Journal of Machine Learning Research 22 (2021) 1-31 Submitted 1/21; Revised 9/21; Published 10/21

Contrastive Estimation Reveals Topic Posterior Information to Linear Models

Christopher Tosh TOSHC@MSKCC.ORG
Department of Epidemiology and Biostatistics
Memorial Sloan Kettering Cancer Center
New York, NY 10065

Akshay Krishnamurthy AKSHAYKR@MICROSOFT.COM
Microsoft Research
New York, NY 10012

Daniel Hsu DJHSU@CS.COLUMBIA.EDU
Department of Computer Science and Data Science Institute
Columbia University
New York, NY 10027

Editor: David Sontag

Abstract

Contrastive learning is an approach to representation learning that utilizes naturally occurring similar and dissimilar pairs of data points to find useful embeddings of data. In the context of document classification under topic modeling assumptions, we prove that contrastive learning is capable of recovering a representation of documents that reveals their underlying topic posterior information to linear models. We apply this procedure in a semi-supervised setup and demonstrate empirically that linear classifiers trained on these representations perform well in document classification tasks with very few training examples.

Keywords: contrastive estimation, latent Dirichlet allocation, representation learning

1. Introduction

Using unlabeled data to find useful embeddings is a central challenge in representation learning. Classical approaches to this task often start by fitting some type of structure to the unlabeled data, such as a generative model or a dictionary, and then embed future data via inference with the fitted structure (Blei et al., 2003; Raina et al., 2007). While principled, this approach is not without its drawbacks. One issue is that learning structures and performing inference is often hard in general (Sontag and Roy, 2011; Arora et al., 2012). Another issue is that we must a priori choose a structure and method for fitting the unlabeled data, and unsupervised methods for learning these structures can be sensitive to model misspecification (Kulesza et al., 2014).

Contrastive learning (also called noise contrastive estimation, or NCE) is an alternative representation learning approach that tries to capture the latent structure in unlabeled data implicitly. At a high level, these methods formulate a classification problem in which the goal is to distinguish examples that naturally occur in pairs, called positive samples, from

randomly paired examples, called negative samples. The particular choice of positive samples depends on the setting. In image representation problems, for example, patches from the same image or neighboring frames from videos may serve as positive examples (Wang and Gupta, 2015; Hjelm et al., 2018). In text modeling, the positive samples may be neighboring sentences (Logeswaran and Lee, 2018; Devlin et al., 2018). The idea is that in the course of learning to distinguish between semantically similar positive examples and randomly chosen negative examples, we will capture some of the latent semantic information.

In this work, we look "under the hood" of contrastive learning and consider its application to document modeling, where the goal is to construct useful vector representations of text documents in a corpus. In this setting, there is a natural source of positive and negative examples: a positive example is simply a document from the corpus, and a negative example is one formed by pasting together the first half of one document and the second half of another (independently chosen) document.[1] We prove that when the corpus is generated by a topic model, learning to distinguish between these two types of documents yields representations that are closely related to their underlying latent variables.

One potential application of contrastive learning is in a semi-supervised setting, where there is a small amount of labeled data as well as a much larger collection of unlabeled data. In these situations, purely supervised methods that fit complicated models may have poor performance due to the limited amount of labeled data. On the other hand, when the labels are well-approximated by some function of the latent structure, our results show that an effective strategy is to fit linear functions, which may be learned with relatively little labeled data, on top of contrastive representations. In our experiments, we verify empirically that this approach produces reasonable results.

1.1 Contributions

The primary goal of this work is to shed light on what contrastive learning techniques uncover in the presence of latent structure. To this end, we focus on the setting of document modeling where latent structure is induced by a topic model. Here, our contrastive learning objective is to distinguish true documents from "fake" documents that are composed by randomly pasting together two document halves from the corpus. We consider two functional forms of solutions for this problem, both trained with logistic loss.

The first of these, on which our theoretical analysis will focus, consists of general functions of the form $f(\cdot, \cdot)$. Here, we have trained f so that $f(x, x')$ indicates the confidence of the model that x and x' are two halves of the same document. To embed a new document x using f, we propose a *landmark embedding* procedure: fix documents $l_1, \ldots, l_M$ (our so-called landmarks) and create the embedding $\phi(x)$ using a function of the predictions $f(x, l_1), \ldots, f(x, l_M)$. In Section 4, we show that the embedding $\phi(x)$ is a linear transformation of the underlying topic posterior moments of x. Moreover, under certain conditions this linear relationship is invertible, so that linear functions of $\phi(x)$ correspond to polynomial functions of the topic posterior moments of document x. In Section 5, we show that errors in f on the contrastive learning objective transfer smoothly to errors in $\phi(x)$ as a linear transformation of the topic posterior of x. Thus, as the quality of f improves, linear

1. In this context, "half" is used liberally; the two halves need not be of the same length.

functions of $\phi(x)$ more closely approximate polynomial functions of the topic posterior of document x.

Unfortunately, the landmark embedding can require quite a few landmarks before our theoretical results kick in. Moreover, embedding a single document requires M evaluations of f, which can be expensive. To circumvent this, in Section 7 we introduce a *direct embedding* procedure that more closely matches what is done in practice. We learn a predictor of the form $f_1(x)^\top f_2(x')$ where f_1, f_2 are functions with d-dimensional outputs, and we train this predictor using the same contrastive learning task as before. To embed a document x, we simply use the evaluation $f_1(x)$. In Section 7, we evaluate this embedding on a semi-supervised learning task, and we show that it has reasonable performance. Indeed, the direct embedding method generally outperforms the landmark embedding method, which raises the question of whether or not anything can be theoretically proven about the direct embedding method. We leave this question to future work.

1.2 Related Work

Reducing an unsupervised problem to a synthetically-generated supervised problem is a well-studied technique. In dynamical systems modeling, Langford et al. (2009) showed that the solutions to a few forward prediction problems can be used to track the underlying state of a dynamical system, generalizing ideas seen in autoregressive models for linear dynamical systems (Yule, 1927). Related ideas can also be found in other multi-view models (Ando and Zhang, 2005, 2007; Lee et al., 2020), which are related to techniques like canonical correlation analysis (Hotelling, 1936; Kakade and Foster, 2007; Foster et al., 2009). In anomaly/outlier detection, a useful technique is to learn a classifier that distinguishes between true samples from a distribution and fake samples from some synthetic distribution (Steinwart et al., 2005; Abe et al., 2006). Similarly, estimating the parameters of a probabilistic model can be reduced to learning to classify between true data and randomly generated noise (Gutmann and Hyvärinen, 2010).

In the context of natural language processing, methods such as skip-gram and continuous bag-of-words turn the problem of finding word embeddings into a prediction problem (Mikolov et al., 2013a b). Modern language representation training algorithms such as BERT and QT also use naturally occurring classification tasks such as predicting randomly masked elements of a sentence or discriminating whether or not two sentences are adjacent (Devlin et al., 2018; Logeswaran and Lee, 2018). Training these models often employs a technique called negative sampling, in which softmax prediction probabilities are estimated by randomly sampling examples; this bears close resemblance to the way that negative examples are produced in contrastive learning.

Most relevant to the current paper, Arora et al. (2019) gave a theoretical analysis of contrastive learning. They specifically study minimization of the contrastive loss

$$L(f) \;=\; \mathbb{E}_{x,x_+,x_-}[\ell\left(f(x)^\top(f(x_+) - f(x_-))\right)],$$

where (x, x_+) is a positive pair and (x, x_-) is a negative pair. 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 f yields good representations for the classification task of distinguishing latent classes.

The main difference between our work and that of Arora et al. (2019) is that we adopt a generative modeling perspective and induce the contrastive distribution *endogenously*, while Arora et al. do not make generative assumptions but do assume that the contrastive distribution is *directly* induced by the downstream classification task. The focus of our work is therefore complementary to theirs: we study the types of functions that can be succinctly expressed with the contrastive representation in our generative modeling setup. In addition, our results apply to semi-supervised regression, but it is unclear how to define their contrastive distribution in this setting; this makes it difficult to apply their results here. Finally, Arora et al. point out the method they study has limitations that arise when the number of latent classes is small and the probability of negative samples having the same class is high. The embedding method we study is quite different from the one studied by Arora et al., and this difference allows us to side-step the class collision issues.

There is a line of works on efficiently (and provably) recovering topic models from unlabeled data under certain structural assumptions (e.g., Arora et al., 2012; Anandkumar et al., 2012; Arora et al., 2013; Ding et al., 2013). Although we assume a topic modeling generative process in this paper, our goal is not to provide another topic recovery algorithm. We instead take a different path and examine contrastive learning in the context of a topic modeling generative process in order to provide some explanations for its recent successes in representation learning.

In a follow-up to the initial version of the present article, we studied the same embedding method proposed here under a multi-view redundancy condition (Tosh et al., 2021). That condition can be viewed as a non-linear generalization of redundancy conditions considered by Kakade and Foster (2007) and Foster et al. (2009). The focus in the present article is different, because here we provide concrete, model-specific semantics to the learned representation, whereas in the follow-up work, we study the predictive capacity of the representation under a specific assumption that involves the downstream task.

1.3 Paper Organization

In Section 3, we present our contrastive learning procedure and the algorithm to embed future documents. In Section 4, we show that under certain topic modeling assumptions, the document embeddings we construct from contrastive learning capture underlying topic structure. In Section 5, we analyze the errors that arise in the finite sample setting and show that whenever we can achieve low prediction error on the contrastive learning task, linear functions learned on the resulting representations must also be high quality. In Section 6, we validate our theoretical findings on a simulated topic recovery task, demonstrating that contrastive learning in our setting leads to recovery of topic posterior information. In Section 7, we apply our contrastive learning procedure to a semi-supervised document classification task. We show that these embeddings generally outperform several natural baselines, particularly in the scarce labeled data regime. All proofs are presented in Section 8.

2. Setup

Let $\mathcal{V}$ denote a finite vocabulary. A *topic* is a distribution over $\mathcal{V}$. We will assume that there are K such topics, and denote the corresponding distributions as $O(\cdot \mid k)$ for $k = 1, \ldots, K$. To generate a length m document x, one first draws a vector w from Δ^{K-1}, the probability

Algorithm 1 Contrastive Estimation with Documents

Input: Corpus $\mathcal{U}$ of unlabeled documents. **Initialize:** $S = \emptyset$.
for $i = 1, \ldots, n$ **do**

Sample x and $\tilde{x}$ independently from unif($\mathcal{U}$); $S \leftarrow S \cup \begin{cases} \{(x^{(1)}, x^{(2)}, 1)\} \text{ w.p. } 1/2 \\ \{(x^{(1)}, \tilde{x}^{(2)}, 0)\} \text{ w.p. } 1/2 \end{cases}$

end for
Solve the optimization problem

$$\hat{f} = \min_f \sum_{(x^{(1)}, x^{(2)}, y) \in S} y \log\left(1 + e^{-f(x^{(1)}, x^{(2)})}\right) + (1-y)\log\left(1 + e^{f(x^{(1)}, x^{(2)})}\right)$$

Select landmark documents $l_1, \ldots, l_M$ and embed $\hat{\phi}(x) = \left(\exp\left(\hat{f}(x, l_i)\right) : i \in [M]\right)$.

simplex in $\mathbb{R}^K$, and then samples each of the m words $x_1, \ldots, x_m$ by first sampling the latent variable z_i from the categorical distribution induced by w, that is $\mathbb{P}(z_i = k) = w_k$, and then drawing $x_i \sim O(\cdot \mid z_i)$. We note that documents are allowed to take different lengths.

We will also be interested in the case where each document has an associated label $\ell \in \mathbb{R}$. One natural restriction to make on a label is that it is conditionally independent of the document given the topic distribution of the document. Thus, we will assume that there is a joint distribution $\mathcal{D}$ of triples (x, w, ℓ), where (x, w) are generated according to the topic model described above, and then ℓ is drawn from some distribution conditioned on w. One of the goals of this paper is to characterize the functional forms of this conditional distribution that are most suited to contrastive learning.

In the representation learning approach to the semi-supervised setting, we are given a large collection $\mathcal{U}$ of documents with no labels, and a small collection $\mathcal{L}$ of labeled documents. Using $\mathcal{U}$, we learn a feature map $\hat{\phi}$ that will form the basis of our predictions. Then, using $\mathcal{L}$, we learn a simple predictor based on $\hat{\phi}$, such as a linear function, to predict the label ℓ given $\hat{\phi}(x)$.

3. Contrastive Learning Algorithm

In contrastive learning, examples come in the form of similar and dissimilar pairs of points, where the exact definition of similar/dissimilar depends on the task at hand. Our construction of similar pairs will take the form of randomly splitting a document into two documents, and our dissimilar pairs will consist of subsampled documents from two randomly chosen documents. In the generative modeling setup, since the words are i.i.d. conditional on the topic distribution, a natural way to split a document x into two is to call the first half of the words $x^{(1)}$ and the second half $x^{(2)}$.

The contrastive representation learning procedure is displayed in Algorithm 1. It uses a finite-sample approximation to the contrastive distribution $\mathcal{D}_{\text{contrast}}$ described as follows:

(a) sample a document x and partition it into $(x^{(1)}, x^{(2)})$,

(b) with probability 1/2 output $(x^{(1)}, x^{(2)}, 1)$,

(c) with probability 1/2, sample a second document $(\tilde{x}^{(1)}, \tilde{x}^{(2)})$ and output $(x^{(1)}, \tilde{x}^{(2)}, 0)$.

For $(x, x', y) \sim \mathcal{D}_{\text{contrast}}$, the parts x and x' are the two halves of a (possibly synthetic) document, and y is the binary label. Our contrastive learning objective is to minimize the binary cross-entropy loss of discriminating between positive and negative examples:

$$L_{\text{contrast}}(f) := \mathbb{E}_{(x,x',y)\sim\mathcal{D}_{\text{contrast}}}\left[y \log\left(1 + e^{-f(x,x')}\right) + (1-y)\log\left(1 + e^{f(x,x')}\right)\right]. \quad (1)$$

In our algorithm, we approximate this expectation via sampling and optimize the empirical objective, which yields an approximate minimizer $\hat{f}$ (chosen from some function class $\mathcal{F}$).

To see why optimizing this contrastive learning objective is so useful, let $f^\star$ be the global minimizer of Eq. (1), which is known to be the log-odds ratio, cf. Section 3 of Buja et al. (2005). By Bayes' theorem we have that $g^\star := \exp(f^\star)$ satisfies the following:

$$g^\star(x, x') := \exp(f^\star(x, x')) = \frac{\mathbb{P}(y = 1 \mid x, x')}{\mathbb{P}(y = 0 \mid x, x')} = \frac{\mathbb{P}(x^{(1)} = x, x^{(2)} = x')}{\mathbb{P}(x^{(1)} = x)\mathbb{P}(x^{(2)} = x')}.$$

Thus, $g^\star(x, x')$ captures the ratio of the probability of x and x' co-occurring as the first and second halves of the same document and the product of their marginal probabilities.

In Eq. (1), we have not imposed any constraints on the functions over which we are optimizing. Thus, we seek to extract a useful embedding from $g^\star$ using only *black box access* to $g^\star$. Our approach is to select some set of fixed documents, which we call *landmarks*, and to embed by utilizing the predictions of $g^\star$ on these landmarks.

Formally, we select documents $l_1, \ldots, l_M$ and represent document x as[2]

$$\phi^\star(x) := (g^\star(x, l_1), \ldots, g^\star(x, l_M)). \quad (2)$$

This yields the final document-level representation, which can be used for downstream tasks. As we shall see in Section 4, when the documents have an underlying topic structure, $\phi^\star(x)$ is related to the posterior information of the topics by a linear transformation and this linear transformation is invertible whenever the landmarks $l_1, \ldots, l_M$ are sufficiently diverse.

In practice, we only have access to an approximate minimizer $\hat{f}$ of Eq. (1). Thus, we define an embedding based only on $\hat{f}$ (or $\hat{g}$, a corresponding approximation of $g^\star$) by

$$\hat{\phi}(x) := (\exp(\hat{f}(x, l_1)), \ldots, \exp(\hat{f}(x, l_M))) = (\hat{g}(x, l_1), \ldots, \hat{g}(x, l_M)). \quad (3)$$

In Section 5 we will see that, under some mild assumptions, our claims about $\phi^\star$ also hold true for $\hat{\phi}$ up to some small errors.

We make two observations about the computational complexity of this approach. First, the optimization problem in Eq. (1) is highly non-convex for many function classes. Nevertheless, first-order optimization methods such as gradient descent routinely find reasonable solutions to such problems in practice. As we will show in Section 6 and Section 7, this

2. Strictly speaking, we should first partition $x = (x^{(1)}, x^{(2)})$, only use landmarks that occur as second-halves of documents, and embed $x \mapsto (g^\star(x^{(1)}, l_1), \ldots, g^\star(x^{(1)}, l_M))$. For the sake of clarity, we will ignore this small technical issue here and in the remainder of the paper.

appears to be true in our situation as well. Second, embedding a single document in the landmark approach requires M evaluations of the learned function $\hat{f}$. Thus, embedding a large corpus with a reasonably large value of M can be quite expensive. Partly motivated by this observation, as well as by what is typically done in practice, Section 7 introduces an alternative approach that only requires a single evaluation of a learned (multidimensional-valued) function in order to embed a document.

Finally, we point out that there is nothing special about the binary cross-entropy loss. We may replace this loss in Eq. (1) with any proper scoring rule (Shuford et al., 1966; Buja et al., 2005), so long as the appropriate non-linear transformation is applied to the resulting predictions.

4. Recovering Topic Structure

In this section, we focus on the expressiveness of our contrastive representation, showing that polynomial functions of the topic posterior can be represented as *linear* functions of the representation. To do so, we ignore statistical issues and assume that we have access to the oracle representations $g^\star(x, \cdot)$. In Section 5, we address statistical issues.

Recall the generative topic model process for a document x. We first draw a topic vector $w \in \Delta^{K-1}$. Then for each word $i = 1, \ldots, \text{length}(x)$, we draw $z_i \sim \text{Categorical}(w)$ and $x_i \sim O(\cdot \mid z_i)$. We will show that when documents are generated according to the above model, the embedding of a document x in Eq. (2) is closely related its underlying topic vector w.

4.1 The Single Topic Case

To build intuition for the embedding in Eq. (2), we first consider the case where each document's probability vector w is supported on a single topic, i.e., $w \in \{e_1, \ldots, e_K\}$ where e_i is the i^{th} standard basis element. Then we have the following lemma.

Lemma 1 *For any documents x, x', we can write $g^\star(x, x') = \eta(x)^\intercal \psi(x')$, where*

$$\eta(x)_k := \mathbb{P}(w = e_k \mid x^{(1)} = x)$$

is the topic posterior distribution and

$$\psi(x')_k := \mathbb{P}(x^{(2)} = x' \mid w = e_k)/\mathbb{P}(x^{(2)} = x').$$

The characterization from Lemma 1 shows that $g^\star$ contains information about the posterior topic distribution $\eta(\cdot)$. To recover it, we must make sure that the $\psi(\cdot)$ vectors for our landmark documents span $\mathbb{R}^K$. Formally, if $l_1, \ldots, l_M$ are the landmarks, and we define the matrix $L \in \mathbb{R}^{K \times M}$ by

$$L := \begin{bmatrix} \psi(l_1) & \cdots & \psi(l_M) \end{bmatrix}, \tag{4}$$

then our representation satisfies $\phi^\star(x) = L^\intercal \eta(x)$. If our landmarks are chosen so that L has rank K, then $L^\dagger \phi^\star(x) = \eta(x)$, where $\dagger$ denotes the matrix pseudo-inverse. Thus, there is a linear transformation of $\phi^\star(x)$ that recovers the posterior distribution of w given x.

There are two observations to be made here. The first is that this argument naturally generalizes beyond the single topic setting to any setting where w can take values in a

finite set S, which may include some mixtures of multiple topics, though the number of landmarks needed would grow at least linearly with $|S|$. The second is that we have made no use of the structure of $x^{(1)}$ and $x^{(2)}$, except for that they are independent conditioned on w. Thus, this argument applies to more exotic ways of partitioning a document beyond the bag-of-words approach.

4.2 The General Setting

In the general setting, we allow document vectors to be any probability vector in Δ^{K-1}, and we do not hope to recover the full posterior distribution over Δ^{K-1}. However, the intuition from the single topic case largely carries over, and we will show that we can still recover the posterior moments.

Let $m_{\max}$ be the length of the longest landmark document. Let

$$S_m := \left\{ \alpha \in \mathbb{Z}_+^K : \sum_{k=1}^K \alpha_k = m \right\}$$

denote the set of non-negative integer vectors that sum to m and let $S_{\leq m_{\max}}$ denote the union of $S_1, S_2, \ldots, S_{m_{\max}}$. Let $\pi(w)$ denote the degree-$m_{\max}$ monomial vector in w:

$$\pi(w) := (w_1^{\alpha_1} \cdots w_k^{\alpha_k} : \alpha \in S_{\leq m_{\max}}).$$

For a positive integer m and a vector $\alpha \in S_m$, we let

$$\binom{[m]}{\alpha} := \left\{ z \in [K]^m : \sum_{i=1}^m \mathbb{1}[z_i = k] = \alpha_k \quad \forall k \in [K] \right\}.$$

For a document x of length m, the degree-m polynomial vector ψ_m is defined by

$$\psi_m(x) \ := \ \left(\sum_{z \in \binom{[m]}{\alpha}} \prod_{i=1}^m O(x_i \mid z_i) : \alpha \in S_m \right),$$

and let $\psi_d(x) = \vec{0}$ for all $d \neq m$. The cumulative polynomial vector ψ is given by

$$\psi(x) \ := \ \frac{1}{\mathbb{P}(x^{(2)} = x)} (\psi_0(x), \psi_1(x), \cdots, \psi_{m_{\max}}(x)). \tag{5}$$

Given these definitions, we have the following general case analogue of Lemma 1.

Lemma 2 *For any documents x, x', we may write $g^\star(x, x') = \eta(x)^\mathsf{T} \psi(x')$ where $\eta(x) := \mathbb{E}[\pi(w) \mid x^{(1)} = x]$ and ψ is defined in Eq (5).*

Thus, we again have $\phi^\star(x) = L^\mathsf{T} \eta(x)$, but the columns of L are now vectors $\psi(l_i)$ from Eq. (5).

Our analysis, so far, shows that if we choose the landmarks such that LL^T is invertible, then our representation captures all moments of the topic posterior up to degree $m_{\max}$. As the next theorem shows, we can ensure that LL^T is invertible whenever each topic has an

associated *anchor word* (Arora et al., 2012), i.e., each topic k satisfies that there is some word j satisfying $O(j \mid k) > 0$ and $O(j \mid k') = 0$ for all $k' \neq k$; word j is said to be an anchor word for topic k in this case. Under this anchor word assumption, there is a set of landmarks $l_1, \ldots, l_M$ such that any polynomial of $\eta(x)$ can be expressed as a linear function of $\phi^\star(x)$.

Theorem 3 *Suppose that (i) each topic has an associated anchor word, and (ii) the marginal distribution of w has positive probability on the interior of Δ^{K-1}. For any $d_o \leq m_{\max}$, there is a collection of $M = O(K^{d_o})$ landmark documents $l_1, \ldots, l_M$ such that if $Q(w)$ is a degree-d_o polynomial in w, then there is a vector $\theta \in \mathbb{R}^M$ such that $\langle \theta, \phi^\star(x) \rangle = \mathbb{E}[Q(w) \mid x^{(1)} = x]$ for all documents x.*

Coupling Theorem 3 with the Stone-Weierstrass theorem (Stone, 1948) shows that, in principle, the posterior mean of any continuous function of w can be approximated using our representation.

5. Error Analysis

Given a finite amount of data, we cannot hope to solve Eq. (1) exactly. Thus, our solution $\hat{f}$ will only be an approximation to $f^\star$. Since $\hat{f}$ is the basis of our representation, one may worry that errors incurred in this approximation will cascade and cause the approximate representation $\hat{\phi}(x)$ to differ so wildly from $\phi^\star(x)$ that the results of Section 4 do not even approximately hold.

In this section, we will show that, under certain conditions, such fears are unfounded. Specifically, we will show that there is an error transformation from the approximation error of $\hat{f}$ to the approximation error of linear functions in $\hat{\phi}$. That is, if the target function is $\eta(x)^\intercal \theta^\star$, then we will show that the best mean squared error achievable using our approximate representation $\hat{\phi}$ (from Eq. (3)), given by

$$R(\hat{\phi}) := \min_v \mathbb{E}_{x \sim \mu^{(1)}} (\eta(x)^\intercal \theta^\star - \hat{\phi}(x)^\intercal v)^2,$$

is bounded in terms of the approximation quality of $\hat{f}$ as well as some other terms. Here, $\mu^{(1)}$ is the marginal distribution over first halves of documents drawn from $\mathcal{D}$. Thus, for the specific setting of semi-supervised learning, an approximate solution to Eq. (1) is good enough.

There are a number of reasonable ways to choose landmark documents. Here we consider a simple method: randomly sample them from $\mu^{(2)}$, the marginal distribution of $x^{(2)}$. We will assume that this distribution satisfies certain regularity properties.

Assumption 1 *There is a constant $\sigma_{\min} > 0$ such that for any $\delta \in (0, 1)$, there is a number M_0 such that for an i.i.d. sample $l_1, \ldots, l_M$ from $\mu^{(2)}$, with $M \geq M_0$, with probability $1 - \delta$, the matrix L in Eq. (4) (with ψ as defined in Lemma 1 or Eq. (5)) has minimum singular value at least $\sigma_{\min}\sqrt{M}$.*

Note that the smallest non-zero singular value of $\frac{1}{\sqrt{M}} L$ is the square-root of the smallest eigenvalue of a certain empirical second-moment matrix. Hence, Assumption 1 holds under

appropriate conditions on the landmark distribution, for instance via tail bounds for sums of random matrices (Tropp, 2012) combined with matrix perturbation analysis (e.g., Weyl's inequality). In the single topic setting with anchor words, it can be shown that for long enough documents, $\sigma_{\min}$ is lower-bounded by a constant when M_0 grows polynomially with K. We defer a detailed proof of this to Section 8.2.

Notice that the magnitude of the predictions made by $\hat{f}$ and $f^\star$ affect the scale of the embeddings $\hat{\phi}$ and $\phi^\star$, which in turn can affect the difficulty of approximating $\eta(x)^\mathsf{T}\theta^\star$. To control this, we will need to assume that the predictions of $\hat{f}$ and $f^\star$ are bounded above by some finite value.

Assumption 2 *There exists some $g_{\max} > 0$ such that each of $\hat{f}(x, l_i)$ and $f^\star(x, l_i)$ is at most $\log g_{\max}$, for all documents x and landmarks l_i.*

Note that if Assumption 2 holds for $f^\star$, then it can be made to hold for $\hat{f}$ by clipping. Moreover, it holds for $f^\star$ whenever the vocabulary and document sizes are constants:

$$f^\star(x, x') = \log \frac{\mathbb{P}(x^{(1)} = x, x^{(2)} = x')}{\mathbb{P}(x^{(1)} = x)\mathbb{P}(x^{(2)} = x')} = \log \frac{\mathbb{P}(x^{(2)} = x' \mid x^{(1)} = x)}{\mathbb{P}(x^{(2)} = x')} \leq \log \frac{1}{\mathbb{P}(x^{(2)} = x')}.$$

Since landmarks are sampled, and the number of possible documents is finite, there exists a constant $p_{\min} > 0$ such that $\mathbb{P}(x^{(2)} = l) \geq p_{\min}$. Thus, Assumption 2 holds for $g_{\max} \leq 1/p_{\min}$.

Given these assumptions, we have the following error transformation guarantee.

Theorem 4 *Fix any $\delta \in (0, 1)$, and suppose Assumption 1 and Assumption 2 hold (with M_0, $\sigma_{\min}$, and $g_{\max}$). Let $\hat{f}$ be the function returned by the contrastive learning algorithm, and let $\varepsilon := L_{\text{contrast}}(\hat{f}) - L_{\text{contrast}}(f^\star)$ denote its excess contrastive loss. If $M \geq M_0$, then with probability at least $1 - \delta$ over the random sample of $l_1, \ldots l_M$,*

$$R(\hat{\phi}) \leq \frac{\|\theta^\star\|_2^2 \, (1 + g_{\max})^4}{\sigma_{\min}^2} \left(\varepsilon + \sqrt{\frac{2 \log(2/\delta)}{M}} \right).$$

We make a few observations here. First, $\|\theta^\star\|_2^2$ is a measure of the complexity of the target function. Thus, if the target function is some reasonable function (e.g., a low-degree polynomial) of the posterior document vector, then we can expect $\|\theta^\star\|_2^2$ to be small. Second, the dependence on $g_{\max}$ is probably not very tight and can likely be improved. Third, note that M can grow and ε can shrink with the number of *unlabeled* documents; indeed, none of the terms in Theorem 4 deal with labeled data.

Finally, it is possible to establish guarantees in a semi-supervised setting using our analysis. If we have n_L i.i.d. labeled examples, and we learn a linear predictor $\hat{v}$ with the representation $\hat{\phi}$ using ERM (say), then the bias-variance decomposition grants

$$\text{mse}(\hat{v}) := \mathbb{E}_{x \sim \mu^{(1)}}(\eta(x)^\mathsf{T}\theta^\star - \hat{\phi}(x)^\mathsf{T}\hat{v})^2 = R(\hat{\phi}) + \mathbb{E}_{x \sim \mu^{(1)}}(\hat{\phi}(x)^\mathsf{T}(v^* - \hat{v}))^2,$$

where v^* is the minimizer of $\text{mse}(\cdot)$. The final term $\mathbb{E}_{x \sim \mu^{(1)}}(\hat{\phi}(x)^\mathsf{T}(v^* - \hat{v}))^2$ is the excess risk in linear regression, which goes to zero as $n_L \to \infty$.

6. Topic Modeling Simulations

To test our theory, we ran simulation experiments where we trained neural networks in the style of Algorithm 1 on synthetic data generated from topic models and tested how well the resulting landmark representations could reconstruct the underlying document-topic vectors.

6.1 Extracting the Document-topic Vectors

Recall the generative model of a document x:

- Draw a weight vector $w \in \Delta^{K-1}$.
- For $i = 1, \ldots, n$, sample $z_i \sim \text{Categorical}(w)$ and draw $x_i \sim O(\cdot \mid z_i)$.

Given an approximation $\hat{g}$ of $g^\star$, the results of Section 4 offer a natural way to approximate $\mathbb{E}[w \mid x]$ using the document x and knowledge of the true topics $O(\cdot \mid \cdot)$. First, choose landmark documents $l_1, \ldots, l_M$ to be *single-word* documents, that is documents of length one consisting of one word chosen from the vocabulary. Then, create the landmark embedding $\hat{\phi}$ according to Eq. (3). Finally, we construct the matrix

$$L := \begin{bmatrix} \frac{O(l_1|1)}{P(l_1)} & \cdots & \frac{O(l_M|1)}{P(l_M)} \\ \vdots & \ddots & \vdots \\ \frac{O(l_1|K)}{P(l_1)} & \cdots & \frac{O(l_M|K)}{P(l_M)} \end{bmatrix}$$

and compute $\hat{\eta}(x) = L^\dagger \hat{\phi}(x)$. The results of Section 4 imply that if L is well-conditioned, then we should have $\hat{\eta}(x) \approx \mathbb{E}[w \mid x]$. In many of our simulation settings, we have access to w, but not to $\mathbb{E}[w \mid x]$. To give a better approximation of w, which is a probability vector, we take the extra step of projecting, in ℓ_2 distance, $\hat{\eta}(x)$ onto the probability simplex Δ^{K-1}.

6.2 Methodology

To solve the optimization problem in Eq. (1), we trained three-layer neural networks with fully-connected layers using 512 nodes per hidden layer. We used ReLU nonlinearities, layer normalization, and the default PyTorch initialization (Paszke et al., 2019). We optimized using Adam with learning rate 10^{-3} for 1k epochs and then continued training using SGD with learning rate 10^{-4} for 100 epochs, the principle of switching from Adam to SGD being suggested by Keskar and Socher (2017). At every epoch, we resampled fresh training data from the generative model, varying the number of documents as a parameter to be explored. In our figures, the number of documents we sampled at each epoch is denoted by N.

For most of our experiments, the underlying document-topic vector w was drawn from a symmetric Dirichlet(β) distribution. We also performed experiments in the single-topic setting, where w was sampled uniformly from the set of one-hot vectors $\{e_1, \ldots, e_K\}$. The lengths of the documents were sampled from a Poisson(100) distribution. As our landmark documents are restricted to single-word documents, we created the document splits by letting $x^{(2)}$ consist of a single randomly selected word from the document, and letting $x^{(1)}$

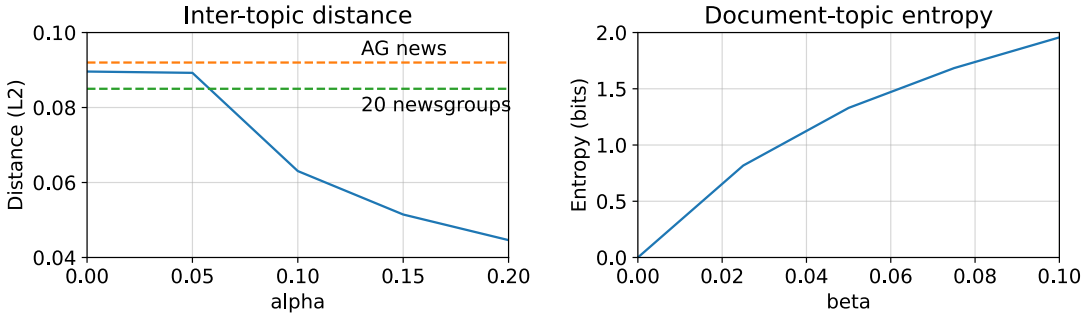

Figure 1: Inter-topic distances and document-topic vector entropies. The solid line in the left panel refers to the fully-synthetic setting. The horizontal dashed lines refer to the topics from the AG News dataset and the 20 Newsgroup datasets. Recall that $\alpha = 0.0$ denotes the block model setting, and $\beta = 0.0$ denotes the single-topic setting.

contain the remainder of the words.[3] To construct the embedding in Eq. (3), we randomly sampled 1k words from the vocabulary according to their frequency.

In all of our experiments, we used $K = 20$ topics. We considered two settings for our topic distributions: *fully synthetic* and *semi-synthetic*.

- In the fully synthetic setting, we sampled the topics independently from a symmetric Dirichlet(α) distribution over a vocabulary of size 5k. Note that the guarantees of Section 4 and Section 5 only apply to recovery of the posterior mean $\mathbb{E}[w \mid x]$ and not the vector w itself. Unfortunately, computing $\mathbb{E}[w \mid x]$, even when the generating topics are known, seems to be intractable in general. Thus, we also considered another fully synthetic setting, which we call the *block model* setting, where each word in the vocabulary is randomly assigned to a single topic, and the topics are uniform over the words assigned to them. In the block model setting, $\mathbb{E}[w \mid x]$ can be computed in closed form, allowing us to directly measure recovery of the posterior means.

- In the semi-synthetic setting, we fit an LDA topic model using batch variational Bayes (Blei et al., 2003) on a dataset with scikit-learn's default choices of $\alpha = \beta = 1/K$ and used the resulting topics to generate the data. We considered two datasets for fitting the topic models: AG News, described in more detail in Section 7, and the classical 20 Newsgroups dataset. After standard preprocessing, these datasets had vocabularies of size 16692 and 19148, respectively.

Note that in the fully synthetic setting, the Dirichlet parameter α governs the sparsity of the topic distributions, effectively determining the similarity of the topics: as α increases the prior concentrates on the interior of the simplex, forcing the topic distributions to be more similar. This is visualized in the left panel of Figure 1, where we display the expected

3. We also experimented with evenly splitting the documents, but the method presented here led to a small but noticeable improvement in reconstruction performance, possibly due to our restriction of landmark documents.

average inter-topic ℓ_2 distance

$$\mathbb{E}\left[\frac{1}{K(K-1)}\sum_{k\neq k'}\|O(\cdot \mid k) - O(\cdot \mid k')\|_2\right]$$

for independently drawn $O(\cdot \mid k) \sim \text{Dirichlet}(\alpha)$ as a function of α. Thus, we would expect topic recovery to be harder as α increases. For comparison, the left panel of Figure 1 also shows the average inter-topic ℓ_2 distance between the topic distributions that we obtained from fitting LDA topic models to the AG News and 20 Newsgroups datasets.

Similarly, in both the fully-synthetic and semi-synthetic settings, the Dirichlet parameter β determines the sparsity of the document-topic vectors w, with smaller values of β being more sparse. This is visualized in the right panel of Figure 1, where we display the expected entropy

$$\mathbb{E}\left[\sum_{k=1}^{K} w_k \log_2 \frac{1}{w_k}\right]$$

for $w \sim \text{Dirichlet}(\beta)$ as a function of β. One would expect that as β increases, the topic recovery task should become harder, up to a point. At some point, the underlying topic-vectors should be close to uniform, and thus predicting the uniform vector (or approximately the uniform vector) should lead to reasonable results.

We tracked two metrics of recovery. The first of these is the top topic accuracy, i.e., the probability of recovery of the underlying top topic; a successful recovery is declared if the highest probability entry in our predicted vector matched the highest probability entry in the underlying document-topic vector. The second is the ℓ_2 distance between the predicted vector and the underlying document-topic vector. For both of these metrics, we used a heldout test set of 5k documents. All results in this section were averaged over 5 independent runs. For heatmap plots, we adjusted color scales to remain consistent across tables.

6.3 Results

6.3.1 Fully Synthetic Setting

The left two panels of Figure 2 show the results for the fully synthetic setting where 120k documents were sampled at every epoch. Here we see the general trend of recovery becoming more difficult as α increases or β increases. For the top topic accuracy metric, this trend appears to be nearly monotonic. Note that random guessing would lead to a top topic accuracy of 0.05. For the ℓ_2 recovery metric, we see that there is a slight deviation in this trend, particularly for large values of α and β. We suspect that this is because when both α and β are large, the model may not be confident in its predictions which leads to reconstructed vectors that are close to uniform and therefore are also reasonable approximations of the underlying vector. Meanwhile, when only α is large, and β is smaller, our model may make confident but wrong predictions, leading to worse ℓ_2 recovery.

The middle two panels of Figure 2 show the results for block model setting where we varied β and the number of training documents sampled in every epoch. The top panel displays the recovery of the underlying top topic. However, the bottom panel displays ℓ_2

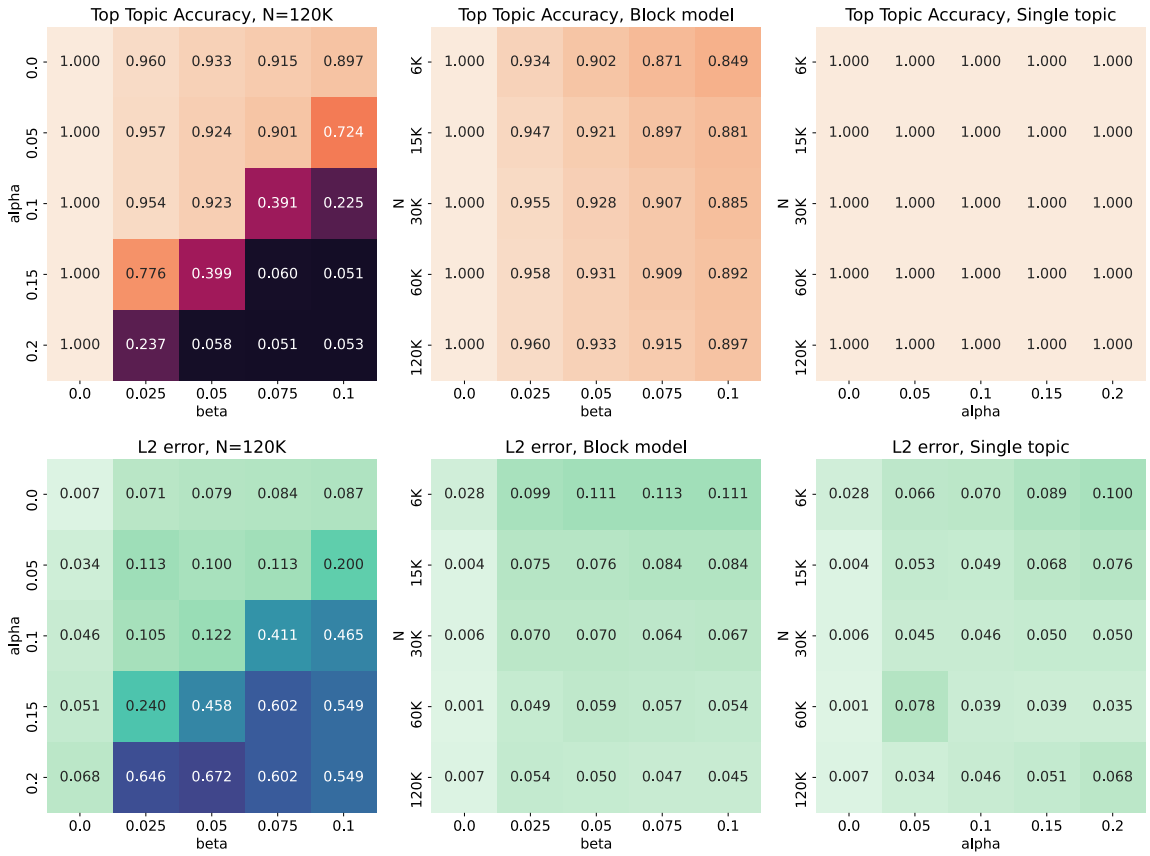

Figure 2: Synthetic topic modeling results. *Top row*: accuracy of recovering the underlying top topic. *Bottom row*: ℓ_2 error of recovering the entire document-topic vector. For the block model setting, this is recovery of the posterior mean of this vector, while for the other two settings it is recovery of the true underlying vector. *Left column*: α and β were varied as 120k documents were sampled at every epoch. *Middle column*: N and β were varied in the block model setting. *Right column*: N and α were varied in the single-topic setting. In all figures, $\alpha = 0$ denote the block model setting, and $\beta = 0$ denotes the single-topic setting.

recovery of the posterior mean vector. In both panels, we again see the general trend where recovery is easier for smaller values of β. Moreover, we also see that as the amount of training data increases, recovery in both metrics generally improves.

The right two panels of Figure 2 show the results for single-topic setting where we varied α and the number of training documents sampled in every epoch. Both panels display recovery with respect to the underlying document-topic vector, which is one-hot in this setting. As seen in the top panel, recovery of the top topic is relatively easy in this setting, even for smaller amounts of training data. The bottom panel shows that, even though the correct topic is consistently recovered, recovery in ℓ_2 distance is somewhat more challenging, although this too generally seems to improve for larger training set sizes.

Figure 3: Semi-synthetic topic modeling results. *Top row*: accuracy of recovering the underlying top topic. *Bottom row*: ℓ_2 error of recovering the entire underlying document-topic vector. *Left column*: varying N and β on the AG News semi-synthetic dataset. *Right column*: varying N and β on the 20 Newsgroups semi-synthetic dataset. In all plots, $\beta = 0.0$ denotes the single-topic setting.

6.3.2 Semi-synthetic Setting

Figure 3 shows the results for the semi-synthetic setting. The results here are very similar in spirit to those of the fully synthetic settings, where we observe better performance for larger datasets and worse performance for larger values of β. Moreover, comparing the last rows of the semi-synthetic tables to the left two panels of Figure 2, we see that, for a fixed dataset size, recovery in the semi-synthetic settings appears to be easier than in the purely synthetic setting for many values of α. This is despite the fact that the vocabularies in the semi-synthetic settings are approximately 3-4x larger than those used in the synthetic setting. However, the situation is perhaps less surprising in light of Figure 1, which shows large inter-topic distances for both AG News and 20 Newsgroups topic models.

7. Semi-supervised Experiments

We also conducted experiments with our document-level contrastive representations in a semi-supervised setting. The goal of these experiments is to demonstrate that the con-

trastive representations yield non-trivial performance, as consistent with the theory. Note that our intention is *not* to show state-of-the-art performance using contrastive learning; that is beyond the scope of the paper.

7.1 A Closely Related Representation

In the worst-case, the guarantees from Section 4 and Section 5 require the number of landmarks to be quite large. To develop a more practical representation, and to more closely mirror what is done in practice, we consider training models of the form $f_1, f_2 : \mathcal{X} \to \mathbb{R}^d$ where $(x, x') \mapsto f_1(x)^\mathsf{T} f_2(x')$. Plugging this into Eq. (1), we solve the following bivariate optimization problem:

$$\min_{f_1, f_2} \mathbb{E}_{\mathcal{D}_{\text{contrast}}} \left[y \log\left(1 + \exp\left(-f_1(x)^\mathsf{T} f_2(x')\right)\right) + (1-y) \log\left(1 + \exp\left(f_1(x)^\mathsf{T} f_2(x')\right)\right) \right]. \tag{6}$$

Given f_1, f_2, we embed a document x according to $f_1(x)$. We call the resulting scheme the *direct embedding* approach (and label it `Direct-NCE`) to distinguish it from the *landmark embedding* approach (labeled `Landmark-NCE`) from Section 3.

7.2 Methodology

To solve the optimization problems in Eqs. (1) and (6), we trained three-layer neural networks with fully-connected layers using 512 nodes per hidden layer. We used ReLU nonlinearities, layer normalization, and the default PyTorch initialization (Paszke et al., 2019). We optimized using Adam with learning rate 10^{-3} and trained until convergence (between 100-500 epochs, depending on the dataset). On each unlabeled dataset, we held out 10% of documents as a validation set and used early stopping based on the contrastive loss on this validation set to choose the final model.

We considered two ways to generate document "halves" from which the positive and negative pairs were created: randomly partitioning a document in half and randomly sampling from the empirical distribution of words in a document. We tested both of these methods as well as a third that alternates between these two method with equal probability. We found that this last method generally outperformed the others. Thus, unless indicated otherwise, `Landmark-NCE` and `Direct-NCE` were trained by alternating between randomly partitioning and sampling from the empirical distribution with replacement. In the appendix, we provide some experiments comparing these different methods.

We compared our representations, `Landmark-NCE` and `Direct-NCE`, against several embedding baselines.

- `BOW`—The standard bag-of-words representation.
- `BOW+SVD`—A bag-of-words representation with dimensionality reduction. We perform SVD on the bag-of-words representation using the unsupervised dataset to compute a low dimensional subspace. Bag-of-words documents are then projected onto this subspace.
- `LDA`—A representation derived from LDA. We fit LDA on the unsupervised dataset using batch variational Bayes (Blei et al., 2003), and our representation is the in-

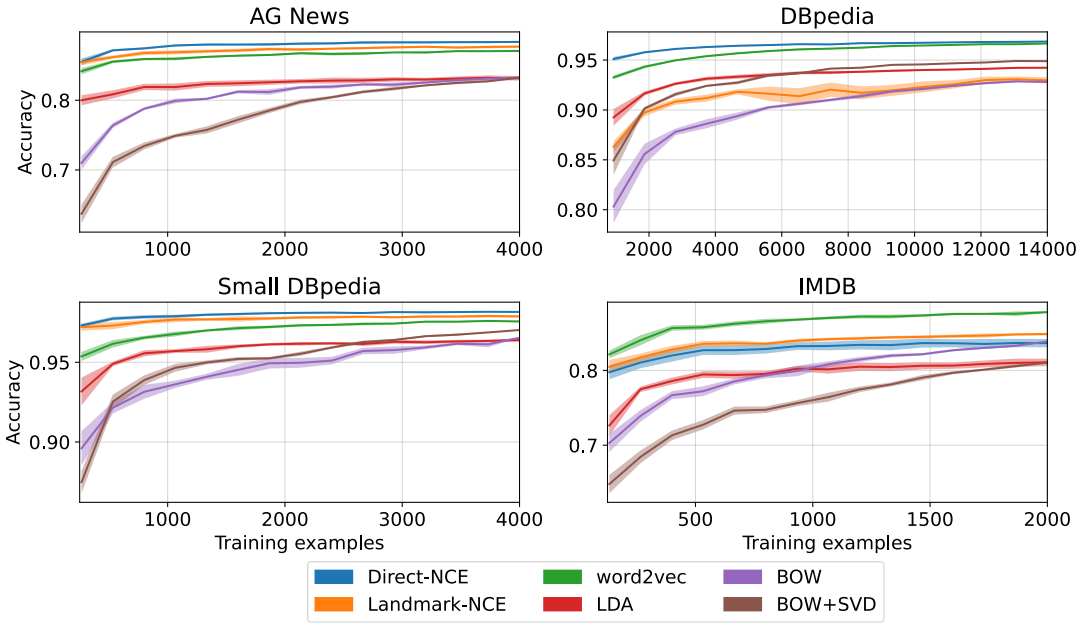

Figure 4: Semi-supervised learning comparisons demonstrating test accuracy of methods as number of supervised training examples increases.

ferred posterior distribution over topics given a document. For the topic modeling hyperparameters, we used the scikit-learn's default choices of $\alpha = \beta = 1/\#\text{ of topics}$.[4]

- `word2vec`—Skip-gram word embeddings (Mikolov et al., 2013b). We fit the skip-gram word embeddings model on the unsupervised dataset. To obtain a document-level representation, we averaged the word embeddings for each word in a document.

We note that all of these methods ignore word order in the final document-level representation, and all of them (with the exception of `word2vec`) ignore word order in their training. Each of these methods, with the exception of `BOW`, comes with a choice of embedding dimension. We tested all methods using embeddings of dimensions 100, 500, and 1000. As inference with LDA can be difficult with a large number of topics, we additionally tested LDA embeddings with 50 dimensions. All experiments in this section were performed using ten independent replicates, and shaded regions correspond to 95% confidence intervals. In all comparison plots where dimensions are not explicitly stated, we took the average performance of each method under each embedding dimension and plotted the point-wise maximum average accuracy.

For all methods, we used ℓ_2-regularized logistic regression to fit a linear classifier on the labeled data, where the regularization parameter was chosen using three-fold cross-validation.

4. We found that tuning the hyperparameters using perplexity on a held-out validation set did not lead to a noticeable improvement in downstream performance.

7.3 Datasets

We evaluated our methods on four datasets, each of which we preprocessed by filtering out uncommon words by document frequency.

- The AG news topic classification dataset is a collection of short news articles. The dataset as compiled by Zhang et al. (2015) has 4 classes (world, sports, business, and science/technology), 30k training examples per class, and 1900 test examples per class. We randomly selected 1k examples per class from the training set as labeled training data and used the remaining training examples for representation learning. After filtering, the vocabulary was of size 16692.

- The IMDB movie review sentiment classification dataset (Maas et al., 2011) is a collection of movie reviews that are classified as either positive or negative in sentiment. The dataset comes pre-separated into an unlabeled dataset of 50k examples, a training set of 12.5k examples per class, and a test set of 12.5k examples per class. We held out 1k examples from each class in the training set as labeled training data and 3.5k examples from each class in the test set as labeled testing data and used the remaining 91k examples for representation learning. After filtering, the vocabulary was of size 35648.

- The DBpedia ontology dataset consists of short descriptions of entities extracted from Wikipedia articles. As compiled by Zhang et al. (2015), the dataset has 14 classes, each with 40k training examples and 5k test examples. We randomly selected 1k examples per class from the training set as labeled training data and used the remaining training examples for representation learning. After filtering, the vocabulary was of size 25707.

- We created the Small DBpedia ontology dataset by taking the subset of the DBpedia dataset corresponding to 4 classes—company, artist, athlete, and office holder. We again held out 1k examples per class from the training set as labeled training data and used the remaining training examples for representation learning. After filtering, the vocabulary was of size 10416.

For `Landmark-NCE`, we randomly selected the landmark documents from the unlabeled data on all datasets.

7.4 Results

In the top left panel of Figure 4, we visualize the semi-supervised performance of `NCE` and the baselines on the AG news dataset. `Direct-NCE` outperforms all the other methods, with dramatic improvements over all except `Landmark-NCE` and `word2vec` in the low labeled data regime. `BOW` is reasonably competitive when there is an abundance of labeled data, but as the dimensionality of this representation is quite large, it performs poorly with limited samples. However, unsupervised dimensionality reduction on this representation appears to be unhelpful and actually degrades performance uniformly. Finally, we point out that word embedding representations (`word2vec`) perform quite well, but our document-level `Direct-NCE` and `Landmark-NCE` procedures are slightly better. This may reflect some

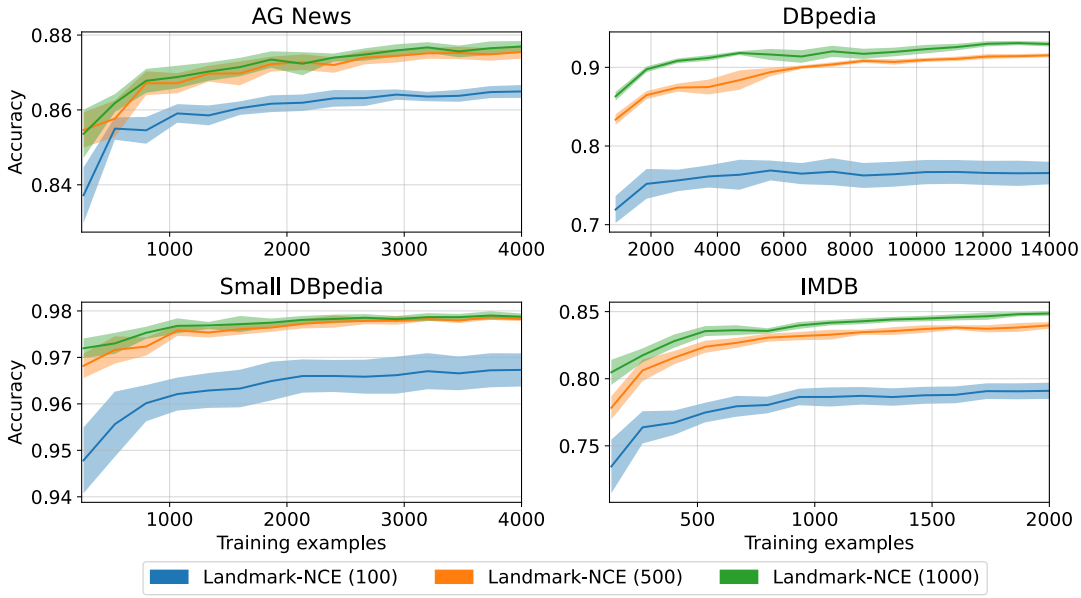

Figure 5: Semi-supervised performance of `Landmark-NCE` as the number of landmarks are varied. Number of landmarks used is in parentheses.

advantage in learning document-level non-linear representations, as opposed to averaging word-level ones.

The top right and bottom left panels of Figure 4 tell a similar story on the DBpedia and Small DBpedia ontology dataset. Again, we see that `Direct-NCE` outperforms the other methods, with good performance by `word2vec`. However, `Landmark-NCE` does markedly worse than some of the baselines on the full DBpedia dataset, while performing quite well on the Small DBpedia dataset. We conjecture that this may be due to the larger number of classes, which could indicate richer topic structure and thus necessitates a larger number (or perhaps more careful selection) of landmark documents for the embedding.

The bottom right panel of Figure 4 shows worse performance for both `Direct-NCE` and `Landmark-NCE` when compared with `word2vec` on the IMDB sentiment dataset. One possible explanation for this relative change in performance of the `NCE` methods across datasets is that the sentiment of a movie review may not be as related to the underlying topic structure of the document as the category of a news or Wikipedia article. Indeed, `LDA` also experiences a similar relative drop in performance when moving from the AG news and DBpedia ontology datasets to the IMDB movie review dataset.

7.4.1 Number of Landmarks

Figure 5 shows the effects of varying the number of landmarks in the `Landmark-NCE` embedding. Across all datasets, a larger number of landmarks leads to better performance on the semi-supervised learning task, consistent with the theory established in Section 4 and Section 5.

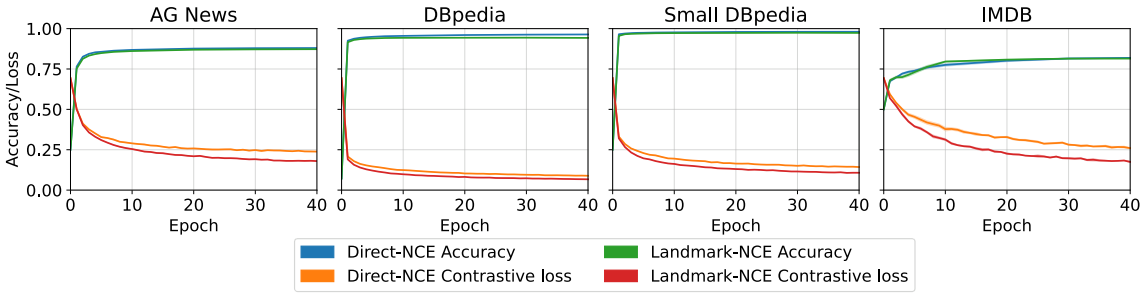

Figure 6: Contrastive loss on held-out validation set plotted against accuracy on downstream classification tasks. On all datasets, `Direct-NCE` used 100-dimensional embeddings and `Landmark-NCE` used 1k-dimensional embeddings. We used the full labeled training sets to train the linear classifier (4k points on AG News and Small DBpedia, 2k points on IMDB, and 14k points on DBpedia).

7.4.2 Accuracy v.s. Contrastive Loss

We also compared the evolution of the contrastive loss of a model and the downstream linear classifcation accuracy of its resulting embeddings. To do so, we tracked the contrastive losses of a model on the held-out validation contrastive dataset. Simultaneously, we checkpointed the models, trained a linear classifier on the embeddings produced by the model, and evaluated their supervised test accuracy. Figure 6 shows how contrastive loss and downstream classification accuracy evolve over training epochs. We see that on all datasets and for both `Direct-NCE` and `Landmark-NCE`, classification accuracy steadily improves as contrastive loss decreases, suggesting that in these settings, contrastive loss (which we can measure using an unlabeled validation set) can be a good surrogate for downstream performance (which may not be measurable until we have a task at hand).

We also observe that on all datasets, the `Landmark-NCE` models achieved lower contrastive loss than their `Direct-NCE` counterparts. This is perhaps due to the difficulty of optimizing models of the form given in Eq. (6). Nevertheless, despite their inferior performance on the contrastive learning tasking, the direct embedding models generally produced embeddings with higher downstream classification accuracy, suggesting that the landmark approach (at least with randomly chosen landmarks) is somewhat inefficient in utilizing the knowledge extracted from the contrastive learning task.

7.4.3 Visualizing Embeddings

For a qualitative perspective, we visualize the embeddings from both `NCE` methods using t-SNE with the default scikit-learn parameters (van der Maaten and Hinton, 2008; Pedregosa et al., 2011). To compare, we also used t-SNE to visualize the document-averaged `word2vec` embeddings. Figure 7 shows these visualizations on the 7,600 test documents of the AG News dataset colored according to their true label. While qualitative, the visualizations of the `Direct-NCE` and `Landmark-NCE` embeddings appear to be more clearly separated into label-homogeneous regions than that of `word2vec`.

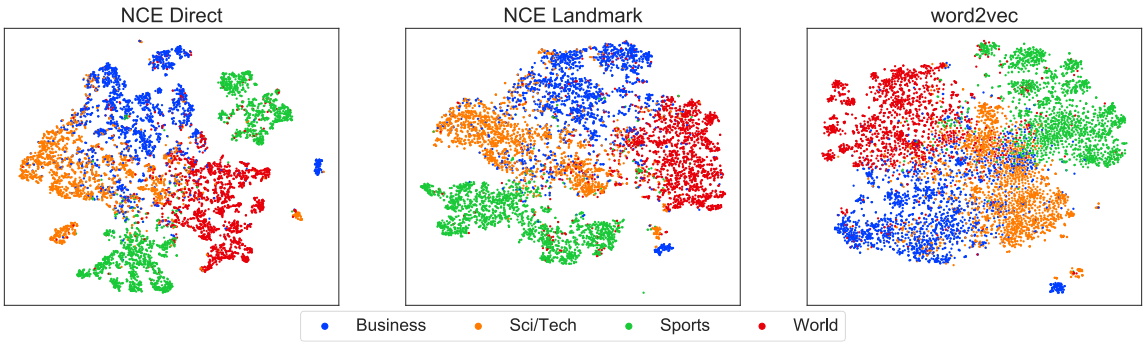

Figure 7: t-SNE visualizations of different embedding methods on the AG news dataset. From left to right: `Direct-NCE`, `Landmark-NCE`, and `word2vec`. All embedding methods used dimension 1k.

8. Proofs

8.1 Proofs from Section 4

Proof [Proof of Lemma 1] Conditioned on the topic vector w, $x^{(1)}$ and $x^{(2)}$ are independent. Thus,

$$\begin{aligned} g^\star(x, x') &= \frac{\mathbb{P}(x^{(1)} = x, x^{(2)} = x')}{\mathbb{P}(x^{(1)} = x)\mathbb{P}(x^{(2)} = x')} \\ &= \sum_{k=1}^{K} \frac{\mathbb{P}(w = e_k)\mathbb{P}(x^{(1)} = x \mid w = e_k)\mathbb{P}(x^{(2)} = x' \mid w = e_k)}{\mathbb{P}(x^{(1)} = x)\mathbb{P}(x^{(2)} = x')} \\ &= \sum_{k=1}^{K} \frac{\mathbb{P}(w = e_k \mid x^{(1)} = x)\mathbb{P}(x^{(2)} = x' \mid w = e_k)}{\mathbb{P}(x^{(2)} = x')} \\ &= \eta(x)^\intercal \psi(x') \end{aligned}$$

where the third equality follows from Bayes' rule. ■

Proof [Proof of Lemma 2] Fix a document x of length m and a document probability vector w. Recall the generative process of a such document spelled out in Section 2: each of the m words $x_1, \ldots, x_m$ are sampled independently by first drawing the latent variable z_i according to $\mathbb{P}(z_i = k) = w_k$, and then by drawing $x_i \sim O(\cdot \mid z_i)$.

Thus, by marginalizing over and conditioning on the z_i's, the probability of a document factorizes as

$$\mathbb{P}(x \mid w) = \sum_{z \in [K]^m} \prod_{i=1}^{m} w_{z_i} O(x_i \mid z_i) = \sum_{z \in [K]^m} \left(\prod_{i=1}^{m} w_{z_i} \right) \left(\prod_{i=1}^{m} O(x_i \mid z_i) \right) = \pi_m(w)^\intercal \psi_m(x),$$

where $\pi_m(w) := (w_1^{\alpha_1} \cdots w_k^{\alpha_k} : \alpha \in S_m^K)$ consists of the degree-m monomials. Thus,

$$\begin{aligned} g^\star(x, x') &= \frac{\mathbb{P}(x^{(1)} = x, x^{(2)} = x')}{\mathbb{P}(x^{(1)} = x)\mathbb{P}(x^{(2)} = x')} \\ &= \frac{\int_w \mathbb{P}(x^{(1)} = x \mid w)\mathbb{P}(x^{(2)} = x' \mid w) \, \mathrm{d}\mathbb{P}(w)}{\mathbb{P}(x^{(1)} = x)\mathbb{P}(x^{(2)} = x')} \\ &= \frac{\int_w \mathbb{P}(x^{(2)} = x' \mid w) \, \mathrm{d}\mathbb{P}(w \mid x^{(1)} = x)}{\mathbb{P}(x^{(2)} = x')} \\ &= \frac{\int_w \pi_m(w)^\mathsf{T} \psi_m(x) \, \mathrm{d}\mathbb{P}(w \mid x^{(1)} = x)}{\mathbb{P}(x^{(2)} = x')} \\ &= \eta(x)^\mathsf{T} \psi(x'). \end{aligned}$$

■

Proof [Proof of Theorem 3| By assumption (i), there exists an anchor word a_k for each topic $k = 1, \ldots, K$. By definition this means that $O(a_k \mid j) > 0$ if and only if $j = k$. For each vector $\alpha \in \mathbb{Z}_+^K$ such that $\sum \alpha_k \leq d_o$, create a landmark document consisting of α_k copies of a_k for $k = 1, \ldots, K$. This will result in $\binom{K+d_o}{d_o}$ landmark documents. Moreover, from assumption (ii), we can see that each of these landmark documents has positive probability of occurring under the marginal distribution of $x^{(2)}$ for $(x^{(1)}, x^{(2)}, y) \sim \mathcal{D}_{\text{contrast}}$, which implies $g^\star(x, l)$ is well-defined for all our landmark documents l.

Let l denote one of our landmark documents and let $\alpha \in \mathbb{Z}_+^K$ be its associated vector. Since l only contains anchor words, $\psi(l)_\beta > 0$ if and only if $\alpha = \beta$. To see this, note that

$$\psi(l)_\alpha = \sum_{z \in \binom{[m]}{\alpha}} \prod_{i=1}^m O(l_i \mid z_i) \;\geq\; \prod_{k=1}^K O(a_k \mid k)^{\alpha_k} > 0.$$

On the other hand, if $\beta \neq \alpha$ but $\sum_k \beta_k = \sum_k \alpha_k$, then there exists an index k such that $\beta_k \geq \alpha_k + 1$. Thus, for any $z \in \binom{[m]}{\beta}$, there will be more than α_k words in l assigned to topic k. Since every word in l is an anchor word and at most α_k of them correspond to topic k, we will have

$$\prod_{i=1}^m O(l_i \mid z_i) \;=\; 0.$$

Rebinding $\psi(l) = (\psi_0(l), \ldots, \psi_{d_0}(l))$ and forming the matrix L using this definition, we see that L^T can be diagonalized and inverted.

For any target degree-d_o polynomial $Q(w)$, there exists a vector v such that $Q(w) = \langle v, \pi_{d_0}(w) \rangle$, where $\pi_{d_0}(w)$ denotes the degree-d_0 monomial vector. Thus, we may take $\theta = L^{-1}v$ and get that for any document x:

$$\langle \theta, g^\star(x, l_{1:M}) \rangle \;=\; (L^{-1}v)^T L^\mathsf{T} \eta(x) \;=\; \mathbb{E}[\langle v, \pi_{d_0}(w) \rangle \mid x^{(1)} = x] \;=\; \mathbb{E}[Q(w) \mid x^{(1)} = x].$$ ■

8.2 Proofs from Section 5

8.2.1 Proof of error transformation guarantee

We first recall and setup some notation. For $(x^{(1)}, x^{(2)}, y) \sim \mathcal{D}_{\text{contrast}}$ (our contrastive distribution defined in Section 3), we let μ_i denote the marginal distribution of $x^{(i)}$. Furthermore, recall the contrastive loss, conditional probability, odds ratio, and oracle representation functions:

$$\begin{aligned}
L_{\text{contrast}}(f) &:= \mathbb{E}_{(x,x',y)\sim\mathcal{D}_{\text{contrast}}}\left[y \log\left(1 + e^{-f(x,x')}\right) + (1-y)\log\left(1 + e^{f(x,x')}\right)\right] \\
f^\star(x,x') &:= \log \frac{\mathbb{P}(y=1 \mid x^{(1)} = x, x^{(2)} = x')}{\mathbb{P}(y=0 \mid x^{(1)} = x, x^{(2)} = x')}, \\
g^\star(x,x') &:= \exp\left(f^\star(x,x')\right) = \frac{\mathbb{P}(x^{(1)} = x, x^{(2)} = x')}{\mathbb{P}(x^{(1)} = x)\mathbb{P}(x^{(2)} = x')}, \\
\phi^\star(x) &:= (g^\star(x, l_1), \ldots, g^\star(x, l_M))
\end{aligned}$$

where $l_1, \ldots, l_M$ are landmark documents. The learned approximation to $f^\star$ is $\hat{f}$, and from it we derive

$$\begin{aligned}
\hat{g}(x,x') &:= \exp\left(\hat{f}(x,x')\right), \\
\hat{\phi}(x) &:= (\hat{g}(x, l_1), \ldots, \hat{g}(x, l_M))
\end{aligned}$$

Let $\eta(x), \psi(x)$ denote the posterior/likelihood vectors from Lemma 1 or the posterior/likelihood polynomial vectors from Lemma 2. Say the length of this vector is $N \geq 1$.

Our goal is to show that linear functions in the representation $\hat{\phi}(x)$ can provide a good approximation to the target function

$$x \mapsto \eta(x)^\intercal \theta^\star$$

where $\theta^\star \in \mathbb{R}^N$ is some fixed vector. To this end, define

$$R(\hat{\phi}) := \min_v \mathbb{E}_{x\sim\mu^{(1)}}(\eta(x)^\intercal\theta^\star - \hat{\phi}(x)^\intercal v)^2,$$

which is the best mean squared error achievable using the representation $\hat{\phi}$.

By Lemma 1 or Lemma 2, we know that for any x, x' we have

$$g^\star(x,x') = \eta(x)^\intercal \psi(x').$$

Recall the matrix

$$L := \left(\psi(l_1), \ldots, \psi(l_M)\right).$$

This matrix is in $\mathbb{R}^{N\times M}$. If L has full row rank, then

$$\eta(x)^\intercal\theta^\star = \eta(x)^\intercal L L^\dagger \theta^\star = \phi^\star(x)^\intercal v^\star$$

where

$$\phi^\star(x) := (g^\star(x, l_1), \ldots, g^\star(x, l_M))$$

and $v^\star = L^\dagger\theta^\star$. Thus, $R(\phi^\star) = 0$. We will show that $R(\hat{\phi})$ can be bounded as well.

Theorem 5 (Restatement of Theorem 4) *Suppose the following assumptions hold.*

(1) There is a constant $\sigma_{\min} > 0$ such that for any $\delta \in (0,1)$, there is a number $M_0(\delta)$ such that for an i.i.d. sample $l_1, \ldots, l_M$ with $M \geq M_0(\delta)$, with probability $1-\delta$, the matrix

$$L = \begin{bmatrix} \psi(l_1) & \cdots & \psi(l_M) \end{bmatrix}$$

has minimum singular value at least $\sigma_{\min}\sqrt{M}$.

(2) There exists a value $g_{\max} > 0$ such that for all documents x and landmarks l_i

$$\max\{\hat{f}(x, l_i), f^\star(x, l_i)\} \leq \log g_{\max}.$$

Let $\hat{f}$ be the function returned by the contrastive learning algorithm, and let

$$\varepsilon := L_{\text{contrast}}(\hat{f}) - L_{\text{contrast}}(f^\star)$$

denote its excess contrastive loss. For any $\delta \in (0,1)$, if $M \geq M_0(\delta/2)$, then with probability at least $1-\delta$ over the random draw of $l_1, \ldots, l_M$, we have

$$R(\hat{\phi}) \leq \frac{\|\theta^\star\|_2^2 (1+g_{\max})^4}{\sigma_{\min}^2} \left(\varepsilon + \sqrt{\frac{2\log(2/\delta)}{M}} \right).$$

Proof We first condition on two events based on the sample $l_1, \ldots, l_M$. The first is the event that L has full row rank and smallest non-zero singular value at least $\sqrt{M}\sigma_{\min} > 0$; this event has probability at least $1 - \delta/2$. The second is the event that

$$\frac{1}{M}\sum_{j=1}^{M} \mathbb{E}_{x\sim\mu^{(1)}} \left(p^\star(x,x') - \hat{p}(x,x')\right)^2 \leq \mathbb{E}_{(x,x')\sim\mu^{(1)}\otimes\mu^{(2)}} \left(p^\star(x,l_j) - \hat{p}(x,l_j)\right)^2 + \sqrt{\frac{2\log(2/\delta)}{M}} \tag{7}$$

where we make the definitions

$$\begin{aligned} \hat{p}(x,x') &:= \frac{1}{1+e^{-\hat{f}(x,x')}} = \frac{\hat{g}(x,x')}{1+\hat{g}(x,x')} \\ p^\star(x,x') &:= \frac{1}{1+e^{-f^\star(x,x')}} = \frac{g^\star(x,x')}{1+g^\star(x,x')}. \end{aligned}$$

By Hoeffding's inequality and the fact that $\hat{p}$ and $p^\star$ have range $[0,1]$, this event also has probability at least $1-\delta/2$. By the union bound, both events hold simultaneously with probability at least $1-\delta$. We condition on these two events for the remainder of the proof.

Since L has full row rank, via Cauchy-Schwarz, we have

$$\begin{aligned} R(\hat{\phi}) &= \min_v \mathbb{E}_{x\sim\mu^{(1)}} (\eta(x)^\intercal\theta^\star - \hat{\phi}(x)^\intercal v)^2 \\ &\leq \mathbb{E}_{x\sim\mu^{(1)}} (\eta(x)^\intercal\theta^\star - \hat{\phi}(x)^\intercal v^\star)^2 \\ &= \mathbb{E}_{x\sim\mu^{(1)}} ((\phi^\star(x)^\intercal - \hat{\phi}(x))^\intercal v^\star)^2 \\ &\leq \mathbb{E}_{x\sim\mu^{(1)}} \left\|v^\star\right\|_2^2 \left\|\phi^\star(x)^\intercal - \hat{\phi}(x)\right\|_2^2 \\ &= \left\|v^\star\right\|_2^2 \cdot \mathbb{E}_{x\sim\mu^{(1)}} \left\|\phi^\star(x)^\intercal - \hat{\phi}(x)\right\|_2^2. \end{aligned}$$

We analyze the two factors on the right-hand side separately.

Analysis of $v^\star$. For $v^\star$, we have

$$\left\|v^\star\right\|_2^2 \le \left\|L^\dagger\right\|_2^2 \left\|\theta^\star\right\|_2^2 \le \frac{1}{M\sigma_{\min}^2} \left\|\theta^\star\right\|_2^2,$$

where we have used the fact that L has smallest non-zero singular value at least $\sqrt{M}\sigma_{\min}$.

Analysis of $\phi^\star - \hat{\phi}$. For the other term, we first note that

$$p^\star(x, x') \;=\; 1/(1 + e^{-f^\star(x,x')}) \;=\; \mathbb{P}(y = 1 \mid x^{(1)} = x, x^{(2)} = x').$$

Thus, we have

$$\begin{aligned}
\varepsilon \;&=\; L_{\text{contrast}}(\hat{f}) - L_{\text{contrast}}(f^\star) \\
&=\; \mathbb{E}_{(x,x',y)\sim\mathcal{D}_{\text{contrast}}} \left[y \log\left(\frac{p^\star(x,x')}{\hat{p}(x,x')}\right) + (1-y)\log\left(\frac{1-p^\star(x,x')}{1-\hat{p}(x,x')}\right)\right] \\
&=\; \mathbb{E}_{(x,x')\sim\mathcal{D}_{\text{contrast}}} \left[p^\star(x,x') \log\left(\frac{p^\star(x,x')}{\hat{p}(x,x')}\right) + (1-p^\star(x,x'))\log\left(\frac{1-p^\star(x,x')}{1-\hat{p}(x,x')}\right)\right] \\
&=\; \mathbb{E}_{(x,x')\sim\mathcal{D}_{\text{contrast}}} \left[\mathrm{KL}(p^\star(x,x'), p(x,x'))\right]
\end{aligned}$$

where $\mathrm{KL}(p, p')$ denotes the KL-divergence between two Bernoulli distributions with biases p and p', respectively. Pinkser's inequality tells us that $\mathrm{KL}(p, p') \ge 2(p - p')^2$. Combining this with the fact that $\mathcal{D}_{\text{contrast}}$ is a mixture distribution that places half its probability mass in $\mu^{(1)} \otimes \mu^{(2)}$ implies

$$\varepsilon \;\ge\; 2\mathbb{E}_{(x,x')\sim\mathcal{D}_{\text{contrast}}} \left[\left(\hat{p}(x,x') - p^\star(x,x')\right)^2\right] \ge \mathbb{E}_{(x,x')\sim\mu^{(1)}\otimes\mu^{(2)}} \left[\left(\hat{p}(x,x') - p^\star(x,x')\right)^2\right].$$

Combining the above with Eq. (7) and the definitions of $\hat{p}, p^\star$, we have

$$\begin{aligned}
&\mathbb{E}_{x\sim\mu^{(1)}} \left\|\phi^\star(x) - \hat{\phi}(x)\right\|_2^2 \\
&=\; \sum_{j=1}^{M} \mathbb{E}_{x\sim\mu^{(1)}} (g^\star(x, l_j) - \hat{g}(x, l_j))^2 \\
&\le\; (1 + g_{\max})^4 \sum_{j=1}^{M} \mathbb{E}_{x\sim\mu^{(1)}} (p^\star(x, l_j) - \hat{p}(x, l_j))^2 \\
&\le\; M(1 + g_{\max})^4 \left(\mathbb{E}_{(x,x')\sim\mu^{(1)}\otimes\mu^{(2)}} \left(p^\star(x, l_j) - \hat{p}(x, l_j)\right)^2 + \sqrt{\frac{2\log(2/\delta)}{M}}\right) \\
&\le\; M(1 + g_{\max})^4 \left(\varepsilon + \sqrt{\frac{2\log(2/\delta)}{M}}\right).
\end{aligned}$$

Wrapping up. To conclude, we have

$$\begin{aligned} R(\hat{\phi}) &\leq \left\|v^\star\right\|_2^2 \cdot \mathbb{E}_{x\sim\mu^{(1)}} \left\|\phi^\star(x)^\mathsf{T} - \hat{\phi}(x)\right\|_2^2 \\ &\leq \left(\frac{1}{M\sigma_{\min}^2}\left\|\theta^\star\right\|_2^2\right)\left(M(1+g_{\max})^4\left(\varepsilon + \sqrt{\frac{2\log(2/\delta)}{M}}\right)\right) \\ &= \frac{\|\theta^\star\|_2^2 (1+g_{\max})^4}{\sigma_{\min}^2}\left(\varepsilon + \sqrt{\frac{2\log(2/\delta)}{M}}\right). \end{aligned}$$

■

8.2.2 Satisfying Assumption 1

Suppose we are in the single topic case where $w \in \{e_1, \ldots, e_K\}$. Assume that $\min_k \Pr(w = e_k) \geq w_{\min}$. Further assumes that each topic k has an anchor word a_k, satisfying $O(a_k \mid z = e_k) \geq a_{\min}$. Then we will show that when M and m are large enough, the matrix L whose columns are $\psi(x)/\mathbb{P}(x)$ will have large singular values.

First note that if document x contains a_k then $\psi(x)$ is one sparse, and satisfies

$$\psi(x) = \frac{\mathbb{P}(x \mid w = e_k)}{\sum_{k'} \mathbb{P}(w = k')\mathbb{P}(x \mid w = k')} e_k = \frac{1}{\mathbb{P}(w = k)} e_k.$$

Therefore, the second moment matrix satisfies

$$\begin{aligned} \mathbb{E}\left[\psi(x)\psi(x)^\mathsf{T}\right] &\succeq \sum_{k=1}^K \mathbb{P}(w = e_k)\mathbb{P}(a_k \in x \mid e_k)\mathbb{E}\left[\psi(x)\psi(x)^\mathsf{T} \mid a_k \in x, w = e_k\right] \\ &= \sum_{k=1}^K \frac{\mathbb{P}(a_k \in x \mid e_k)}{\mathbb{P}(w = e_k)} e_k e_k^\mathsf{T}. \end{aligned}$$

Now, if the number of words per document is $m \geq 1/a_{\min}$ then

$$\begin{aligned} \mathbb{P}(a_k \in x \mid e_k) &= 1 - (1 - O(a_k \mid e_k))^m \geq 1 - \exp(-mO(a_k \mid e_k)) \\ &\geq 1 - \exp(-ma_{\min}) \\ &\geq 1 - 1/e. \end{aligned}$$

Finally, using the fact that $\mathbb{P}(w = e_k) \leq 1$, we see that the second moment matrix satisfies

$$\mathbb{E}\left[\psi(x)\psi(x)^\mathsf{T}\right] \succeq (1 - 1/e) I_{K\times K}.$$

For the empirical matrix, we perform a crude analysis and apply the Matrix-Hoeffding inequality (Tropp, 2012). We have $\left\|\psi(x)\psi(x)^\mathsf{T}\right\|_2 \leq K w_{\min}^{-2}$ and so with probability at least $1 - \delta$, we have

$$\left\|\frac{1}{M}\sum_{i=1}^M \psi(l_i)\psi(l_i)^\mathsf{T} - \mathbb{E}\left[\psi(x)\psi(x)^\mathsf{T}\right]\right\|_2 \leq \sqrt{\frac{8K\log(K/\delta)}{M w_{\min}^2}}.$$

If we take $M \geq \Omega(K\log(K/\delta)/w_{\min}^2)$ then we will have that the minimum eigenvalue of the empirical second moment matrix will be at least $1/2$.

9. Discussion

Our analysis shows that document-level contrastive learning under topic modeling assumptions yields a representation that exposes posterior topic information to linear predictors, and hence is suitable for certain downstream supervised learning tasks. We validated our theoretical results in a simulated topic modeling study, demonstrating that neural networks trained on the contrastive learning task can recover underlying topic posterior information. In semi-supervised learning experiments, we show that our contrastive learning procedure yields representations that improve classification accuracy over several natural baselines, and the improvement is most striking when we have few labeled examples.

There are a number of interesting future directions for this line of work.

- How robust or amenable to transfer learning are the representations studied in this work? For example, what guarantees can be made if we build a representation by training a contrastive model on one distribution but learn a classifier on documents from a different distribution? Such scenarios are common in document modeling settings where it is easy to scrape large corpora from websites such as Wikipedia but the classification problem of interest involves a smaller domain-specific dataset. It is natural to expect a deviation in distribution between the two datasets in this scenario, and one might reasonably wonder how contrastive learning behaves under such distributional shifts.

- Is there a better way to select landmarks? In this work, we have mostly focused on the setting where landmarks were randomly selected. However, this may not be optimal. Indeed, the results of Section 4 suggest that there are certain qualities that we might want our landmarks to have in order to guarantee some type of "coverage" of the underlying topics. This naturally opens up the question of how to go about finding good landmarks with which to embed future documents.

- In what other settings can contrastive learning be shown to extract useful underlying information? While we have focused on document representations and topic modeling assumptions in this work, our analysis more generally sheds light on the power of contrastive learning, which is empirically known to be useful in many settings. Aspects of our analysis may help characterize the expressiveness of contrastive learning representations under other modeling assumptions, for example in time-series data (see, e.g., Liu et al., 2021).

Acknowledgments

We thank Miro Dudík for initial discussions and suggesting the landmark embedding technique. This work was initiated while CT and DH were visiting Microsoft Research NYC, and was completed while CT was at Columbia University, and was supported in part by NSF grant CCF-1740833 and a JP Morgan Faculty Award.

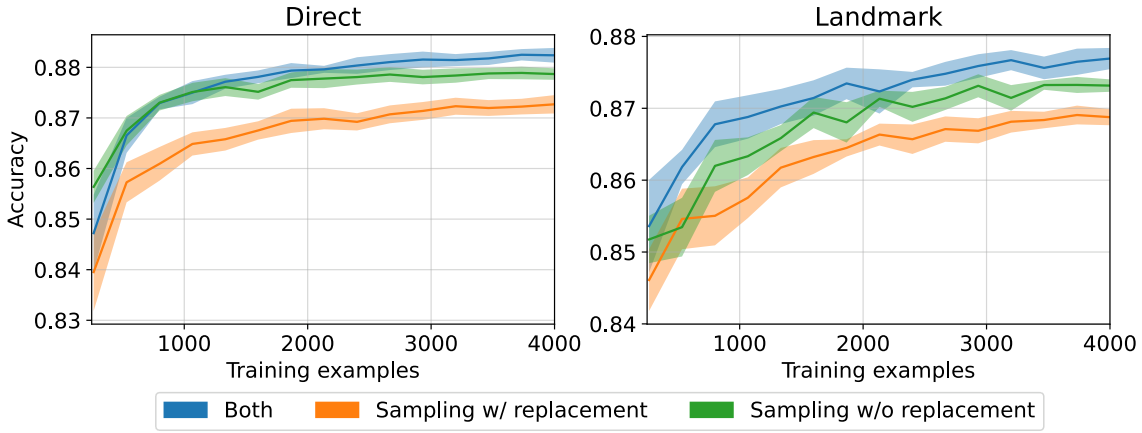

Figure 8: Effect of the choice of contrastive learning signal on downstream classification task for the AG News dataset. Both `Direct-NCE` and `Landmark-NCE` used 1k-dimensional embeddings.

Appendix A. Other Semi-supervised Experimental Results

A.1 Contrastive Learning Signal

On the AG News dataset, we also investigated the effect of contrastive learning signal on downstream accuracy. Recall that we considered two ways of creating document halves: randomly partitioning a document in half (which we call "sampling without replacement") and randomly sampling from the empirical distribution of words in a document (which we call "sampling with replacement"). We tested both of these methods as well as a third that alternates between these two method with equal probability (which we call "both"). Figure 8 shows a comparison of these methods, where we can see that sampling without replacement outperforms sampling with replacement, but each are generally outperformed by alternating between the two methods.

A.2 Effect of Dimension on `Direct-NCE`

Figure 9 shows the results of varying the dimension of the embedding used by `Direct-NCE` on the downstream classification accuracy. In contrast with Figure 5, Figure 9 does not show a strong effect of dimension on downstream classification accuracy.

References

N. Abe, B. Zadrozny, and J. Langford. Outlier detection by active learning. In *International Conference on Knowledge Discovery and Data Mining*, 2006.

A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y.-K. Liu. A spectral algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, 2012.

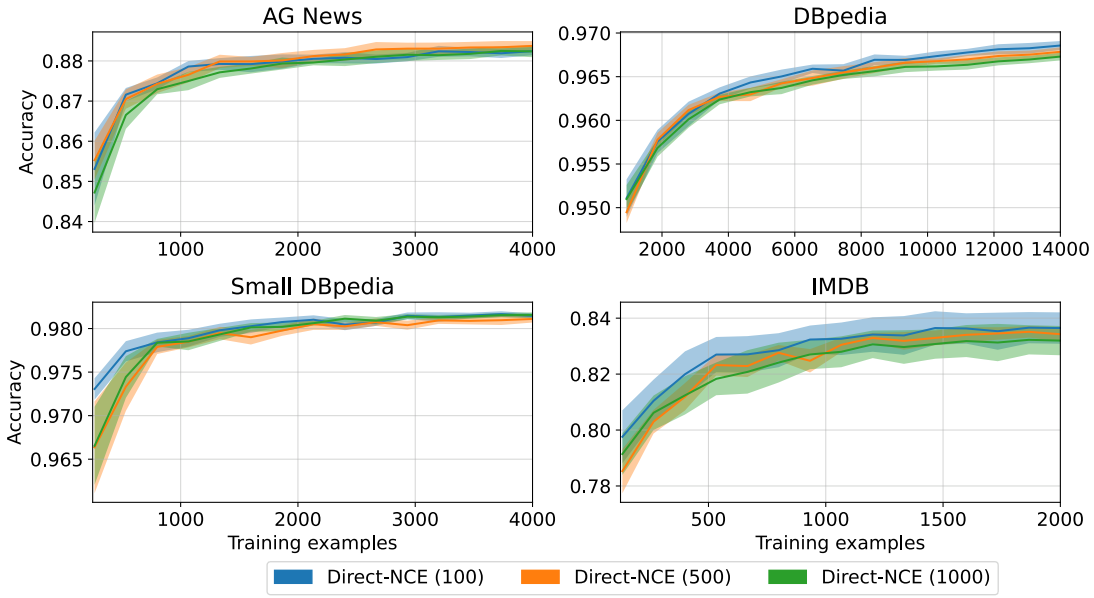

Figure 9: Semi-supervised performance of `Direct-NCE`. Dimension of the embedding is in parentheses. Top left: AG news dataset. Top right: DBpedia ontology dataset. Bottom left: Small DBpedia ontology dataset. Bottom right: IMDB sentiment dataset.

R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 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, R. Ge, and A. Moitra. Learning topic models–going beyond SVD. In *Symposium on Foundations of Computer Science*, 2012.

S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, 2013.

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.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November 3*, 2005.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

W. Ding, M. H. Rohban, P. Ishwar, and V. Saligrama. Topic discovery through data dependent and random projections. In *International Conference on Machine Learning*, 2013.

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. 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. *arXiv preprint arXiv:1808.06670*, 2018.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3):321–377, 1936.

S. M. Kakade and D. P. Foster. Multi-view regression via canonical correlation analysis. In *International Conference on Computational Learning Theory*, 2007.

N. S. Keskar and R. Socher. Improving generalization performance by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*, 2017.

A. Kulesza, N. R. Rao, and S. Singh. Low-rank spectral learning. In *International Conference on Artificial Intelligence and Statistics*, 2014.

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 preprint arXiv:2008.01064*, 2020.

B. Liu, P. Ravikumar, and A. Risteski. Contrastive learning of strong-mixing continuous-time stochastic processes. In *International Conference on Artificial Intelligence and Statistics*, 2021.

L. Logeswaran and H. Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018.

A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Annual meeting of the Association for Computational Linguistics: Human language technologies*, 2011.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013b.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: Transfer learning from unlabeled data. In *International Conference on Machine Learning*, 2007.

E. H. Shuford, A. Albert, and H. E. Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966.

D. Sontag and D. Roy. Complexity of inference in latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, 2011.

I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005.

M. H. Stone. The generalized Weierstrass approximation theorem. *Mathematics Magazine*, 1948.

C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive learning, multi-view redundancy, and linear models. In *International Conference on Algorithmic Learning Theory*, 2021.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 2012.

L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *International Conference on Computer Vision*, 2015.

G. U. Yule. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 226(636-646): 267–298, 1927.

X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 2015.