Masked Prediction: A Parameter Identifiability View

Bingbin Liu
Carnegie Mellon University
bingbinl@cs.cmu.edu

Daniel Hsu
Columbia University
djhsu@cs.columbia.edu

Pradeep Ravikumar
Carnegie Mellon University
pradeepr@cs.cmu.edu

Andrej Risteski
Carnegie Mellon University
aristesk@andrew.cmu.edu

Abstract

The vast majority of work in self-supervised learning have focused on assessing recovered features by a chosen set of downstream tasks. While there are several commonly used benchmark datasets, this lens of feature learning requires assumptions on the downstream tasks which are not inherent to the data distribution itself. In this paper, we present an alternative lens, one of parameter identifiability: assuming data comes from a parametric probabilistic model, we train a self-supervised learning predictor with a suitable parametric form, and ask whether the parameters of the optimal predictor can be used to extract the parameters of the ground truth generative model.

Specifically, we focus on latent-variable models capturing sequential structures, namely Hidden Markov Models with both discrete and conditionally Gaussian observations. We focus on masked prediction as the self-supervised learning task and study the optimal masked predictor. We show that parameter identifiability is governed by the task difficulty, which is determined by the choice of data model and the amount of tokens to predict. Technique-wise, we uncover close connections with the uniqueness of *tensor rank decompositions*, a widely used tool in studying identifiability through the lens of the method of moments.

1 Introduction

Self-supervised learning (SSL) is a relatively new approach to unsupervised learning, where the learning algorithm learns to predict auxiliary labels generated automatically from the data without human annotators. The hope is that with a properly designed prediction task, a successfully learned predictor would capture some knowledge about the underlying data. While SSL has been enjoying a rapid growth on the empirical front, theoretical understanding of why and when SSL works is still nascent. In no small part, this is because formalizing the desired guarantees seems challenging. For instance, the focus of SSL has largely been on learning *good features*, which in practice has been quantified by downstream performance on various benchmark datasets [Wang et al., 2018, 2019, Deng et al., 2009, Zhai et al., 2019, Tamkin et al., 2021]. To provide theoretical underpinning to this, one needs to make extra assumptions on the relationship between the self-supervised prediction task and the downstream tasks [Arora et al., 2019, Saunshi et al., 2020, HaoChen et al., 2021, Lee et al., 2021a, Wang et al., 2021, Wei et al., 2021, Wen and Li, 2021].

While associating SSL with downstream supervised tasks is a useful perspective and has led to several very interesting theoretical results, we take a step back and revisit a more general goal of SSL, which is to learn some informative functionals of the data distribution. Naturally, the key question here is

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

what functionals should be considered *informative*. While downstream performance is a notable valid choice, in this work, we choose an alternative criterion that is meaningful even without referencing any downstream tasks.

The alternative lens we are interested in is whether the functionals of the data distribution extracted by the SSL predictors can be simply stitched together to obtain the data distribution itself, given additional side-information about the family from which the data distribution is drawn. While this might seem like a tall order, masked prediction based SSL algorithms (which is essentially what pseudo-likelihood corresponds to) have classically been used for learning parametric graphical models such as Ising models [Ravikumar et al., 2010, Bresler, 2015, Vuffray et al., 2016]. But can this be done for broader classes of parametric models?

In this paper, we take a preliminary step towards this and ask the question of *parameter identifiability*: assuming the data comes from a ground truth parametric probabilistic model, can common self-supervised tasks uniquely identify the parameters of the ground truth model? More precisely, are the parameters of the model uniquely determined by the optimal predictor for the SSL task (Definition 1)? An appeal of this identifiability perspective is that when a SSL task is sufficient for parameter identifiability, the model parameters can then be recovered straightforwardly from the parameters from the optimal SSL predictor. Parameter identification also has the desirable property of being independent of any downstream task.

A priori, it is unclear whether we can achieve such model parameter identifiability via self-supervised tasks, since it requires recovering the full (parametric) generative model which is arguably more difficult than learning generic latent representations. This work provides a positive answer for broad classes of HMMs: we show that the commonly-used *masked prediction task* [Pathak et al., 2016, Devlin et al., 2018, He et al., 2021, Lee et al., 2021a], wherein a model is trained to predict a masked-out part of a sample given the rest of the sample, can identify the parameters of a HMM. As noted earlier, while such masked prediction for parameter learning has been applied in classical settings such as Ising models [Ravikumar et al., 2010, Bresler, 2015, Vuffray et al., 2016], the HMM setup in this work is more challenging due to the presence of latent variables. HMMs are also more suitable for modeling practical sequential data, and have been commonly adopted in theoretical analyses as a clean proxy for languages [Wei et al., 2021, Xie et al., 2021].

Concretely, the two HMM models we consider in this work are 1) the classic HMM with discrete latent and discrete observables, and 2) a HMM variant with discrete latents and continuous observables that are conditionally Gaussian given the latent, which we abbreviate as G-HMMs. We show that:

- Parameter identifiability is governed by the difficulty of the masked prediction task. The task difficulty is related to the amount of information provided by the combination of the model and the prediction task—where the difficulty can be increased by using a more complicated model, or by predicting more tokens. For instance, predicting the conditional mean of one token given another does not yield identifiability for a discrete HMM (Theorem 2), but does so when data comes from a G-HMM (Theorem 3). Moreover, the identifiability in the latter case quite strongly leverages structural properties of the posterior of the latent variables (Section 3.1).

- Tools for characterizing the uniqueness of tensor decompositions (e.g., Kruskal's Theorem [Kruskal, 1977, Allman et al., 2009]) can be leveraged to prove identifiability: For both HMM (Theorem 5) and G-HMM (Theorem 6), if we have predictors of the tensor product of tokens (e.g., $\mathbb{E}[x_2 \otimes x_3 | x_1]$), we can use the predictor output to construct a 3-tensor whose rank-1 components are uniquely determined and reveal the parameters of the model.

The rest of the paper is structured as follows. Section 2 provides relevant definitions, preliminaries and assumptions. Section 3 states the main results of this work. Main proofs, including the identifiability proof via tensor decomposition, are provided in Section 4, with the rest deferred to the appendix. We then discuss related works in Section 5. Finally, we emphasize that this work is a first-cut study on this lens of parameter recovery for analyzing SSL tasks based on masked prediction, and our encouraging results suggest interesting open directions in this thread of analyzing self-supervised learning via parameter recovery, which are discussed briefly in the conclusion.

2 Setup

This work focuses on two classes of latent-variable sequence models. The first are fully discrete hidden Markov models (HMMs), and the second are HMMs whose observables marginally follow a mixtures of Gaussians with identity covariance. We denote the observations and hidden states respectively by $\{x_t\}_{t\geq 1}$ and $\{h_t\}_{t\geq 1}$ for both classes. The hidden states $h_1 \to h_2 \to \cdots$ form a Markov chain, and conditional on h_t, the observable x_t is independent of all other variables. Throughout, we refer to $\{x_t\}_{t\geq 1}$ as tokens, following the nomenclature from language models.

2.1 Models

Discrete Hidden Markov Model We first describe the parameterization of the standard HMMs with discrete latents and observations. Let $\mathcal{X} := \{1, \dots, d\} = [d]$ denote the observation space, and let $\mathcal{H} := [k]$ be the state space.[1] The parameters of interest are the *transition matrix* $T \in \mathbb{R}^{k\times k}$ and the *emission matrix* $O \in \mathbb{R}^{d\times k}$, defined in the standard way as

$$P(h_{t+1} = i \mid h_t = j) = T_{ij}, \qquad P(x_t = i \mid h_t = j) = O_{ij}.$$

Conditionally-Gaussian HMM (G-HMM) We next describe the parameterization of *conditionally-Gaussian HMMs (G-HMMs)*. The state space $\mathcal{H} := [k]$ is the same as in the previous case, while the observation space is now continuous with $\mathcal{X} := \mathbb{R}^d$. The parameters of interest are $T \in \mathbb{R}^{k\times k}$, the transition matrix, and $\{\mu_i\}_{i\in[k]} \subset \mathbb{R}^d$, the means of the k identity-covariance Gaussians. Precisely,

$$P(h_{t+1} = i \mid h_t = j) = T_{ij}, \qquad P(x_t = x \mid h_t = i) = (2\pi)^{-\frac{d}{2}} \exp\big(-\|x - \mu_i\|^2/2\big).$$

We use $M := [\mu_1, \dots, \mu_k] \in \mathbb{R}^{d\times k}$ to denote the matrix whose columns are the Gaussian means.

2.2 Masked prediction tasks

We are interested in the (regression) task of predicting one or more "masked out" tokens as a function of another observed token, with the goal of minimizing expected squared loss under a distribution given by an HMM or G-HMM (equation 1). In the case of the discrete HMMs, we will specifically be predicting the *one-hot encoding vectors* of the observations. Thus, both for HMM and G-HMM, predicting a single token will correspond to predicting a vector. For notational convenience, we will simply associate the discrete states or observations via their one-hot vectors $\{e_1, e_2, \dots\}$ in the appropriate space and interchangeably write $h = i$ or $h = e_i$, and similarly for x. For the task of predicting the tensor product of (one-hot encoding vectors of) tokens $\otimes_{\tau\in\mathcal{T}} x_\tau$ from another token x_t (where $\mathcal{T}$ is some index set and $t \notin \mathcal{T}$), the optimal predictor with respect to the squared loss calculates the conditional expectation:

$$f(x_t) = \arg\min_{\tilde{f}} \mathbb{E}_{\{x_\tau\}_{\tau\in\mathcal{T}}} \|\mathrm{vec}(\otimes_{\tau\in\mathcal{T}} x_\tau) - \mathrm{vec}(\tilde{f}(x_t))\|_2^2 = \mathbb{E}[\otimes_{\tau\in\mathcal{T}} x_\tau \mid x_t] \in (\mathbb{R}^d)^{\otimes|\mathcal{T}|}, \quad (1)$$

where "vec" returns the vectorized form of a tensor.

We use the shorthand "$\otimes_{\tau\in\mathcal{T}} x_\tau | x_t$" to refer to this prediction task. For instance, consider the case of predicting x_2 given x_1 under the HMM with parameters (O, T). The optimal predictor, denoted by $f^{2|1}$, can be written in terms of (O, T) as [2]

$$f^{2|1}(x) = \mathbb{E}[x_2 \mid x_1 = x] = \sum_{i\in[k]} \mathbb{E}[x_2 \mid h_2 = i] P(h_2 = i \mid x_1 = x)$$

$$= \sum_{i\in[k]} \sum_{j\in[k]} \mathbb{E}[x_2 \mid h_2 = i] P(h_2 = i \mid h_1 = j) \underbrace{P(h_1 = j \mid x)}_{:=[\phi(x)]_j} = \sum_{i\in[k]} \sum_{j\in[k]} O_i T_{ij} \underbrace{\frac{O_{x,j}}{\sum_{l\in[k]} O_{x,l}}}_{:=[\phi(x)]_j}.$$

[1] Our results will assume $d \geq k$; see Section 2.3.

[2] The computation here relies on Assumption 1, given in Section 2.3.

Here $\phi : \mathbb{R}^d \to \mathbb{R}^k$ denotes the posterior distribution of a hidden state h_t given the corresponding observation x_t, i.e., $\phi(x_t) = \mathbb{E}[h_t \mid x_t]$. [3]

Our goal is to study the parameter identifiability from the prediction tasks, when the predictors have the correct parametric form. Formally, we define identifiability from a prediction task as follows:

Definition 1 (Identifiability from a prediction task, HMM)**.** *A prediction task suffices for identifiability if, for any two HMMs with parameters (O, T) and $(\tilde{O}, \tilde{T})$, equality of their optimal predictors for this task implies that there is a permutation matrix Π such that $O = \tilde{O}\Pi$ and $T = \Pi^\top \tilde{T}\Pi$.*

In other words, the mapping from (the natural equivalence classes of) HMM distributions to optimal predictors for a task is injective, up to a permutation of the hidden state labels. By identifiability from a collection of prediction tasks, we refer to the injectiveness of the mapping from HMM distributions to the collections of optimal predictors for the tasks. Identifiability for G-HMMs is defined analogously with $O, \tilde{O}$ changed to $M, \tilde{M}$.

2.3 Assumptions

We now state the assumptions used in our results. The first assumption is that the transition matrices of the HMMs are doubly stochastic.

Assumption 1 (Doubly stochastic transitions)**.** *The transition matrix T is doubly stochastic, and the marginal distribution of the initial hidden state h_1 is stationary with respect to T.*

This assumption guarantees that the stationary distribution of the latent distribution is uniform for any t, and the transition matrix for the reversed chain is simply $T^\top$. Moreover, this assumption reduces the parameter space and hence will make the non-identifiability results stronger.

We require the following conditions on the parameters for the discrete HMM:

Assumption 2 (Non-redundancy, discrete HMM)**.** *Every row of O is non-zero.*

Assumption 2 can be interpreted as requiring each token to have a non-zero probability of being observed, which is a mild assumption. We also require the following non-degeneracy condition:

Assumption 3 (Non-degeneracy, discrete HMM)**.** *$rank(T) = rank(O) = k \leq d$.*

Note that Assumption 3 only requires the parameters to be non-degenerate, rather than have singular values bounded away from 0. The reason is that this work will focus on population level quantities and make no claims on finite sample behaviors or robustness.

For G-HMM, we similarly require the parameters to be non-degenerate:

Assumption 4 (Non-degeneracy, G-HMM)**.** *$rank(T) = rank(M) = k \leq d$.*

Moreover, we assume that the norms of the means are known and equal:

Assumption 5 (Equal norms of the means)**.** *For each $i \in [k]$, μ_i is a unit vector.*[4]

Assumptions 1-4 are fairly standard [see, e.g., Anandkumar et al., 2012]; in particular, Assumption 3, 4 are required to enable efficient learning, since learning degenerate HMMs can be computationally hard [Mossel and Roch, 2005]. Assumption 5 may be an artifact of our proofs, and it would be interesting to relax in future work.

Our notion of identifiability from a prediction task (or a collection of prediction tasks) will restrict attention to HMMs satisfying Assumptions 1, 2, 3 and G-HMMs satisfying Assumptions 1, 4, 5.

2.4 Uniqueness of tensor rank decompositions

Some of our identifiability results rely on the uniqueness of *tensor rank-1 decompositions* [Hitchcock, 1927]. An *order-t tensor* (or *t-tensor*) is an t-way multidimensional array; a matrix is a 2-tensor. The *tensor rank* of a tensor W is the minimum number R such that W can be written as a sum of R

[3] For discrete HMMs, $\phi(x_t) = \frac{O^\top x_t}{\|O^\top x_t\|_1}$. For GHMMs, $[\phi(x_t)]_i = \frac{\exp\left(-\frac{\|x_t - \mu_i\|_2^2}{2}\right)}{\sum_{j\in[k]} \exp\left(-\frac{\|x_t-\mu_j\|_2^2}{2}\right)}$, $\forall i \in [k]$. ϕ does not need to be indexed by t due to the stationarity assumption in Section 2.3.

[4] Assumption 5 can be changed to $\|\mu_i\|_2 = c$ for all $i \in [k]$, for any other fixed number $c > 0$.

rank-1 tensors. That is, if a t-tensor W has rank-R, it means that $W = \sum_{i\in[R]} \otimes_{j\in[t]} U_i^{(j)}$ for some matrices $U^{(j)} \in \mathbb{R}^{n_j \times R}$, where $U_i^{(j)}$ denotes the i^{th} column of matrix $U^{(j)}$.

In this work, we only need to work with 3-tensors of the form $W = \sum_{i\in[R]} A_i \otimes B_i \otimes C_i$ for some matrices $A \in \mathbb{R}^{n_1\times R}$, $B \in \mathbb{R}^{n_2\times R}$, $C \in \mathbb{R}^{n_3\times R}$, as 3-tensors will suffice for identifiability in all of our settings of interest.[5] A classic work by Kruskal [1977] gives a sufficient condition under which A, B, C can be recovered up to column-wise permutation and scaling. The condition is stated in terms of the *Kruskal rank*, which is the maximum number r such that every r columns of the matrix are linearly independent. Let k_A denote the Kruskal rank of matrix A, then:

Proposition 1 (Kruskal's theorem, Kruskal [1977])**.** *The components A, B, C of a 3-tensor $W := \sum_{i\in[R]} A_i \otimes B_i \otimes C_i$ are identifiable up to a shared column-wise permutation and column-wise scaling if $k_A + k_B + k_C \geq 2R + 2$.*

We note that this work focuses on identifiability results rather than providing an algorithm or sample complexity bounds, though the proofs can be adapted into algorithms [see, e.g., Harshman, 1970] under slightly more restrictive conditions (which will be satisfied by all of our identifiability results).

3 Identifiability from masked prediction tasks

We now present the main (non-)identifiability results, and show that the combination of the data generative models and the prediction task directly impacts the sufficiency of identifiability.

3.1 Pairwise prediction

We begin with the simplest prediction task: namely predicting one token from another. We refer to such tasks as *pairwise prediction tasks*. For HMMs, this task fails to provide parameter identifiability:

Theorem 2 (Nonidentifiability of HMM from predicting $x_t|x_1$)**.** *For any $t \in \mathbb{Z}, t \geq 2$, there exists a pair of HMM distributions with parameters (O, T) and $(\tilde{O}, \tilde{T})$, each satisfying Assumptions 1, 2 and 3, such that the optimal predictors for the task $x_t|x_1$ are the same under each distribution, but there is no permutation matrix $\Pi \in \mathbb{R}^{k\times k}$ such that $\tilde{O} = O\Pi$ and $\tilde{T} = \Pi^\top T \Pi$ are both satisfied.*

Theorem 2 follows from the fact that the optimal predictor has the form of a product of (stochastic) matrices, and generally, one cannot uniquely recover matrices from their product sans additional conditions [Donoho and Elad, 2003, Candes et al., 2006, Spielman et al., 2012, Arora et al., 2014, Georgiev et al., 2005, Aharon et al., 2006, Cohen and Gillis, 2019]. Specifically, by equation 1, the optimal predictor is $f(x_1) = \mathbb{E}[x_t|x_1] = OT^{t-1}\phi(x_1)$ (where $\phi(x_1) := \mathbb{E}[h_1|x_t]$ is the posterior). When $t = 2$, we can find a non-permutation matrix R such that $\tilde{O} = OR, \tilde{T} = R^\top T R$ give the same predictor as O, T. For $t > 2$, even if $\tilde{O} = O$, we show that the matrix power T^{t-1} is not identifiable:

Claim 1 (Nonidentifiability of matrix powers)**.** *For any $t \in \mathbb{Z}, t \geq 2$, there exist stochastic matrices $T, \tilde{T}$ satisfying Assumption 1, 3, such that $T \neq \tilde{T}$ and $T^t = \tilde{T}^t$.*

On the other hand, pairwise prediction actually *does* suffice for identifiability for G-HMM:

Theorem 3 (Identifiability of G-HMM from predicting $x_2|x_1$)**.** *Under Assumption 1, 4, and 5, if the optimal predictors for the task $x_2|x_1$ under the G-HMM distributions with parameters (M, T) and $(\tilde{M}, \tilde{T})$ are the same, then $(M, T) = (\tilde{M}, \tilde{T})$ up to a permutation of the hidden state labels.*

Comparing Theorem 2 and 3 shows that the specific parametric form of the generative model matters. Note that HMM and G-HMM have a similar form when conditioning on the latent variable; that is, with $t = 2$, the predictor conditioned on the hidden variable h_2 is $P(x_2|h_2 = i) = OT_i$ for HMM, and $P(x_2|h_2 = i) = MT_i$ for G-HMM. The salient difference between these two setups lies in the posterior function: while the posterior function for HMM is linear in the observable, the posterior function for G-HMM is more complicated and "reveals" more information about the parameter.

[5] To apply our results on higher order tensors, one can consider an order-3 slice of the higher order tensor.

To formalize the above intuition, first recall that the GHMM posterior has entries $[\phi(x_t)]_i = \frac{\exp\left(-\frac{\|x_t-\mu_i\|_2^2}{2}\right)}{\sum_{j\in[k]}\exp\left(-\frac{\|x_t-\mu_j\|_2^2}{2}\right)}$, $\forall i \in [k]$. We will show that for G-HMM, even matching the posterior function *nearly* suffices to identify M: if $M, \tilde{M}$ parameterize two posterior functions $\phi, \tilde{\phi}$ where $\phi = \tilde{\phi}$, then up to a permutation, $\tilde{M}$ must be equal to either M or a unique (and somewhat special) transformation of M. The next step is to further exclude the (special) transformation, which is achieved using the constraint that $T, \tilde{T}$ are stochastic matrices. The first step of the proof sketch is captured by following lemma:

Lemma 1. *For $d \geq k \geq 2$, under Assumption 4, 5, $\phi = \tilde{\phi}$ implies $\tilde{M} = M$ or $\tilde{M} = HM$, where H is a Householder transformation of the form $H := I_d - 2\hat{v}\hat{v}^\top \in \mathbb{R}^{d\times d}$, with $\hat{v} := \frac{(M^\dagger)^\top \mathbf{1}}{\sqrt{\mathbf{1}^\top M^\dagger (M^\dagger)^\top \mathbf{1}}}$.*

To provide some geometric intuition about how H acts on M, note that $\hat{v}$ is a unit vector in the column space of M and perpendicular to the affine hull of $\mathcal{A} := \{\mu_i : i \in [k]\}$, which means $\hat{v}^\top \mu_i$ is the same for all $i \in [k]$. As a result, $\tilde{M} = [\tilde{\mu}_1, ..., \tilde{\mu}_k] = [H\mu_1, ..., H\mu_k] = M - 2(\hat{v}^\top \mu_1)[\hat{v}, ..., \hat{v}]$ is a translation of M along the direction of $\hat{v}$, such that the translated points $\{\tilde{\mu}_i\}_{i\in[k]}$ lie on the opposite side of the origin. It is non-trivial to argue that HM is the only solution (other than M itself) that preserves ϕ, and we defer the proof to Appendix A.1.2. It is, however, easy to see that HM indeed results in a matching posterior, whose sufficient conditions are 1) $\tilde{M}$ is a translation of M, and 2) $\|\tilde{\mu}_i\|^2 - \|\mu_i\|^2$ is the same for all $i \in [k]$. $\tilde{M} := HM$ indeed satisfies both conditions.

Proof sketch for Theorem 3: We first show that if M, T and $\tilde{M}, \tilde{T}$ produce the same predictor, then their posterior function must be equal up to a permutation (Lemma 2). We can then apply Lemma 1 to recover M up to a permutation and a Householder transformation H. Then, we show that if $\tilde{M} = HM$, then the corresponding $\tilde{T}$ must have negative entries and thus would not be a valid stochastic matrix. Hence it must be that $\tilde{M}, M$ are equal up to permutation.

Finally, by way of remarks, another way to think of the difference between the two setups is that for HMM, $P(x_2|x_1)$ is a mixture of categorical distributions, which itself is also a categorical distribution. This also implies that the nonidentifiability from pairwise prediction in the HMM case cannot be resolved by changing the squared loss to another proper loss function. On the other hand, for G-HMM, the conditional distribution $P(x_2|x_1)$ is a mixture of Gaussians, which is well known to be identifiable. In fact, if we were given access to the *entire* conditional distribution $P(x_2|x_1)$ (instead of just the conditional mean), it is even easier to prove identifiability for G-HMM. Though this is already implied from identifiability from the conditional means, we provided a (much simpler) proof in Appendix A.3 assuming access to the full conditional distribution.

3.2 Beyond pairwise prediction

The conclusion from Theorem 2 is that a single pairwise prediction task does not suffice for identifiability on HMMs. The next question is then: can we modify the task to obtain identifiability? A natural idea is to force the model to "predict more", and one straightforward way to do so is to combine multiple pairwise prediction tasks. It turns out that this does not resolve the nonidentifiability issue, as we can show that the parameters are not identifiable even when considering *all* possible pairwise tasks involved 3 (adjacent) tokens:

Theorem 4 (Nonidentifiability of HMM from all pairwise predictions on 3 tokens)**.** *There exists a pair of HMM distributions with parameters (O, T) and $(\tilde{O}, \tilde{T})$, each satisfying Assumptions 1, 2 and 3, and also $\tilde{O} \neq O$, such that, for each of the tasks $x_2|x_1$, $x_1|x_2$, $x_3|x_1$, and $x_1|x_3$, the optimal predictors are the same under each distribution.*[6]

We briefly remark that the reason for only considering adjacent time steps is that when the tokens are at least two time steps apart, matching predictors only matches powers of the transition matrices, which in general does not ensure the transition matrices themselves are matched as shown in Claim 1.

[6] These 4 pairwise tasks cover all possible pairwise tasks on 3 adjacent tokens. In particular, there is no need to consider $x_2|x_3$ or $x_3|x_2$, since they are the same as $x_1|x_2$ and $x_2|x_1$.

For the intuition of the nonidentifiability result in Theorem 4, recall that the limitation of pairwise predictions on HMMs comes from non-uniqueness of matrix factorization. While adding additional pairwise prediction tasks introduces more equations on the product of matrices, these equations are highly dependent, and the proof works by providing counterexamples that can simultaneously satisfy all these equations.

The above intuition leads to another way of forcing the model to "predict more", that is, to increase the number of predicted tokens. The hope is that doing so results in equations on tensors—as opposed to matrices— for which there is a lot of classical machinery delineating tensors for which the rank-1 decomposition is unique, as discussed in Section 2.4. This intuition proves to be true and we show that increasing the number from 1 to 2 already suffices for identifiability:

Theorem 5 (Identifiability from masked prediction on three tokens, HMM)**.** *Let (t_1, t_2, t_3) be any permutation of $(1, 2, 3)$, and consider the prediction task $x_{t_2} \otimes x_{t_3}|x_{t_1}$. Under Assumption 1, 2, 3, if the optimal predictors under the HMM distributions with parameters (O, T) and $(\tilde{O}, \tilde{T})$ are the same, then $(O, T) = (\tilde{O}, \tilde{T})$ up to a permutation of the hidden state labels.*

Compared to prior results on identifiability from third order moments [Allman et al., 2009, Anandkumar et al., 2012, 2014], the difficulty in our setup is that we only have access to the conditional 2-tensors (i.e. matrices) given by the predictors. The proof idea is to construct a third-order tensor by linearly combining the conditional 2-tensors for each possible value of the token being conditioned on, such that Kruskal's theorem applies and gives identifiability. Note, importantly, that the weights for the linear combination cannot depend on the marginal probabilities of the token being conditioned on, since we do not have access to these marginals, and it is unclear whether we could extract unique marginals given the conditional probabilities we are predicting. Thus, the above theorem cannot be simply derived from results showing parameter identifiability from the 3rd order moments.

It can be show that this tensor decomposition argument can also be applied to G-HMM, with the help of Lemma 1. We leave the details to Theorem 6 in Appendix A.

4 Proofs

We now discuss proofs for some of the main results. Section 4.1 proves the identifiability of HMM parameters from the task of predicting two tokens (Theorem 5) using ideas from tensor decomposition, and Section 4.2 shows the identifiability proof of pairwise prediction on G-HMM. The rest of the proofs are deferred to the appendix.

4.1 Proof of Theorem 5: identifiability of predicting two tokens for HMM

There are three cases for the two-token prediction task, i.e. 1) $x_2 \otimes x_3|x_1$, 2) $x_1 \otimes x_3|x_2$, and 3) $x_1 \otimes x_2|x_3$. We will prove for the first two cases, as the third case is proved the same way as the first case by symmetry. In all cases, the idea is to use the predictor to construct a 3-tensor whose components are each of rank-k, so that applying Kruskal's theorem gives identifiability.

Case 1, $x_2 \otimes x_3|x_1$: O, T and $\tilde{O}, \tilde{T}$ producing the same predictor means $f^{2\otimes 3|1}(x_1) := \mathbb{E}[x_2 \otimes x_3|x_1] = \tilde{\mathbb{E}}[x_2 \otimes x_3|x_1] := \tilde{f}^{2\otimes 3|1}(x_1)$, where $\mathbb{E}, \tilde{\mathbb{E}}$ are parameterized by the corresponding parameters. Let $\mathcal{X} := \{e_i : i \in [d]\}$, and consider the following 3-tensor:

$$\begin{aligned} W &:= \sum_{x_1\in\mathcal{X}} x_1 \otimes \mathbb{E}[x_2 \otimes x_3|x_1] = \sum_{x_1\in\mathcal{X}} x_1 \otimes \mathbb{E}_{h_2|x_1}[\mathbb{E}[x_2 \otimes x_3|x_1]|h_2] \\ &= \sum_{i\in[k]} \sum_{x_1\in\mathcal{X}} P(h_2 = i|x_1) x_1 \otimes \mathbb{E}[x_2|h_2 = i] \otimes \mathbb{E}[x_3|h_2 = i] \\ &= \sum_{i\in[k]} \underbrace{\Big(\sum_{x_1\in\mathcal{X}} (T\phi(x_1))^\top e_i^{(k)} x_1 \Big)}_{:=a_i} \otimes O_i \otimes (OT)_i, \end{aligned} \tag{2}$$

where O_i denotes the i^{th} column of O, and similarly for $(OT)_i$. Note that W can also be written as

$$W = \sum_{x_1 \in \mathcal{X}} x_1 \otimes \tilde{\mathbb{E}}[x_2 \otimes x_3 | x_1] = \sum_{i \in [k]} \Big(\sum_{x_1 \in \mathcal{X}} (\tilde{T}\tilde{\phi}(x_1))^\top e_i^{(k)} x_1 \Big) \otimes \tilde{O}_i \otimes (\tilde{O}\tilde{T})_i. \tag{3}$$

We want to apply Kruskal's theorem for identifiability. In particular, we will show that each component in equation 2 forms a matrix of Kruskal rank k. The second and third components clearly satisfy this condition by Assumption 3. For the first component, recall that $\phi(x) = \frac{O^\top x}{\|O^\top x\|_1}$ and write a_i as

$$a_i = \sum_{j \in [d]} \big(T\phi(e_j^{(d)})\big)^\top e_i^{(k)} \cdot e_j^{(d)} = \text{diag}\left([\frac{1}{\|(e_j^{(d)})^\top O\|_1}]_{j \in [d]} \right) OT^\top e_i^{(k)}. \tag{4}$$

Putting a_i into a matrix form, we get $A := [a_1, ..., a_k] = \text{diag}\big([1/\|(e_j^{(d)})^\top O\|_1]_{j \in [d]}\big) OT^\top$,[7] which is of rank k by Assumption 3. Hence components W are all of Kruskal rank k, and columns of OT, O are identified up to column-wise permutation and scaling by Kruskal's theorem. The indeterminacy in scaling is further removed noting that columns of O, T need to sum up to 1. Lastly, T is recovered as $T = O^\dagger OT$.

Case 2, $x_1 \otimes x_3 | x_2$: The optimal predictor for the task of predicting x_1, x_3 given x_2 takes the form

$$\mathbb{E}[x_1 \otimes x_3 | x_2] = (OT^\top)\text{diag}(\phi(x_2))(OT)^\top. \tag{5}$$

Similarly as the previous case, we would like to construct a 3-tensor whose components can be uniquely determined by Kruskal's theorem. Let $\mathcal{X}$ be the same as before, and consider the 3-tensor

$$\begin{aligned} W &:= \sum_{x_2 \in \mathcal{X}} x_2 \otimes \mathbb{E}[x_1 \otimes x_3 | x_2] = \sum_{x_2 \in \mathcal{X}} x_2 \otimes \mathbb{E}_{h_2|x_2}(\mathbb{E}[x_1|h_2] \otimes \mathbb{E}[x_3|h_2]) \\ &= \sum_{i \in [k]} \underbrace{\sum_{x_2 \in \mathcal{X}} (\phi(x_2))^\top e_i^{(k)} x_2}_{:=a_i} \otimes \mathbb{E}[x_1|h_2] \otimes \mathbb{E}[x_3|h_2] = \sum_{i \in [k]} a_i \otimes (OT^\top)_i \otimes (OT)_i, \end{aligned} \tag{6}$$

where the first component can be simplified to

$$a_i = \sum_{j \in [d]} \frac{(e_j^{(d)})^\top O}{\|(e_j^{(d)})^\top O\|_1} e_i^{(k)} \cdot e_j^{(d)} = \Big(\text{diag}\big([\|O_j^\top\|_1]_{j \in [d]}\big)\Big)^{-1} O e_i^{(k)} := D^{-1} O e_i^{(k)}. \tag{7}$$

The matrix $A := [a_1, ..., a_k] = D^{-1}O$ is of rank k, hence we can identify (up to permutation) columns of each component of W by Kruskal's theorem. This means if O, T and $\tilde{O}, \tilde{T}$ produce the same predictor, then we have $OT = \tilde{O}\tilde{T}$, $OT^\top = \tilde{O}\tilde{T}^\top$, and that $O, \tilde{O}$ are matched up to a scaling of rows (i.e. D^{-1}). Next, to determine D, note that $T, \tilde{T}$ are doubly stochastic by Assumption 1, which means the all-one vector $\mathbf{1} \in \mathbb{R}^k$ satisfies $T\mathbf{1} = \tilde{T}\mathbf{1} = \mathbf{1}$. Hence $\tilde{O}\tilde{T}\mathbf{1} = OT\mathbf{1} = O\mathbf{1} = [\|O_j^\top\|_1]_{j \in [d]}$. We can then compute D as $D = \text{diag}(OT\mathbf{1})$, and recover O as $O = DA$. Finally, T is also recovered since $\tilde{O}\tilde{T} = O\tilde{T} = OT \Rightarrow \tilde{T}T^{-1} = I_k \Rightarrow \tilde{T} = T$.

4.2 Proof of Theorem 3: identifiability of predicting x_2 given x_1 for G-HMM

For G-HMM, the predictor for x_2 given x_1 is parameterized as $f^{2|1}(x_1) = \mathbb{E}[x_2|x_1] = MT\phi(x_1)$. If M, T and $\tilde{M}, \tilde{T}$ produce the same predictor, then

$$f^{2|1}(x) = MT\phi(x) = \tilde{M}\tilde{T}\tilde{\phi}(x) = \tilde{f}^{2|1}(x),\ \forall x \in \mathbb{R}^d. \tag{8}$$

Let $R := (\tilde{M}\tilde{T})^\dagger(MT) \in \mathbb{R}^{k \times k}$, then $\tilde{\phi}(x) = R\phi(x)$. The following lemma (proof deferred to Appendix A.1) says that $\phi, \tilde{\phi}$ must then be equal up to a permutation of coordinates:

Lemma 2. *If there exists a non-singular matrix $R \in \mathbb{R}^{k \times k}$ such that $\phi(x) = R\tilde{\phi}(x)$, $\forall x \in \mathbb{R}^d$, then R must be a permutation matrix.*

[7] We use $[\alpha_i]_{i \in [d]}$ to denote a d-dimensional vector whose i^{th} entry is α_i.

Combined with Lemma 1, we have $\tilde{M}$ is equal to (up to a permutation) either M or HM, where H is the Householder reflection given in Lemma 1.

The remaining step is to show that HM can be ruled out by requiring $\tilde{T}$ to be a stochastic matrix. Note that matching both the predictor and the posterior function means we also have $\tilde{M}\tilde{T} = MT$, or $\tilde{T} = (\tilde{M}^\dagger M)T$. Recall that $H := I_d - 2\hat{v}\hat{v}^\top$ for $\hat{v} = \frac{(M^\dagger)^\top \mathbf{1}}{\sqrt{\mathbf{1}^\top M^\dagger (M^\dagger)^\top \mathbf{1}}}$. When $\tilde{M} = H\tilde{M}$, the column sum of $\tilde{M}^\dagger M$ is

$$\begin{aligned} &\mathbf{1}^\top \tilde{M}^\dagger M = \mathbf{1}^\top M^\dagger H^{-1} M = \mathbf{1}^\top M^\dagger (I - 2\hat{v}\hat{v}^\top) M = \mathbf{1}^\top (I - 2M^\dagger \hat{v}\hat{v}^\top M) \\ =&\mathbf{1}^\top - 2 \cdot \mathbf{1}^\top \frac{M^\dagger (M^\dagger)^\top \mathbf{1}\mathbf{1}^\top M^\dagger M}{\mathbf{1}^\top M^\dagger (M^\dagger)^\top \mathbf{1}} = \mathbf{1}^\top - 2 \cdot \frac{\mathbf{1}^\top M^\dagger (M^\dagger)^\top \mathbf{1}}{\mathbf{1}^\top M^\dagger (M^\dagger)^\top \mathbf{1}} \mathbf{1}^\top = \mathbf{1}^\top - 2 \cdot \mathbf{1}^\top = -\mathbf{1}^\top. \end{aligned} \tag{9}$$

This means the column sum of $\tilde{T}$ is $\mathbf{1}^\top \tilde{T} = \mathbf{1}^\top (\tilde{M}^\dagger M) T = -\mathbf{1}^\top T = -\mathbf{1}^\top$, which violates the constraint that $\tilde{T}$ should be a stochastic matrix with positive entries and column sum 1. Hence it must be that $M = \tilde{M}$ and hence also $T = \tilde{T}$ (up to permutation), proving the theorem statement.

5 Related works

Self-supervised learning On the empirical side, self-supervised methods have gained a great amount of popularity across many domains, including language understanding [Mikolov et al., 2013, Vaswani et al., 2017, Devlin et al., 2018], visual understanding [Doersch et al., 2015, Pathak et al., 2016], and distribution learning [Gutmann and Hyvärinen, 2010, Gao et al., 2020]. Classic ideas such as contrastive learning [Hadsell et al., 2006, Gutmann and Hyvärinen, 2010, Dosovitskiy et al., 2014] and masked prediction [Mikolov et al., 2013] remain powerful in their modern realizations [Hénaff et al., 2019, Chen et al., 2020b, Devlin et al., 2018, Radford et al., 2019, Chen et al., 2020a, He et al., 2021], pushing the state of the art performance and even surpassing supervised pretraining in various aspects [Lee et al., 2021b, Liu et al., 2021].

On the theoretical front, there have been analyses on both masked predictions [Lee et al., 2021a, Zhang and Hashimoto, 2021] and contrastive methods [Arora et al., 2019, Tosh et al., 2020a,b, Wang and Isola, 2020, HaoChen et al., 2021, Wen and Li, 2021], with a focus on characterizing the quality of the learned features for downstream tasks [Saunshi et al., 2020, Wei et al., 2021]. These approaches usually rely on quite strong assumptions to tie the self-supervised learning objective to the downstream tasks of interest. In contrast, our work takes the view of parameter identifiability, for which there is no need for downstream assumptions but instead the specific parametric form is key. Note also that while the parameter recovery lens is a new contribution of our work, Wen and Li [2021] argue (as a side-product of their analysis) that some aspects of a generative model are recovered in their setup. Their data model, however, is substantially different from ours and has very different identifiability properties (owing to its basis in sparse coding).

Latent variable models and tensor methods Latent variable models have been widely studied in the literature. One important area of research is independent component analysis (ICA), where the data is assumed to be given as a transformation (mixing) of unknown independent sources which ICA aims to identify. In nonlinear ICA data models, both the sources and the mixing function are generally not identifiable. However, identifiability of the sources can be shown under some additional assumptions (e.g. on the dependency structure of different time steps) [Hyvarinen and Morioka, 2016, 2017, Hälvä and Hyvarinen, 2020]. Similar ideas have also been applied in the self-supervised setting, where the latent variables can be identified under suitable assumptions on the conditional distribution of the latent [Zimmermann et al., 2021] or on data augmentations [Von Kügelgen et al., 2021]. Unlike our setup though, the mixing function in these models is deterministic and not the object of recovery.

More related to this work is the line of work on learning latent variable models with tensor methods. Specific to learning HMMs, Mossel and Roch [2005] and Anandkumar et al. [2012, 2014] provide algorithms based on third-order moments. A major difference between these prior works on tensor methods and ours is that previous results operate on joint moments, while the results in this work are based on conditional moments given by the optimal predictors for the masked tokens.

6 Conclusion

In this work, we take a model parameter identifiability view of self-supervised learning, which offers a complementary perspective to the current focus of feature learning for downstream performance. By analyzing the masked prediction task in the setup of HMMs and its conditionally-Gaussian variant G-HMM, we showed that parameter recovery is determined by the task difficulty, which can be tuned by both changing the parametric form of the data generative model, and by changing the masked prediction task.

We emphasize that this is a first-cut effort in the research program of analyzing SSL through the lens of model identifiability; we aim to build on this foundation to extend our analyses from HMMs to more complicated latent sequence and latent variable models. We also note that we have focused here on population analyses, and model identifiability. It would be of interest to build off this to develop and analyze the corresponding finite-sample learning algorithms for parametric generative models given SSL tasks, with sample complexity results, both in the realizable case, as well as in the agnostic case where we have model mis-specification. Given the use of conditional MLEs and regressions in SSL, and the natural robustness of these to model mis-specifications, we conjecture that these approaches will be much more robust when compared to say spectral methods.

Overall, we hope this work on an alternative lens to analyze SSL inspires further research.

Acknowledgement

The authors would like to thank Elan Rosenfeld for insightful discussions and comments. We acknowledge the support of ONR via N000141812861, NSF via IIS-1909816, IIS-1955532, IIS-2211907, IIS-1563785, a JP Morgan Faculty Award, and the CMU/PwC DT&I Center.

References

M. Aharon, M. Elad, and A. M. Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear algebra and its applications*, 416(1):48–67, 2006.

E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 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*, pages 33–1. JMLR Workshop and Conference Proceedings, 2012.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014.

S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pages 779–806. PMLR, 2014.

S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

G. Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 771–782, 2015.

E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020a.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020b.

J. E. Cohen and N. Gillis. Identifiability of complete dictionary learning. *SIAM Journal on Mathematics of Data Science*, 1(3):518–536, 2019.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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.

C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014.

R. Gao, E. Nijkamp, D. P. Kingma, Z. Xu, A. M. Dai, and Y. N. Wu. Flow contrastive estimation of energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7518–7528, 2020.

P. Georgiev, F. Theis, and A. Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE transactions on neural networks*, 16(4):992–996, 2005.

M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

H. Hälvä and A. Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR, 2020.

J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34, 2021.

R. Harshman. Foundations of the parafac procedure: Model and conditions for an explanatory factor analysis. *Technical Report UCLA Working Papers in Phonetics 16, University of California, Los Angeles, Los Angeles, CA*, 1970.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. Eslami, and A. v. d. Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems*, pages 3765–3773, 2016.

A. Hyvarinen and H. Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 460–469. PMLR, 20–22 Apr 2017. URL `https://proceedings.mlr.press/v54/hyvarinen17a.html`.

J. B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34, 2021a.

K.-H. Lee, A. Arnab, S. Guadarrama, J. Canny, and I. Fischer. Compressive visual representations. *Advances in Neural Information Processing Systems*, 34, 2021b.

B. G. Lindsay and P. Basak. Multivariate normal mixtures: a fast consistent method of moments. *Journal of the American Statistical Association*, 88(422):468–476, 1993.

H. Liu, J. Z. HaoChen, A. Gaidon, and T. Ma. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021.

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

E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden markov models. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 366–375, 2005.

D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

N. Saunshi, S. Malladi, and S. Arora. A mathematical exploration of why language models help solve downstream tasks. In *International Conference on Learning Representations*, 2020.

D. A. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory*, pages 37–1. JMLR Workshop and Conference Proceedings, 2012.

A. Tamkin, V. Liu, R. Lu, D. Fein, C. Schultz, and N. Goodman. Dabs: A domain-agnostic benchmark for self-supervised learning. *arXiv preprint arXiv:2111.12062*, 2021.

C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive learning, multi-view redundancy, and linear models, 2020a.

C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive estimation reveals topic posterior information to linear models. *arXiv preprint arXiv:2003.02234*, 2020b.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

J. Von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

M. Vuffray, S. Misra, A. Lokhov, and M. Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. *Advances in neural information processing systems*, 29, 2016.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020.

X. Wang, X. Chen, S. S. Du, and Y. Tian. Towards demystifying representation learning with non-contrastive self-supervision. *arXiv preprint arXiv:2110.04947*, 2021.

C. Wei, S. M. Xie, and T. Ma. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. *Advances in Neural Information Processing Systems*, 34, 2021.

Z. Wen and Y. Li. Toward understanding the feature learning process of self-supervised contrastive learning. *International Conference on Machine Learning*, page 11112–11122, 2021.

S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

T. Zhang and T. Hashimoto. On the inductive bias of masked language modeling: From statistical to syntactic dependencies. *arXiv preprint arXiv:2104.05694*, 2021.

R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The limitations and future directions are discussed in the conclusion section.
   (c) Did you discuss ctions any potential negative societal impacts of your work? [N/A] This work is purely theoretical.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Assumptions are listed in Section 2.3, together with a discussion on their implications.
   (b) Did you include complete proofs of all theoretical results? [Yes] Proofs for Theorem 3 and 5 are presented in Section 4. Missing proofs are provided in Appendix A for GHMMs, and Appendix B for HMMs.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Missing proofs for G-HMM

This section provides missing proofs for results on G-HMM. We will first prove the two lemmas on properties of the posterior function (Lemma 1, 2), then show the proof for the three-token prediction task (Theorem 6) using a tensor decomposition idea similar to that of Theorem 5. At the end, we show the identifiability from pairwise conditional distributions (as opposed to conditional expectation as in masked prediction tasks), which is proved by reducing parameter recovery to the identifiability of Gaussian mixtures (Theorem 7).

A.1 Proofs of helper lemmas

A.1.1 Proof for Lemma 2

Given the form of the predictor, matching two predictors $f, \tilde{f}$ means that the corresponding posteriors $\phi, \tilde{\phi}$ are matched up to a linear transformation. We will now prove the following lemma, which says that in this case, $\phi, \tilde{\phi}$ can in fact only differ by a permutation of coordinates:

Lemma (Lemma 2 restated)**.** *If there exists a non-singular matrix $R \in \mathbb{R}^{k\times k}$ such that $\phi(x) = R\tilde{\phi}(x)$, $\forall x \in \mathbb{R}^d$, then R must be a permutation matrix.*

Proof. We will prove the lemma by matching the Jacobian w.r.t. x on both sides. Let's first quickly recall the Jacobian of the posterior vector $\phi(x) \in \mathbb{R}^k$, where $[\phi(x)]_i = \frac{\exp(-\frac{\|x-\mu_i\|^2}{2})}{\sum_{j\in[k]}\exp(-\frac{\|x-\mu_j\|^2}{2})}$. Denote $o(x) := \big[-\frac{\|x-\mu_1\|^2}{2}, ..., -\frac{\|x-\mu_k\|^2}{2}\big] \in \mathbb{R}^k$, then $\nabla_x\phi(x) = \nabla_{o(x)}\text{softmax}(o(x)) \cdot \nabla_x o(x)$, where

$$\begin{aligned}\nabla_o[\text{softmax}(o)]_i &= [\text{softmax}(o)]_i \cdot (e_i - \text{softmax}(o)) = [\phi(x)]_i \cdot (e_i - \phi(x)),\\ \nabla_o\text{softmax}(o) &= \text{diag}(\phi(x)) - \phi(x)\phi(x)^\top,\\ \nabla_x o(x) &= -\left[x-\mu_1, ..., x-\mu_k\right]^\top.\end{aligned} \tag{10}$$

Hence the Jacobian is

$$\nabla_x\phi(x) = \big(\text{diag}(\phi(x)) - \phi(x)\phi(x)^\top\big) \cdot (M - [x, x, ..., x])^\top. \tag{11}$$

Denote $\Delta := M - [x, x, ..., x] \in \mathbb{R}^{d\times k}$, and similarly $\tilde{\Delta} = \tilde{M} - [x, x, ..., x]$. Matching $\nabla_x\tilde{\phi}(x) = \nabla_x R\phi(x)$ gives

$$\text{diag}(R\phi(x))\tilde{\Delta}^\top - R\phi(x)(\tilde{\Delta}R\phi(x))^\top = R\text{diag}(\phi(x))\Delta^\top - R\phi(x)(\Delta\phi(x))^\top. \tag{12}$$

Let's take $x = x_c^{(i)} := c\mu_i$ for $c > 1$. We claim that this $x_c^{(i)}$ satisfies $\lim_{c\to\infty}\phi(x_c^{(i)}) \to e_i$. This is because $\forall j \neq i$,

$$\begin{aligned}\lim_{c\to\infty}\frac{[\phi(x_c^{(i)})]_j}{[\phi(x_c^{(i)})]_i} &= \lim_{c\to\infty}\exp\Big(\frac{\|c\mu_i - \mu_i\|^2}{2} - \frac{\|c\mu_i - \mu_j\|^2}{2}\Big)\\ = \lim_{c\to\infty}\exp\Big(-\frac{\big((2c-1)\mu_i - \mu_j\big)^\top(\mu_i - \mu_j)}{2}\Big) &= \lim_{c\to\infty}\exp\Big(-\frac{2c\mu_i^\top(\mu_i-\mu_j)}{2}\Big) = 0\end{aligned} \tag{13}$$

where the last equality is because $\mu_i^\top(\mu_i - \mu_j) > 0$ for any μ_i, μ_j lying on the same hypersphere.

With such choices of x, the two sides of equation 12 are now:

$$\begin{aligned}LHS &= \text{diag}(R_i)\tilde{\Delta}^\top - R_iR_i^\top\tilde{\Delta}^\top = (\text{diag}(R_i) - R_iR_i^\top)\tilde{\Delta}^\top\\ = RHS &= \sum_{j\in[k]}[e_i]_jR_j(\Delta_j)^\top - Re_i(\Delta e_i)^\top = R_i(\Delta_i)^\top - R_i(\Delta_i)^\top = 0.\end{aligned} \tag{14}$$

Since $x := c\mu_i$ for $c \to \infty$ lies outside the affine hull of $\{\tilde{\mu}_i\}_{i\in[k]}$, $\tilde{\Delta}$ is of full rank due to the following claim:

Claim 2. *Given a linearly independent set $\{u_i\}_{i\in[k]}$, if $\{u_i - v\}_{i\in[k]}$ is not linearly independent, then $v = \sum_{i\in[k]} \beta_i \cdot u_i$ where $\sum_{i\in[k]} \beta_i = 1$.*

Proof. Since $\{u_i - v\}_{i\in[k]}$ is linearly dependent, we can write some $u_j - v$ as the linear combination of other $\{u_i - v\}_{i\in[k], i\neq j}$. Let's take $j = k$ wlog, and denote the coefficients of the linear combination as $\{\alpha_i\}_{i\in[k-1]}$. Then

$$u_k - v = \sum_{i\in[k-1]} \alpha_i(u_i - v) \Rightarrow \big(1 - \sum_{i\in[k-1]} \alpha_i\big)v = -\sum_{i\in[k-1]} \alpha_i \cdot u_i + u_k \tag{15}$$

The right hand side is non-zero since $\{u_i\}_{i\in[k]}$ are linearly independent by assumption, hence $1 - \sum_{i\in[k-1]} \alpha_i \neq 0$, and we get

$$v = \sum_{i\in[k-1]} \underbrace{\frac{-\alpha_i}{1 - \sum_{i\in[k-1]} \alpha_i}}_{:=\beta_i} \cdot u_i + \underbrace{\frac{1}{1 - \sum_{i\in[k-1]} \alpha_i}}_{:=\beta_k} u_k. \tag{16}$$

Note that $\sum_{i\in[k]} \beta_i = 1$, hence v is an affine combination of $\{u_i : i \in [k]\}$. □

Since $\tilde{\Delta}$ is full rank, it must be $\text{diag}(R_i) - R_i R_i^\top = 0$, which implies R is a permutation matrix. This is because for any non-zero v s.t. $\text{diag}(v) - vv^\top = 0$, the entries of v satisfy $v_i^2 = 1$, $v_i v_j = 0$ for $i \neq j$. Hence v has exactly one non-zero entry which is ± 1. Since $R\phi(x) = \tilde{\phi}(x)$ where $\phi(x), \tilde{\phi}(x)$ are both probability vectors with non-negative entries, this non-zero entry has to be 1 (and not -1). Since R is of rank-k by Assumption 4, this non-zero entry is at different positions for different R_i, hence R is a permutation matrix.

□

A.1.2 Proof of Lemma 1

We show that if $M, \tilde{M}$ parameterize $\phi, \tilde{\phi}$ respectively and that $\phi = \tilde{\phi}$, then $\tilde{M}$ must equal to either M or a unique (and somewhat special) transformation of M:

Lemma (Lemma 1 restated)**.** *For $d \geq k \geq 2$, then under Assumption 4, 5, $\phi = \tilde{\phi}$ implies $\tilde{M} = M$ or $\tilde{M} = HM$, where H is a Householder transformation of the form $H := I_d - 2\hat{v}\hat{v}^\top \in \mathbb{R}^{d\times d}$, with $\hat{v} := \frac{(M^\dagger)^\top \mathbf{1}}{\sqrt{\mathbf{1}^\top M^\dagger (M^\dagger)^\top \mathbf{1}}}$.*

Proof. Let's start with $d = k$. First, let's check the conditions for $\phi = \tilde{\phi}$. For any $x \in \mathbb{R}^d$, we have

$$\begin{aligned}
&[\phi(x)]_i = \frac{\exp\big(-\frac{\|x-\mu_i\|^2}{2}\big)}{\sum_{j\in[k]} \exp\big(-\frac{\|x-\mu_j\|^2}{2}\big)} = \frac{\exp\big(-\frac{\|x-\tilde{\mu}_i\|^2}{2}\big)}{\sum_{j\in[k]} \exp\big(-\frac{\|x-\tilde{\mu}_j\|^2}{2}\big)} = [\tilde{\phi}(x)]_i,\ \forall i \in [k] \\
&\Rightarrow \frac{\exp\big(-\frac{\|x-\mu_i\|^2}{2}\big)}{\exp\big(-\frac{\|x-\tilde{\mu}_i\|^2}{2}\big)} = \frac{\exp\big(-\frac{\|x-\mu_j\|^2}{2}\big)}{\exp\big(-\frac{\|x-\tilde{\mu}_j\|^2}{2}\big)}, \forall i, j \in [k] \\
&\Rightarrow \|x-\mu_i\|^2 - \|x-\tilde{\mu}_i\|^2 = \|x-\mu_j\|^2 - \|x-\tilde{\mu}_j\|^2,\ \forall i, j \in [k] \\
&\Rightarrow 2\big((\tilde{\mu}_i - \mu_i) - (\tilde{\mu}_j - \mu_j)\big)^\top x = \big(\|\mu_j\|^2 - \|\tilde{\mu}_j\|^2\big) - \big(\|\mu_i\|^2 - \|\tilde{\mu}_i\|^2\big),\ \forall i, j \in [k].
\end{aligned} \tag{17}$$

Since the left hand side is linear in $x \in \mathbb{R}^d$ and the right hand side is a constant, it must be that both sides are 0. That is, the necessary conditions for $\phi = \tilde{\phi}$ are that for any $i, j \in [k]$, 1) $\tilde{\mu}_i - \mu_i = \tilde{\mu}_j - \mu_j$, and 2) $\|\mu_i\|^2 - \|\tilde{\mu}_i\|^2 = \|\mu_j\|^2 - \|\tilde{\mu}_j\|^2$. It can be checked that these two conditions are also sufficient for $\phi = \tilde{\phi}$.

Denote $v := \mu_i - \tilde{\mu}_i$. The norms of the means are known and equal by Assumption 5, which gives

$$\|\mu_i\|^2 - \|\tilde{\mu}_i\|^2 = \|\mu_i\|^2 - \|\mu_i - v\|^2 = (2\mu_i - v)^\top v = 0,\ \forall i \in [k]. \tag{18}$$

The last equality in equation 18 holds for a non-zero v when the span of $\{2\mu_i - v : i \in [k]\}$ is $(d-1)$-dimensional subspace. On the other hand, the span of $\{2\mu_i - v : i \in [k]\}$ is at least $(k-1)$ by Assumption 4. When $d = k$, it must be that the dimension is exactly $(d-1)$, which means v is an affine combination of $\{2\mu_i : i \in [k]\}$ by Claim 2.

Moreover, v has to be orthogonal to $\{2\mu_i - v : i \in [k]\}$, which leads to the unique choice of v that is the projection of the origin onto the $(d-1)$-dimensional subspace specified by the affine combinations of $\{2\mu_i : i \in [k]\}$.

Claim 3. *v is the projection of the origin to the hyperplane defined by $\{2\mu_i : i \in [k]\}$, and is the only solution to equation 18.*

Proof. It is clear that this choice of v satisfies $(2\mu_i - v)^\top v = 0, \forall i \in [k]$. To see that this is the unique choice, suppose there exists some v' lying in the hyperplane of $\{2\mu_i\}$, and denote $\delta := v' - v$.

Note that $\delta^\top v = 0$: let the hyperplane specified by $\{2\mu_i\}_{i\in[k]}$ be specified as $\{x : \langle u, x\rangle = c\}$ for some $u \in \mathbb{R}^d$ and $c \in \mathbb{R}$. Then v, the projection of the origin, can be written as $v = \frac{c}{\|u\|} \cdot \frac{u}{\|u\|}$, i.e. v is proportional to the normal vector u. For any v' in the hyperplane, it satisfy $\langle u, v'\rangle = c$, and

$$\begin{aligned}\delta^\top v =&(v'-v)^\top v = \langle \frac{c}{\|u\|}\frac{u}{\|u\|}, v'\rangle - \left\|\frac{c}{\|u\|}\frac{u}{\|u\|}\right\|^2 \\ =&\frac{c}{\|u\|^2}\cdot\langle u, v'\rangle - \frac{c^2}{\|u\|^2}\frac{\|u\|^2}{\|u\|^2} = \frac{c^2}{\|u\|^2} - \frac{c^2}{\|u\|^2} = 0.\end{aligned} \tag{19}$$

Then for any v' satisfying equation 18,

$$\begin{aligned}&(2\mu_i - v')^\top v' = (2\mu_i - v - \delta)^\top (v + \delta) \\ =&\underbrace{(2\mu_i - v)^\top v}_{0} + 2\mu_i^\top \delta - \underbrace{v^\top \delta}_{0} - \underbrace{\delta^\top v}_{0} - \delta^\top\delta = (2\mu_i - \delta)^\top \delta = 0,\ \forall i \in [k].\end{aligned} \tag{20}$$

Since $\{2\mu_i - \delta\}_{i\in[k]}$ spans the $(k-1)$-dimensional hyperplane and that δ lies in the hyperplane, it must be that $\delta = 0$, i.e. $v' = v$. □

Note that this choice of v also satisfies $\|\mu_i - v\| = \|\mu_i\|$, since v and the origin are reflections w.r.t. the hyperplane that is the affine hull of $\{\mu_i : i \in [k]\}$. In other words, $\{\mu_i - v\}_{i\in[k]}$ is related to $\{\mu_i\}_{i\in[k]}$ via the Householder transformation of the form $H := I_d - 2\frac{vv^\top}{\|v\|^2}$, i.e. $\mu_i - v = H\mu_i$. Denote $\hat{v} := \frac{v}{\|v\|_2}$. An explicit formula for $\hat{v}$ is $\hat{v} := \frac{M^{-\top}\mathbf{1}}{\sqrt{\mathbf{1}^\top M^{-1} M^{-\top}\mathbf{1}}}$. This finishes the proof for $d = k$.

For $d > k$, the above argument still applies and H remains the only indeterminacy (up to permutation), where $H := I_d - 2\hat{v}\hat{v}^\top$ for $\hat{v} := \frac{(M^\dagger)^\top \mathbf{1}}{\sqrt{\mathbf{1}^\top M^\dagger (M^\dagger)^\top \mathbf{1}}}$. The reason is that even though the ambient dimension d is larger, $\{\mu_i - v : i \in [k]\}$ has to have the same span as $\{\mu_i : i \in [k]\}$, since having the same predictor requires the column space of $M, \tilde{M}$ to match. Hence we only need to consider v in the k-dimensional column space of M, which reduces to the case of $d = k$.

□

A.2 Identifiability of predicting $x_{t_2} \otimes x_{t_3} | x_{t_1}$, G-HMM

Theorem 5 shows that triplet prediction tasks (i.e. predict 2 tokens given 1) suffices for the identifiability of HMM, using tools from the uniqueness of tensor decomposition. The next theorem shows that the same conclusion also applies for G-HMM:

Theorem 6 (Identifiability from masked prediction on three tokens, G-HMM)**.** *Let (t_1, t_2, t_3) be any permutation of $(1, 2, 3)$, and consider the prediction task $x_{t_2} \otimes x_{t_3} | x_{t_1}$. Under Assumption 1, 4, 5, if the optimal predictors under the G-HMM distributions with parameters (M, T) and $(\tilde{M}, \tilde{T})$ are the same, then $(M, T) = (\tilde{M}, \tilde{T})$ up to a permutation of the hidden state labels.*

Proof. Similar to the discrete case, we will prove $x_2 \otimes x_3 | x_1$ and $x_1 \otimes x_3 | x_2$ separately; the proof for $x_1 \otimes x_2 | x_3$ is analogous to $x_2 \otimes x_3 | x_1$ by symmetry and hence omitted. The proofs also follow a similar strategy as in the proof for Theorem 5, that is, to construct a 3-tensor using the predictor, on which applying Kruskal's theorem provides identifiability.

Case 1, $x_2 \otimes x_3 | x_1$**:** Let $\mathcal{X} := \{x^{(i)} \in \mathbb{R}^d : i \in [k]\}$ be a linearly independent set, and consider the following 3-tensor:

$$\begin{aligned} W :=& \sum_{x_i \in \mathcal{X}} x_1 \otimes \mathbb{E}[x_2 \otimes x_3 | x_1] = \sum_{x_1 \in \mathcal{X}} x_1 \otimes \mathbb{E}_{h_2|x_1}\big[\mathbb{E}[x_2 \otimes x_3|x_1]|h_2\big] \\ =& \sum_{x_1 \in \mathcal{X}} x_1 \otimes \sum_{h_2} P(h_2|x_1)\mathbb{E}[x_2|h_2] \otimes \mathbb{E}[x_3|h_2] \\ =& \sum_{i \in [k]} \sum_{x_1 \in \mathcal{X}} P(h_2 = i|x_1) x_1 \otimes \mathbb{E}[x_2|h_2 = i] \otimes \mathbb{E}[x_3|h_2 = i] \\ =& \sum_{i \in [k]} \underbrace{\Big(\sum_{x_1} (T\phi(x_1))^\top e_i^{(k)} x_1\Big)}_{:=a_i} \otimes M_i \otimes (MT)_i. \end{aligned} \tag{21}$$

The matrices formed by second and third components are both of rank-k by Assumption 4. Hence in order to apply Kruskal's theorem on W, it suffices to show that there exists a choice of $\mathcal{X}$ such that the matrix $A := [a_1, ..., a_k]$ is of rank k. One such choice is to let $x^{(i)} = \mu_i$, which gives

$$\begin{aligned} a_i :=& \sum_{j \in [k]} \phi(x_1 = \mu_j)^\top T^\top e_i^{(k)} \mu_j = M[\phi(\mu_1), ..., \phi(\mu_k)]^\top T^\top e_i^{(k)}, \\ A :=& [a_1, ..., a_k] = M[\phi(\mu_1), ..., \phi(\mu_k)]^\top T^\top. \end{aligned} \tag{22}$$

Since M, T are both of rank k by Assumption 4, we only need to argue that the matrix $\Phi := [\phi(\mu_1), ..., \phi(\mu_k)] \in \mathbb{R}^{k \times k}$ is of full rank. Recall that for a mixture of k Gaussians with identify covariance and mean $\{\mu_i \in \mathbb{R}^d : i \in [k]\}$, the posterior function ϕ is defined entrywise as

$$[\phi(x)]_i = \frac{\exp\big(-\frac{\|x-\mu_i\|_2^2}{2}\big)}{\sum_{j\in[k]} \exp\big(-\frac{\|x-\mu_j\|_2^2}{2}\big)}, \ \forall i \in [k]. \tag{23}$$

To show Φ is of full rank, we can equivalently show that a columnwise scaled version of Φ is full rank. In particular, let's look at the matrix $\hat{\Phi} \in \mathbb{R}^{k \times k}$, where $\hat{\Phi}_{ij} = \exp(-\frac{\|\mu_i - \mu_j\|^2}{2})$; that is, each column of $\hat{\Phi}$ can be considered as a scaled version of the column in Φ without the normalization for a unit ℓ_1 norm. It can be seen that $\hat{\Phi}$ is a Gaussian kernel matrix which is known to be full rank.

Therefore we have shown that each component of the tensor $W := \sum_{i \in [k]} a_i \otimes M_i \otimes (MT)_i$ has Kruskal rank k, which allows to recover columns of M, MT up to permutation and scaling by Kruskal's theorem. The indeterminacy in scaling is further removed since the norms of $\{M_i\}_{i \in [d]}$ are known by Assumption 5.

On the other hand, for any $\tilde{M}, \tilde{T}$ that form the same predictor as M, T, W can also be written as

$$\begin{aligned} W =& \sum_{x_1 \in \mathcal{X}} x_1 \otimes \mathbb{E}[x_2 \otimes x_3|x_1] = \sum_{x_1 \in \mathcal{X}} x_1 \otimes \tilde{\mathbb{E}}[x_2 \otimes x_3 | x_1] \\ =& \sum_{i \in [k]} \Big(\sum_{x_1} (\tilde{T}\tilde{\phi}(x_1))^\top e_i^{(k)} x_1\Big) \otimes \tilde{M}_i \otimes (\tilde{M}\tilde{T})_i. \end{aligned} \tag{24}$$

Hence columns of $M, \tilde{M}$ and $MT, \tilde{M}\tilde{T}$ are both matched up to a shared permutation, which proves identifiability.

Case 2, $\mathbb{E}[x_1 \otimes x_3 | x_2]$**:** For the task of predicting x_1, x_3 given x_2, the predictor takes the form

$$\mathbb{E}[x_1 \otimes x_3|x_2] = (OT^\top)\text{diag}(\phi(x_2))(OT)^\top. \tag{25}$$

Let $\mathcal{X} := \{\mu_i : i \in [k]\}$ as in the previous case, and consider the 3-tensor

$$\begin{aligned}W := &\sum_{x_2 \in \mathcal{X}} x_2 \otimes \mathbb{E}[x_1 \otimes x_3 | x_2] = \sum_{x_2 \in \mathcal{X}} x_2 \otimes \mathbb{E}_{h_2|x_2}(\mathbb{E}[x_1|h_2] \otimes \mathbb{E}[x_3|h_2]) \\ =&\sum_{h_2} \sum_{x_2 \in \mathcal{X}} p(h_2|x_2) x_2 \otimes \mathbb{E}[x_1|h_2] \otimes \mathbb{E}[x_3|h_2] \\ =&\sum_{i \in [k]} \underbrace{\Big(\sum_{x_2 \in \mathcal{X}} (\phi(x_2))^\top e_i^{(k)} x_2 \Big)}_{:=a_i} \otimes (MT^\top)_i \otimes (MT)_i,\end{aligned} \tag{26}$$

The first component is of rank-k as shown in the proof for $x_2 \otimes x_3|x_1$, and the other two components are of rank-k by Assumption 4. Thus Kruskal's theorem applies and the columns of $MT, MT^\top$ are recovered up to a shared permutation.

The first component $\{a_i\}_{i\in[k]}$ are also recovered, which means that if $\tilde{M}, \tilde{T}$ form the same predictor as M, T, then for any linearly independent set $\mathcal{X}$ with k elements (not necessarily the previous choice of $\{\mu_i\}_{i\in[k]}$) such that $\mathcal{X}$ leads to a full rank A, we have $A = \tilde{A}$ where $\tilde{A}$ is parameterized by $\tilde{M}, \tilde{T}$. For any such $\mathcal{X} = \{x^{(i)} : i \in [k]\}$, denote $X := [x^{(1)}, ..., x^{(k)}]$, then

$$A = X[\phi(x^{(1)}), ..., \phi(x^{(k)})]^\top T^\top = X[\tilde{\phi}(x^{(1)}), ..., \tilde{\phi}(x^{(k)})]^\top \tilde{T}^\top = \tilde{A}. \tag{27}$$

Since X is of rank-k by the choice of $\mathcal{X}$, this means

$$[\tilde{\phi}(x^{(1)}), ..., \tilde{\phi}(x^{(k)})] = \underbrace{\tilde{T}^{-1}T}_{:=R}[\phi(x^{(1)}), ..., \phi(x^{(k)})] \Rightarrow \tilde{\phi}(x^{(i)}) = R\phi(x^{(i)}),\ \forall i \in [k]. \tag{28}$$

Moreover, for any valid choice of $\mathcal{X}$, matrices defined with points in sufficiently small neighborhoods of $x^{(i)}$ are still of full rank by the upper continuity of matrix rank. Hence the equality in equation 28 holds for points in these neighborhoods, and thus the Jacobian on both sides should be equal. Then, the exact same proof of Lemma 2 applies, and we get $\tilde{\phi}$, ϕ are equal up to a permutation of coordinates. Thus $\tilde{M}$ must be equal to (up to permutation) either M or HM for a Householder reflection H by Lemma 1. Finally, the solution of HM is eliminated since it would lead to a $\tilde{T}$ that is not a valid stochastic matrix, as shown in the proof of Theorem 3.

□

A.3 Identifiability from pairwise conditional distribution

We show that matching the entire conditional *distribution* for G-HMM provides identifiability. Though this is implied by Theorem 3, which states that matching the conditional *expectation* already suffices, having access to the full conditional distribution allows an even simpler proof.

Theorem 7 (Identifiability of conditional distribution)**.** *Let M, T and $\tilde{M}, \tilde{T}$ be two set of parameters satisfying Assumption 1 and 4. If $p(x_2|x_1; M, T) = p(x_2|x_1; \tilde{M}, \tilde{T})$, $\forall x_1, x_2 \in \mathbb{R}^d$, then $M = \tilde{M}$, $T = \tilde{T}$ up to a permutation of labeling.*

Proof. First note that the conditional distribution of x_2 given x_1 is a mixture of Gaussian, with means $\{\mu_i\}_{i\in[k]}$ and mixture weights given by $P(h_2|x_1) = TP(h_1|x_1)$, hence we can directly apply the identifiability of Gaussian mixtures to recover the means $\{\mu_i\}_{i\in[k]}$:

Lemma 3 (Proposition 4.3 in Lindsay and Basak [1993])**.** *Let Q_k denote a Gaussian mixture with means $\{\xi_j\}_{j\in[k]} \in \mathbb{R}^d$. Suppose $\exists l \in [d]$ such that the set $\{[\xi_j]_l\}$ has distinct values, then one can recover $\{\xi_j\}_{j\in[k]}$ from moments of Q_k.*

We note that the assumption on the existence of a coordinate $l \in [k]$ is with out loss of generality, since we can first rotate the means to a different coordinate system in which this condition holds, then rotation back the means. Such rotation is guaranteed to exist since finding such rotation is equivalent to finding a vector v s.t. $v^\top(\mu_i - \mu_j) \neq 0$ for every $i, j \in [k]$, for which the solution set is $\mathbb{R}^d \setminus \cup_{i,j\in[k]}\{u : u^\top(\mu_i - \mu_j) = 0\} \neq \emptyset$.

Recovering $\{\mu_i\}_{i\in[k]}$ means the scaled likelihood and the posterior both match, i.e. $\psi = \tilde{\psi}$, and $\phi(x) = P(h|x) = \frac{\psi}{\|\psi\|_1}$. The conditional distribution is

$$p(x_2|x_1) = \sum_{i,j\in[k]} p(x_2|h_2)p(h_2|h_1)p(h_1|x_1) = \frac{1}{(2\pi)^{d/2}}\psi(x_2)^\top T\phi(x_1). \tag{29}$$

Choose a set $\mathcal{X} := \{x^{(i)}\}_{i\in[k]}$ such that $\Psi_{\mathcal{X}} := [\psi(x^{(1)}), ..., \psi(x^{(k)})] \in \mathbb{R}^{k\times k}$ is full rank. $\Phi_{\mathcal{X}} := [\phi(x^{(1)}), ..., \phi(x^{(k)})] \in \mathbb{R}^{k\times k}$ is also full rank since its columns are nonzero scalings of columns of $\Psi_{\mathcal{X}}$. Then we have

$$\Psi_{\mathcal{X}}^\top T\Phi_{\mathcal{X}} = \tilde{\Psi}_{\mathcal{X}}^\top \tilde{T}\tilde{\Phi}_{\mathcal{X}} = \Psi_{\mathcal{X}}^\top \tilde{T}\Phi_{\mathcal{X}} \Rightarrow T = \tilde{T}. \tag{30}$$

□

B Proof of Theorem 4: non-identifiability of HMM from multiple pairwise predictions

Theorem (Theorem 4 restated: nonidentifiability of HMM from multiple pairwise predictions)**.** *There exists a pair of HMM distributions with parameters (O, T) and $(\tilde{O}, \tilde{T})$, each satisfying Assumptions 1, 2 and 3, and also $\tilde{O} \neq O$, such that, for each of the tasks $x_2|x_1$, $x_1|x_2$, $x_3|x_1$, and $x_1|x_3$, the optimal predictors are the same under each distribution.*

Proof. We provide an example to show the nonidentifiability result in Theorem 4. The goal is to find $\tilde{O} \neq O, \tilde{T} \neq T$ that produce the same predictors for predicting both $x_2|x_1$ and $x_3|x_1$. We will choose $T, \tilde{T}$ to be symmetric, so that O, T and $\tilde{O}, \tilde{T}$ also form the same predictors for the reversed direction, i.e., for predicting x_1 given x_2 and x_1 given x_3, since the reverse chain has transition matrix $T^\top = T$.

Let's consider the case where the all row sums of O and $\tilde{O}$ are k/d. Consequently, the posterior function is simply $\phi(x) = \frac{O^\top x}{\|O^\top x\|_1} = \frac{d}{k}O^\top x$, and similarly we have $\tilde{\phi}(x) = \frac{d}{k}\tilde{O}^\top x$. The predictors are of the form:

$$f^{2|1}(x) = OT\phi(x) = \frac{d}{k}OTO^\top x,\ f^{3|1}(x) = OT^2\phi(x) = \frac{d}{k}OT^2O^\top x. \tag{31}$$

Matching $f^{2|1}(x) = \tilde{f}^{2|1}(x)$ on all $x \in \mathcal{X} := \{e_i\}_{i\in[d]}$ means

$$OTO^\top I_d = OTO^\top = \tilde{O}\tilde{T}\tilde{O}^\top \Rightarrow \tilde{T} = \tilde{O}^\dagger O \cdot T \cdot (\tilde{O}^\dagger O)^\top. \tag{32}$$

Similarly, matching $f^{3|1} = \tilde{f}^{3|1}$ gives $OT^2O^\top = \tilde{O}\tilde{T}^2\tilde{O}^\top$, hence

$$\begin{aligned} \tilde{O}\tilde{T}^2\tilde{O}^\top &= \tilde{O}\tilde{O}^\dagger OT(\tilde{O}^\dagger O)^\top \cdot \tilde{O}^\dagger OT(\tilde{O}^\dagger O)^\top\tilde{O}^\top \\ &\overset{(i)}{=} OT\cdot(\tilde{O}^\dagger O)^\top\tilde{O}^\dagger O\cdot TO^\top = OT\cdot TO \Rightarrow (\tilde{O}^\dagger O)^\top\cdot\tilde{O}^\dagger O = I_k, \end{aligned} \tag{33}$$

where step (i) uses $\tilde{O}\tilde{O}^\dagger O = O$, since $\tilde{O}, O$ share the same column space.

Denote $R := \tilde{O}^\dagger O$; R is orthogonal by the last equality in equation 33. To construct the desired example, consider $k = 3$, and let R represent a rotation with axis of rotation $\frac{1}{3}(e_1 + e_2 + e_3)$. This axis is the direction pointing from the origin to the projection of the origin on the hyperplane $\mathcal{P}_c := \{v \in \mathbb{R}^d : \sum_{i\in[d]} v_i = c\}$ for any positive constant c (i.e. $\mathcal{P}_c$ is parallel to the hyperplane in which probability vectors lie). This means such rotation guarantees $Rv \in \mathcal{P}_c, \forall v \in \mathcal{P}_c$, and has the following property:

Claim 4. *Each row and each column of R sums up to 1.*

Define $\tilde{O} := OR, \tilde{T} := R^\top TR$, Claim 4 ensures that row sum and column sum of $\tilde{O}, \tilde{T}$ remain the same as those of O, T. When the rotation angle represented by R is sufficiently small, entries $\tilde{O}, \tilde{T}$ remain in $[0, 1]$, hence such $\tilde{O}, \tilde{T}$ form a valid example. We will provide a concrete example in the subsequent subsection.

□

B.1 Example for Theorem 4

The intuition of the nonidentifiability result in Theorem 4 is related to the non-uniqueness of matrix factorization: while adding additional pairwise prediction tasks introduces more equations on the product of matrices, these equations can be highly dependent, and there are cases where different set of matrices can simultaneously satisfy all the equations.

We now provide a concrete example for the non-identifiability of predicting $x_2|x_1$, $x_1|x_2$, $x_3|x_1$, and $x_1|x_3$, by finding 2 set of O, T such that the corresponding predictors (of the form specified in equation 31) match. Let $d = 4$, $k = 3$,

$$\begin{aligned} &O = \begin{bmatrix} 0.23016003 & 0.3549092 & 0.16493077 \\ 0.30716059 & 0.06962305 & 0.37321636 \\ 0.2580854 & 0.26965425 & 0.22226035 \\ 0.20459398 & 0.3058135 & 0.23959252 \end{bmatrix}, \tilde{O} = \begin{bmatrix} 0.24120928 & 0.35062535 & 0.15816537 \\ 0.28937626 & 0.07433156 & 0.38629218 \\ 0.26077674 & 0.26749114 & 0.22173212 \\ 0.20863772 & 0.30755194 & 0.23381033 \end{bmatrix}, \\ &T = \begin{bmatrix} 0.56893146 & 0.35811118 & 0.07295736 \\ 0.35811118 & 0.10805638 & 0.53383243 \\ 0.07295736 & 0.53383243 & 0.39321021 \end{bmatrix}, \tilde{T} = \begin{bmatrix} 0.59740926 & 0.30452087 & 0.09806987 \\ 0.30452087 & 0.1331689 & 0.56231024 \\ 0.09806987 & 0.56231024 & 0.33961989 \end{bmatrix}, \\ &\det(O) = \det(\tilde{O}) = 0.0110, \det(T) = \det(\tilde{T}) = -0.1611. \end{aligned} \tag{34}$$

Note that $T, \tilde{T}$ are both symmetric as desired by the proof of Theorem 4, which means this is also a valid counter example for learning to predict $x_1|x_2$ and $x_1|x_3$, and hence for all of $x_2|x_1$, $x_1|x_2$, $x_3|x_1$, and $x_1|x_3$.

B.2 Proof of Claim 4

Proof. We would like to show that each row and each column of R sums up to 1. Denote the d-dimensional simplex by Δ_d, i.e. $\Delta_d := \{x \in \mathbb{R}^d : \sum_{i\in[d]} x_i = 1\}$, and let $\mathcal{P}_c := \{v \in \mathbb{R}^d : \sum_{i\in[d]} v_i = c\}$ for some positive constant c denote a hyperplane parallel to the hyperplane in which probability vectors lie.

Let's first check that the columns of R sum up to 1. Any $v \in \mathcal{P}_c$ can be written as $v = c \cdot [\alpha_1, \alpha_2, ..., \alpha_{d-1}, 1 - \sum_{i\in[d-1]} \alpha_i]$ for some $[\alpha_1, ..., \alpha_{d-1}] \in \Delta_{d-1}$. Let r_i denote the i_{th} row of R, then $Rv \in \mathcal{P}_c$ means $\sum_{i\in[d]} \langle r_i, v\rangle = \langle \sum_{i\in[d]} r_i, v\rangle = c$. Let β_j denote the j_{th} coordinate of $\sum_{i\in[d]} r_i$, then

$$\begin{aligned} &\sum_{i\in[d-1]} \beta_i\alpha_i + \beta_d\big(1 - \sum_{i\in[d-1]} \alpha_i\big) = 1, \ \forall[\alpha_1, ..., \alpha_{d-1}] \in \Delta_{d-1} \\ \Rightarrow &\sum_{i\in[d-1]} (\beta_i - \beta_d)\alpha_i + \beta_d = 1, \ \forall[\alpha_1, ..., \alpha_{d-1}] \in \Delta_{d-1} \\ \Rightarrow &\ \beta_i = 1, \ \forall i \in [d]. \end{aligned} \tag{35}$$

It then follows that $R^{-1} = R^\top$ also has columns summing up to 1, since

$$\sum_{i\in[d]} (RR^{-1})_{ij} = \langle \sum_{i\in[d]} r_i, (R^{-1})_j\rangle = \langle \mathbf{1}, (R^{-1})_j\rangle = 1, \ \forall j \in [d]. \tag{36}$$

□

C Nonidentifiability from large time gaps

As noted earlier, there is an inherent obstacle when using prediction tasks on tokens that are more than 1 time gaps apart. For instance, if we are predicting x_{t+1} given x_1 for some $t > 1$ with G-HMM, then we are still able to identify M from the posterior function, however it remains to to recover T from T^t. For general matrices, it is clear that matching a power of a matrix does not imply the matrix itself is matched. For our case, even though requiring T to be stochastic adds additional constraints, matching the matrix power still does not suffice to identify the underlying matrix, as formalized in the following claim.

Claim 5 (Nonidentifiability of matrix powers (Claim 1 restated))**.** *For any positive integer t, there exist stochastic matrices $T, \tilde{T}$ satisfying Assumption 1, 3, such that $T \neq \tilde{T}$ and $T^t = \tilde{T}^t$.*

Proof. As in Theorem 4, the nonidentifiability comes from the non-uniqueness of matrix factorization. Specifically for this case, we will set $\tilde{T}$ to be equal to T up to a special rotation that gets composed when taking the matrix power. That is, we want $\tilde{T} = RT = TR$ for some matrix R that implicitly performs a rotation, so that $\tilde{T}^t = T^t R^t$. Since R corresonds to a rotation, we can choose the rotation angle properly so that $R^t = I$, and hence $\tilde{T}^t = T^t$ but $\tilde{T} \neq T$.

Precisely, using notations for the G-HMM setup, set $a \in [0, 1]$, and let the parameters (T, M) be given by

$$T = \begin{bmatrix} a & 0 & 1-a \\ 1-a & a & 0 \\ 0 & 1-a & a \end{bmatrix}, \; M = \begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & -\sqrt{3}/2 & \sqrt{3}/2 \\ 1/\sqrt{2} & 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

Let θ be some rotation angle, and denote by $R(\theta) := \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$ a rotation that acts on the first two dimensions. We will show that for any $\theta \in \mathbb{R}$, we have

$$\tilde{T} := \big(M^{-1}\big(R(\theta)\big)^{-1}M\big) \cdot T = T \cdot \big(M^{-1}\big(R(\theta)\big)^{-1}M\big). \tag{37}$$

Assuming equation 37, since $R(\theta)$ represents a rotation of angle θ, $\big(R(\theta)\big)^{\tau}$ corresponds to a rotation of angle $\tau\theta$ for any integer τ (τ could be negative). Setting $\theta := \frac{2\pi}{t}$, we then have

$$\begin{aligned} \tilde{T}^t =& \big(M^{-1}\big(R(\theta)\big)^{-1}M \cdot T\big)^t = T^t\big(M^{-1}\big(R(\theta)\big)^{-1}M\big)\big)^t = T^t M^{-1}\big(R(\theta)\big)^{-t}M \\ =& T^t M^{-1} \cdot R(2\pi) \cdot M = T^t. \end{aligned} \tag{38}$$

For $\tilde{T}$ to serve as a valid example for our theorem, it remains to check that for every t, there exists a choice of a such that $\tilde{T} := RT$, where $R := M^{-1}\big(R(\frac{2\pi}{t})\big)^{-1}M$, is a valid stochastic matrix. That is, $\tilde{T}$ has 1) columns and rows each summing up to 1, and 2) entries bounded in $[0, 1]$. Let's first show that the columns and rows each sum up to 1. Noting that $M^{-1} = \frac{1}{3}\begin{bmatrix} 2 & 0 & \sqrt{2} \\ -1 & -\sqrt{3} & \sqrt{2} \\ -1 & \sqrt{3} & \sqrt{2} \end{bmatrix}$, the column sums are

$$\mathbf{1}^\top \tilde{T} = \mathbf{1}^\top M^{-1}R(\theta)^{-1}MT \overset{(i)}{=} \mathbf{1}^\top T M^{-1}R(\theta)^{-1}M = \sqrt{2}e_3^\top R(\theta)M = \sqrt{2}e_3^\top M = \sqrt{2}\frac{1}{\sqrt{2}}\mathbf{1} = \mathbf{1}, \tag{39}$$

where step (i) uses equation 37. Similarly, the row sums are

$$\tilde{T}\mathbf{1} = M^{-1}R(\theta)^{-1}M\mathbf{1} = M^{-1}R(\theta)^{-1} \cdot \frac{3}{\sqrt{2}}e_3 = M^{-1} \cdot \frac{3}{\sqrt{2}}e_3 = \mathbf{1}. \tag{40}$$

To show that there exists a choice of T such that entries of $\tilde{T}$ are non-negative, we provide a concrete example where T is defined with $a = \frac{1}{2}$. It can be checked that $\tilde{T} := M^{-1}(R(\frac{2\pi}{t}))^{-1}M$ has non-negative entries for $t \in \{2, 3, 4, ..., 10\}$. For larger t, let $\theta = \frac{2\pi}{t}$, then we have by the Taylor expansion of $R(\frac{2\pi}{t})$:

$$\begin{aligned} R(\theta) :=& \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 - \theta^2/2 + c_1\theta^4 & -\theta + c_2\theta^2 & 0 \\ \theta + c_2\theta^2 & 1 - \theta^2/2 + c_1\theta^4 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ =& I + \theta \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \theta^2 \begin{bmatrix} -1/2 + c_1\theta^2 & c_2 & 0 \\ c_2 & -1/2 + c_1\theta^2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{aligned} \tag{41}$$

for some constants $c_1 \in [-\frac{1}{4!}, \frac{1}{4!}]$, $c_2 \in [-\frac{1}{2}, \frac{1}{2}]$. Substituting this into $\tilde{T} := M^{-1}R(\theta)^{-1}M$ gives

$$\begin{aligned}
\tilde{T} =& \begin{bmatrix} a - \frac{\theta}{\sqrt{3}}(1-a) & \frac{\theta}{\sqrt{3}}(1-2a) & 1-a+\frac{\theta}{\sqrt{3}}a \\ 1-a+\frac{\theta}{\sqrt{3}}a & a-\frac{\theta}{\sqrt{3}}(1-a) & \frac{\theta}{\sqrt{3}}(1-2a) \\ \frac{\theta}{\sqrt{3}}(1-2a) & 1-a+\frac{\theta}{\sqrt{3}}a & a-\frac{\theta}{\sqrt{3}}(1-a) \end{bmatrix} \\
&+ \frac{1}{3}\begin{bmatrix} -1+2c_1\theta^2 & \frac{1}{2}-c_1\theta^2-\sqrt{3}c_2 & \frac{1}{2}-c_1\theta^2+\sqrt{3}c_2 \\ \frac{1}{2}-c_1\theta^2-\sqrt{3}c_2 & -1+2c_1\theta^2+\sqrt{3}c_2 & \frac{1}{2}-c_1\theta^2 \\ \frac{1}{2}-c_1\theta^2+\sqrt{3}c_2 & \frac{1}{2}-c_1\theta^2 & -1+2c_1\theta^2-\sqrt{3}c_2 \end{bmatrix} \cdot \begin{bmatrix} a & 0 & 1-a \\ 1-a & a & 0 \\ 0 & 1-a & a \end{bmatrix} \\
=& \frac{1}{2}\begin{bmatrix} 1-\frac{\theta}{\sqrt{3}} & 0 & 1+\frac{\theta}{\sqrt{3}} \\ 1+\frac{\theta}{\sqrt{3}} & 1-\frac{\theta}{\sqrt{3}} & 0 \\ 0 & 1+\frac{\theta}{\sqrt{3}} & 1-\frac{\theta}{\sqrt{3}} \end{bmatrix} + \frac{\theta^2}{6}\begin{bmatrix} -\frac{1}{2}+c_1\theta^2-\sqrt{3}c_2 & 1-2c_1\theta^2 & -\frac{1}{2}+c_1\theta^2+\sqrt{3}c_2 \\ -\frac{1}{2}+c_1\theta^2 & -\frac{1}{2}+c_1\theta^2+\sqrt{3}c_2 & 1-2c_1\theta^2-\sqrt{3}c_2 \\ 1-2c_1\theta^2+\sqrt{3}c_2 & -\frac{1}{2}+c_1\theta^2-\sqrt{3}c_2 & -\frac{1}{2}+c_1\theta^2 \end{bmatrix} \\
\overset{(i)}{\geq}& \frac{1}{2}\begin{bmatrix} 1-\frac{\theta}{\sqrt{3}} & 0 & 1+\frac{\theta}{\sqrt{3}} \\ 1+\frac{\theta}{\sqrt{3}} & 1-\frac{\theta}{\sqrt{3}} & 0 \\ 0 & 1+\frac{\theta}{\sqrt{3}} & 1-\frac{\theta}{\sqrt{3}} \end{bmatrix} + \theta^2\begin{bmatrix} -0.25 & -0.16 & -0.25 \\ -0.09 & -0.25 & 0.01 \\ 0.01 & -0.25 & -0.09 \end{bmatrix}
\end{aligned} \tag{42}$$

where the inequality (i) is taken entry-wise. It can be checked that all entries are non-negative for $\theta \leq \frac{2\pi}{10}$.

Proof of equation 37 Let's conclude the proof by proving the commutativity in equation 37. Denote $R_2(\theta) := \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$, i.e. $R(\theta) = \begin{bmatrix} R_2(\theta) & 0 \\ 0 & 1 \end{bmatrix}$. Denote $U := \begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & -\sqrt{3}/2 & \sqrt{3}/2 \end{bmatrix}$, i.e. $M = \begin{bmatrix} U \\ \mathbf{1}^\top/\sqrt{2} \end{bmatrix}$. We can write

$$M^\top R(\theta)^\top M = \begin{bmatrix} U^\top & \mathbf{1}/\sqrt{2} \end{bmatrix} \begin{bmatrix} R_2(\theta)^\top & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} U \\ \mathbf{1}^\top/\sqrt{2} \end{bmatrix} = U^\top R_2(\theta)^\top U + \frac{\mathbf{1}\mathbf{1}^\top}{2}. \tag{43}$$

Let $R_2(\theta)$ denote a clockwise rotation of angle θ, then

$$U = [v_1, R_2(\frac{2\pi}{3})v_1, R_2(\frac{4\pi}{3})v_1] = [R_2(\frac{4\pi}{3})v_2, v_2, R_2(\frac{2\pi}{3})v_2] = [R_2(\frac{2\pi}{3})v_3, R_2(\frac{4\pi}{3})v_3, v_3], \tag{44}$$

where $v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $v_2 = \begin{bmatrix} -1/2 \\ -\sqrt{3}/2 \end{bmatrix}$, $v_3 = \begin{bmatrix} -1/2 \\ \sqrt{3}/2 \end{bmatrix}$. Denote $\alpha_{ij} := v_i^\top R_2^\top v_j$ for $i, j \in [3]$. Noting $T = aI + (1-a)\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} := aI + (1-a)P$, we have

$$\begin{aligned}
& M^\top R(\theta)^\top M T = T M^\top R(\theta)^\top M \\
\Leftrightarrow\ & U^\top R_2(\theta)^\top U(aI + (1-a)P) + \frac{\mathbf{1}\mathbf{1}^\top}{2}T = (aI + (1-a)P)U^\top R_2(\theta)^\top U + T\frac{\mathbf{1}\mathbf{1}^\top}{2} \\
\overset{(i)}{\Leftrightarrow}\ & U^\top R_2(\theta)^\top U P = P U^\top R_2(\theta)^\top U \\
\Leftrightarrow\ & \begin{bmatrix} \alpha_{31} & \alpha_{32} & \alpha_{33} \\ \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \end{bmatrix} \overset{(*)}{=} \begin{bmatrix} \alpha_{12} & \alpha_{13} & \alpha_{11} \\ \alpha_{22} & \alpha_{23} & \alpha_{21} \\ \alpha_{32} & \alpha_{33} & \alpha_{31} \end{bmatrix}.
\end{aligned} \tag{45}$$

where step (i) uses $\mathbf{1}\mathbf{1}^\top T = T\mathbf{1}\mathbf{1}^\top = \mathbf{1}\mathbf{1}^\top$. The equality $(*)$ is true due to equation 44. □

D Simulation

We empirically verify the identifiability results for HMM (Theorem 5) and G-HMM (Theorem 3) on simulation data, by checking whether matching the optimal predictor implies matching the parameters of the data generative model.

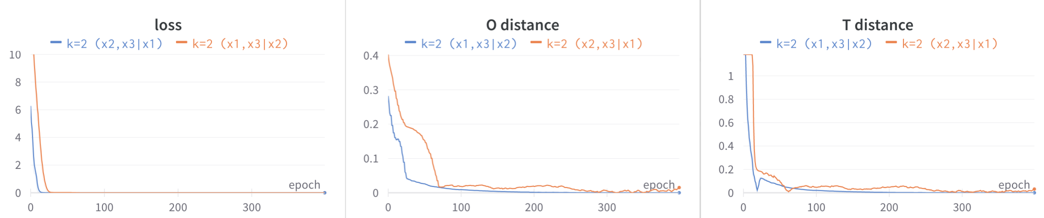

Figure 1: HMM: left: objective; middle: $\|O - O^*\|_F$; right: $\|T - T^*\|_F$.

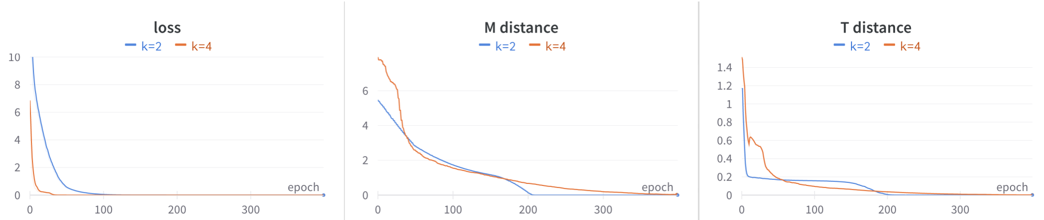

Figure 2: G-HMM: left: objective; middle: $\|M - M^*\|_F$; right: $\|T - T^*\|_F$.

In particular, we aim to recover $O \in \mathbb{R}^{d\times k}, T \in \mathbb{R}^{k\times k}$ for HMM, and $M \in \mathbb{R}^{d\times k}, T \in \mathbb{R}^{k\times k}$ for G-HMM, with $d = 10$ and $k \in \{2, 4\}$. Given a batch of samples $\mathcal{B} := \{x_1^{(i)}\}_{i\in|\mathcal{B}|}$, the objective to minimize for HMM is

$$\ell(O, T) := \frac{1}{|\mathcal{B}|} \sum_{x_1 \in \mathcal{B}} \|f^{2,3|1}_{O^*,T^*}(x_1) - f^{2,3|1}_{O,T}(x_1)\|_F^2, \tag{46}$$

where $f^{2,3|1}_{O,T}(x_1) := \mathbb{E}_{O,T}[x_2 \otimes x_3 | x_1]$; the objective for the task of predicting $x_1 \otimes x_3$ given x_2 is defined analogously. Note that though equation 46 differs from equation 1 by a constant, [8] it suffices for verifying parameter identifiability since both losses are minimized at $f_{O,T} = f_{O^*,T^*}$. We choose to use the form in equation 46 since it is more stable to optimize for and that its minimal loss value is 0, making it easy to check for optimality. Similarly, the objective to minimizer for G-HMM is

$$\ell(M, T) := \frac{1}{|\mathcal{B}|} \sum_{x_1 \in \mathcal{B}} \|f^{2|1}_{M^*,T^*}(x_1) - f^{2|1}_{M,T}(x_1)\|_2^2, \tag{47}$$

where $f^{2|1}_{M,T} := \mathbb{E}_{M,T}[x_2|x_1]$.

For both HMM and G-HMM, we optimize for O (or M) and T alternatingly in different epochs. We found that it is usually helpful to use a larger learning rate for T than for O (or M), and that normalized gradient descent helps speed up training [9]

Figure 1 and 2 show the results for HMM and G-HMM. It can be seen that as the objective value approaches the optimum, the parameter distances indeed go to zero, corroborating Theorem 5 and 3.

[8] In equation 1, the population loss is defined as $\mathbb{E}_{x_1}\mathbb{E}_{x_2,x_3|x_1}\|x_2 \otimes x_3 - f(x_1)\|_F^2$, whereas for equation 46 the population loss is $\mathbb{E}_{x_1}\|\mathbb{E}_{x_2,x_3|x_1}[x_2 \otimes x_3] - f(x_1)\|_F^2$, i.e. the expectation over x_2, x_3 is moved to within the Frobenius norm. The two losses differ by a constant that is independent of the parameters of f.

[9] That is, we normalize each gradient to have Frobenius norm 1. The gradients are otherwise too small which will result in slow convergence.