Z_1)} \mathbb{E}_q[Y_1 | Z_1] g^*(X, Z_1) \right).$$

Proof We partition the m coordinates of the embedding into blocks of $n := \lfloor m / \log_2(1/\delta) \rfloor$ coordinates per block. We first consider the part of the embedding corresponding to the first block, say, $\varphi_{q,1:n}^* : \mathcal{X} \rightarrow \mathbb{R}^n$. Define the weight vector $v \in \mathbb{R}^n$ by

$$v = v(Z_1, \dots, Z_n) := \frac{p_Z(Z_i)}{q_Z(Z_i)n} (\mathbb{E}_q[Y_1 | Z_1], \dots, \mathbb{E}_q[Y_n | Z_n])$$

Define $A_i(x) = \frac{p_Z(Z_i)}{q_Z(Z_i)} \mathbb{E}_q[Y_i | Z_i] g^*(x, Z_i) - \mu(x)$ for all $x \in \mathcal{X}$. Note that for any x we have

$$\begin{aligned}\mathbb{E}_q \left[\frac{p_Z(Z_i)}{q_Z(Z_i)} \mathbb{E}_q[Y_i | Z_i] g^*(x, Z_i) \right] &= \mathbb{E}_p [\mathbb{E}_q[Y | Z = Z_i] g^*(x, Z_i)] \\ &= \mathbb{E}_p \left[\mathbb{E}_q[Y | Z = Z_i] \frac{p_{X,Z}(x, Z_i)}{p_X(x)p_Z(Z_i)} \right] \\ &= \mathbb{E}_p \left[\mathbb{E}_q[Y | Z = Z_i] \frac{p_{Z|X}(Z_i | X = x)}{p_Z(Z_i)} \right] \\ &= \mathbb{E}_q [\mathbb{E}_q[Y | Z = Z_i] | X = x] \\ &= \mu_q(x)\end{aligned}$$

where the second-to-last line follows from the assumption that $p_{Z|X}(z|x) = q_{Z|X}(z|x)$ for all x and z . Thus, $A_1(x), \dots, A_n(x)$ are i.i.d. mean-zero random variables. This implies

$$\mathbb{E}_q[(v^\top \varphi_{q,1:n}^*(x) - \mu(x))^2] = \mathbb{E}_q \left[\left(\frac{1}{n} \sum_{i=1}^n A_i(x) \right)^2 \right] = \frac{\mathbb{E}_q[A_1(x)^2]}{n}.$$

Now replacing x with X and taking expectations gives

$$\mathbb{E}_q[(v^\top \varphi_{q,1:n}^*(X) - \mu(X))^2] = \frac{\mathbb{E}[A_1(X)^2]}{n} = \frac{1}{n} \text{var} \left(\frac{p_Z(Z_1)}{q_Z(Z_1)} \mathbb{E}_q[Y_1 | Z_1] g^*(X, Z_1) \right)$$

By Markov's inequality, the event

$$\mathbb{E}_q[(v^\top \varphi_{q,1:n}^*(X) - \mu_q(X))^2 | Z_1, \dots, Z_m] \leq \frac{2}{n} \text{var} \left(\frac{p_Z(Z_1)}{q_Z(Z_1)} \mathbb{E}_q[Y_1 | Z_1] g^*(X, Z_1) \right).$$

has probability at least $1/2$. We can analogously define such a ‘‘good’’ event for each block of coordinates. With probability at least $1 - \delta$, at least one of these good events occurs; in this event, we can pick any such ‘‘good’’ block, set the corresponding weights in w according to the construction above, and set the remaining weights in w to zero. This produces the desired guarantee. \blacksquare

The construction φ_q^* is not unique. Indeed, for any distribution α_Z over Z , we could sample Z_1, \dots, Z_m i.i.d. from α_Z and embed according to

$$\varphi_\alpha^*(x) := (g^*(x, Z_1), \dots, g^*(x, Z_m)).$$

The same arguments of Lemma 15 apply here as well to give us that with probability $1 - \delta$ there exists a $w \in \mathbb{R}^m$ satisfying

$$\mathbb{E}_q[(w^\top \varphi_\alpha^*(X) - \mu_q(X))^2 | Z_1, \dots, Z_m] \leq \frac{2}{\lfloor m/\log_2(1/\delta) \rfloor} \text{var} \left(\frac{p_Z(Z_1)}{\alpha(Z_1)} \mathbb{E}_q[Y_1 | Z_1] g^*(X, Z_1) \right)$$

where the variance is taken with respect to $X \sim q_X$ independently of Z_1 .

Given that φ_α^* produces an error bound analogous to the one in Lemma 15 for any valid distribution α , it is natural to ask whether one can find an α whose corresponding variance term is smaller than the variance induced by q . As the following lemma shows, there is not much room from improvement over q .

Lemma 16 *Let α_Z be any distribution over \mathcal{Z} . Suppose that $X \sim q_X$, $Z \sim q_Z$, and $\tilde{Z} \sim \alpha_Z$, all independently. Then*

$$\text{var} \left(\frac{p_Z(Z)}{q_Z(Z)} \mathbb{E}[Y | Z] g^*(X, Z) \right) \leq \text{var} \left(\frac{p_Z(\tilde{Z})}{\alpha_Z(\tilde{Z})} \mathbb{E}[Y | \tilde{Z}] g^*(X, \tilde{Z}) \right) + \text{var}(\mathbb{E}_q[Y | Z]).$$

Proof To simplify notation, let $f(z) := \mathbb{E}_q[Y | Z]$. Since $f(Z)g^*(X, Z)$ and $f(\tilde{Z})g^*(X, \tilde{Z})$ have the same mean, it suffices to show

$$\mathbb{E} (f(Z)g^*(X, Z))^2 \leq \mathbb{E} (f(\tilde{Z})g^*(X, \tilde{Z}))^2 + \text{var}(\mathbb{E}_q[Y | Z]).$$

Writing out expectations in integral form, we have

$$\begin{aligned}
 & \mathbb{E} \left(f(\tilde{Z}) \frac{p_Z(\tilde{Z})}{\alpha_Z(\tilde{Z})} g^*(X, \tilde{Z}) \right)^2 - \mathbb{E} \left(\frac{p_Z(Z)}{q_Z(Z)} f(Z) g^*(X, Z) \right)^2 \\
 &= \int \int f(z)^2 p_Z(z)^2 g^*(x, z)^2 q_X(x) \left(\frac{\alpha_Z(z)}{\alpha_Z(z)^2} - \frac{q_Z(z)}{q_Z(z)^2} \right) dx dz \\
 &= \int \int f(z)^2 \frac{p_{X,Z}(x, z)^2}{p_X(x)^2} q_X(x) \left(\frac{1}{\alpha_Z(z)} - \frac{1}{q_Z(z)} \right) dx dz \\
 &= \int f(z)^2 \left(\frac{1}{\alpha_Z(z)} - \frac{1}{q_Z(z)} \right) \left(\int p_{Z|X}(z | x)^2 q_X(x) dx \right) dz \\
 &\geq \int f(z)^2 \left(\frac{1}{\alpha_Z(z)} - \frac{1}{q_Z(z)} \right) \left(\int p_{Z|X}(z | x) q_X(x) dx \right)^2 dz \\
 &= \int f(z)^2 \left(\frac{1}{\alpha_Z(z)} - \frac{1}{q_Z(z)} \right) q_Z(z)^2 dz \\
 &= \int f(z)^2 \frac{q_Z(z)^2}{\alpha_Z(z)} dz - \int f(z)^2 q_Z(z) dz \\
 &\geq \frac{(\int f(z) q_Z(z) dz)^2}{\int \alpha_Z(z) dz} - \int f(z)^2 q_Z(z) dz \\
 &= \left(\int f(z) q_Z(z) dz \right)^2 - \int f(z)^2 q_Z(z) dz \\
 &= -\text{var}(f(Z)) = -\text{var}(\mathbb{E}_q[Y|Z]).
 \end{aligned}$$

In the above, the first inequality follows from Jensen's inequality, and the second inequality is Titu's lemma (a simple corollary of Cauchy-Schwarz). ■