A Proof of Theorem 1

Theorem 1 (restated). *Assume Θ is a non-empty open connected subset of $[0,1]^n$ and $\mu\colon \mathbb{R}^n \to \mathbb{R}^m$ is a polynomial map. With probability 1, the following holds.*

- CHECKIDENTIFIABILITY *returns "no" $\Rightarrow$ for almost all $\theta_0 \in \Theta$ and any open neighborhood $N(\theta_0)$ around θ_0, $|\mathcal{S}_\Theta(\theta_0) \cap N(\theta_0)|$ is infinite (not locally identifiable).*
- CHECKIDENTIFIABILITY *returns "yes" $\Rightarrow$ (i) for almost all $\theta_0 \in \Theta$, there exists an open neighborhood $N(\theta_0)$ around θ_0 such that $|\mathcal{S}_\Theta(\theta_0) \cap N(\theta_0)| = 1$ (locally identifiable); and (ii) there exists a set $\mathcal{E} \subset \Theta$ with measure zero such that $|\mathcal{S}_{\Theta \setminus \mathcal{E}}(\theta_0)|$ is finite for every $\theta_0 \in \Theta \setminus \mathcal{E}$ (identifiability of $\Theta \setminus \mathcal{E}$).*

The proof of Theorem 1 crucially relies on the following lemma from [16] which holds even in the case that μ is merely an analytic function (see Lemma 9 of [17] for a simpler proof in the case μ is a polynomial map); it states that the Jacobian achieves its maximal rank almost everywhere in Θ. To state this precisely, first define $r_{\max} \overset{\text{def}}{=} \max\{\operatorname{rank}(J(\theta)) : \theta \in \Theta\}$ and $\Theta_{\max} \overset{\text{def}}{=} \{\theta \in \Theta : \operatorname{rank}(J(\theta)) = r_{\max}\}$.

Lemma 2. *The set $\Theta \setminus \Theta_{\max}$ has Lebesgue measure zero. That is, $\Theta_{\max}$ is almost all of Θ.*

Proof of Theorem 1. By Lemma 2, CHECKIDENTIFIABILITY chooses a point $\tilde{\theta} \in \Theta_{\max}$ with probability 1. We henceforth condition on this event, so $\operatorname{rank}(J(\tilde{\theta})) = r_{\max}$.

Case 1: $\operatorname{rank}(J(\tilde{\theta})) < n$ (*i.e.*, "no" is returned). In this case, we have $r_{\max} < n$. We now employ an argument from the proof of Proposition 20 of [16]. Fix any $\theta_0 \in \Theta_{\max}$. Since Θ is open, Weyl's theorem implies that there is an open neighborhood U around θ_0 in Θ on which $\operatorname{rank}(J(\theta)) = r_{\max}$ for all $\theta \in U$ (*i.e.*, $\operatorname{rank}(J(\cdot))$ is constant on U). Therefore, by the constant rank theorem, there is an open neighborhood $N(\theta_0)$ around θ_0 in Θ such that $\mu^{-1}(\mu(\theta_0)) \cap N(\theta_0)$ is homeomorphic with an open set in $\mathbb{R}^{n - r_{\max}}$. Therefore $\mathcal{S}_\Theta(\theta_0) \cap N(\theta_0)$ is uncountably infinite.

Case 2: $\operatorname{rank}(J(\tilde{\theta})) = n$ (*i.e.*, "yes" is returned). In this case, we have $r_{\max} = n$. Therefore for every $\theta_0 \in \Theta_{\max}$, the Jacobian $J(\theta_0)$ has full column rank, and thus by the inverse function theorem, μ is injective on a neighborhood of θ_0. This in turn implies that for all $\theta_0 \in \Theta_{\max}$, there exists an open neighborhood $N(\theta_0)$ around θ_0 such that $\mathcal{S}_\Theta(\theta_0) \cap N(\theta_0) = \{\theta_0\}$. This proves (i).

To show (ii), define $\mathcal{E} \overset{\text{def}}{=} \Theta \setminus \Theta_{\max}$, and now claim that for every $\theta_0 \in \Theta_{\max}$, the equivalence class $\mathcal{S}_{\Theta_{\max}}(\theta_0)$ is finite. Observe that by (i), the set $\mathcal{S}_{\Theta_{\max}}(\theta_0)$ contains only geometrically isolated solutions to the system of polynomial equations given by $\mu(\theta) = \mu(\theta_0)$. Therefore the claim follows immediately from Bézout's Theorem, which implies that the number of geometrically isolated solutions is finite. $\square$

Remark. All the models considered in this paper have moments μ which correspond to a polynomial map. However, for some models (e.g., exponential families), μ will not be a polynomial map, but rather, a general analytic function. In this case, Theorem 1 holds with one modification to (ii). If CHECKIDENTIFIABILITY returns "yes", then we have the following weaker guarantee in place of (ii): $\mathcal{S}_{\Theta_{\max}}(\theta_0)$ is *countable* (but not necessarily finite) for all $\theta_0 \in \Theta_{\max}$. The above proof does not require the fact that μ is a polynomial map except in the invocation of Bézout's Theorem. In place of Bézout's Theorem, we use the following argument. If $\mathcal{S}_{\Theta_{\max}}(\theta_0)$ is uncountable, then it contains a limit point $\theta^* \in \mathcal{S}_{\Theta_{\max}}(\theta_0)$; thus for any small enough neighborhood $N(\theta^*)$ of θ^*, there is some $\theta \in \mathcal{S}_{\Theta_{\max}}(\theta_0) \cap N(\theta^*)$. This contradicts (i) as applied to θ^*, and thus we conclude that $\mathcal{S}_{\Theta_{\max}}(\theta_0)$ is countable.

B Additional results from the identifiability checker

PCFG models with $d < k$. The PCFG models that we've considered so far assume that the number of words d is at least the number of hidden states k, which is a realistic assumption for natural language. However, there are applications, e.g., computational biology, where the vocabulary size d is relatively small. In this regime, identifiability becomes trickier because the data doesn't reveal as much about the hidden states, and brings us closer to the boundary between identifiability and non-identifiability. In this section, we consider the $d < k$ regime.

The following table gives additional identifiability results from CHECKIDENTIFIABILITY for values of d, k, and L where $d < k$ (recall that the results reported in Section 4.4 only considered values where $d \geq k$). In each cell, we show the (k, d, L) values for which CHECKIDENTIFIABILITY returned "yes"; the values checked were $k \in \{3, 4, \ldots, 8\}$, $d \in \{2, \ldots, k-1\}$, $L \in \{3, 4, \ldots, 9\}$.

	ϕ_{12}	ϕ_{**}	ϕ_{123e_1}	ϕ_{123}	ϕ_{***e_1}	ϕ_{***}
PCFG	None					
PCFG-I	None	$(3,2,\geq 6)$ $(4,2,\geq 8)$ $(4,3,\geq 5)$ $(5,3,\geq 6)$ $(5,4,\geq 4)$ $(6,3,\geq 7)$ $(6,4,\geq 5)$ $(6,5,\geq 4)$ $(7,3,\geq 8)$ $(7,4,\geq 6)$ $(7,5,\geq 5)$ $(7,6,\geq 4)$	None	$(5,4,\geq 4)$ $(6,5,\geq 4)$ $(7,5,\geq 4)$ $(7,6,\geq 4)$	$(3,2,\geq 5)$ $(4,2,\geq 6)$ $(4,3,\geq 4)$ $(5,2,\geq 7)$ $(5,\geq 3,\geq 4)$ $(6,2,\geq 8)$ $(6,3,\geq 5)$ $(6,\geq 4,\geq 4)$ $(7,2,\geq 9)$ $(7,3,\geq 5)$ $(7,\geq 4,\geq 4)$	
PCFG-IE	None	$(3,2,\geq 6)$ $(4,2,\geq 8)$ $(4,3,\geq 5)$ $(5,3,\geq 6)$ $(5,4,\geq 5)$ $(6,3,\geq 7)$ $(6,4,\geq 5)$ $(6,5,\geq 4)$ $(7,3,\geq 8)$ $(7,4,\geq 6)$ $(7,5,\geq 5)$ $(7,6,\geq 4)$	$(5,4,\geq 4)$ $(6,5,\geq 4)$ $(7,5,\geq 5)$ $(7,6,\geq 4)$	$(4,3,\geq 4)$ $(5,4,\geq 4)$ $(6,\geq 4,\geq 4)$ $(7,\geq 5,\geq 4)$	$(3,2,\geq 5)$ $(4,2,\geq 6)$ $(4,3,\geq 4)$ $(5,2,\geq 7)$ $(5,3,\geq 5)$ $(5,4,\geq 4)$ $(6,2,\geq 8)$ $(6,3,\geq 5)$ $(6,\geq 4,\geq 4)$ $(7,2,\geq 9)$ $(7,3,\geq 5)$ $(7,\geq 4,\geq 4)$	$(3,2,\geq 5)$ $(4,2,\geq 6)$ $(4,3,\geq 4)$ $(5,2,\geq 7)$ $(5,\geq 3,\geq 4)$ $(6,2,\geq 8)$ $(6,3,\geq 5)$ $(6,\geq 4,\geq 4)$ $(7,2,\geq 9)$ $(7,3,\geq 5)$ $(7,\geq 4,\geq 4)$

Fixed topology models. We now present some results for latent class models (LCMs) and hidden Markov models (HMMs). While identifiability for these models are more developed than for parsing models, we show that the identifiability checker can refine the results even for the classic models.

The parameters of an HMM are $\theta = (\pi, T, O)$, where $\pi \in \mathbb{R}^k$ specifies the initial state distribution, $T \in \mathbb{R}^{k \times k}$ specifies the state transition probabilities, and $O \in \mathbb{R}^{d \times k}$ specifies the emission distributions. The probability over a sentence $\mathbf{x}$ is:

$$\mathbb{P}_\theta(\mathbf{x}) = \mathbf{1}^\top T \operatorname{diag}(O^\top x_L) \cdots T \operatorname{diag}(O^\top x_2) T \operatorname{diag}(O^\top x_1) \pi. \tag{5}$$

The parameters of an LCM are $\theta = (\pi, O)$—the same as that of an HMM except with $T \equiv I$. The probability over a sentence $\mathbf{x}$ is also given by (5) (with $T = I$).

The following table summarizes some identifiability results obtained by CHECKIDENTIFIABILITY (for $d \geq k$); these results have all been proven analytically in previous work (e.g., [8, 10, 11, 20, 21]) except for the identifiability of HMMs from ϕ_{**}.

	ϕ_{12}	ϕ_{**}	ϕ_{123e_1}	ϕ_{123}	ϕ_{***e_1}	ϕ_{***}
LCM	No		Yes iff $L \geq 3$			
HMM	No	Yes iff $L \geq 3$				

It is known that LCMs are not identifiable from ϕ_{**} for any value of L [8]. However, LCMs constitute a subfamily of HMMs arising from a measure zero subset of the HMM parameter space. Therefore the identifiability of HMMs from ϕ_{**} (for $L \geq 3$) does not contradict this result. The result does not appear to be covered by application of Kruskal's theorem in previous work [11], so we prove the result rigorously below.

It can be checked using (5) that

$$\mathbb{E}_\theta[\phi_{12}(\mathbf{x})] = O\,\mathrm{diag}(\pi)T^\top O^\top$$
$$\mathbb{E}_\theta[\phi_{34}(\mathbf{x})] = O\,\mathrm{diag}(T\pi)T^\top O^\top.$$

Let $M_1 \stackrel{\mathrm{def}}{=} O$, $M_2 \stackrel{\mathrm{def}}{=} OT\,\mathrm{diag}(\pi)$, and $D \stackrel{\mathrm{def}}{=} \mathrm{diag}(T\pi)\,\mathrm{diag}(\pi)^{-1}$. Provided that

1. $\pi > 0$,
2. O has full column rank,
3. T is invertible,
4. the ratios of probabilities $(T\pi)_i/\pi_i$, ranging over $i \in [k]$, are distinct

(all of which are true for all but a measure zero set of parameters in Θ), the matrices M_1 and M_2 have full column rank and the diagonal matrix D has distinct diagonal entries. Therefore Lemma 1 can be applied with $X = \mathbb{E}_\theta[\phi_{12}(\mathbf{x})] = M_1 M_2^\top$ and $Y = \mathbb{E}_\theta[\phi_{34}(\mathbf{x})] = M_1 D M_2^\top$ to recover $M_1 = O$. It is easy to see that π and T can also easily be recovered.

Note that the fourth condition above, that $T\pi$ be entry-wise distinct from π, is violated when a LCM distribution is cast as an HMM distribution (by setting $T = I$ so $T\pi = \pi$). However, the set of HMM parameters satisfying this equation is a measure zero set.

Discussion. CHECKIDENTIFIABILITY tests for local identifiability. If it finds that a model family is not locally identifiable, then it is not globally identifiable. However the inverse claim is not necessarily true: if it finds that a model family is locally identifiable, it is not necessarily globally identifiable. Theorem 1 provides the somewhat weaker guarantee that a restricted model family is globally identifiable, where the equivalence classes $\mathcal{S}_{\Theta\setminus\mathcal{E}}(\theta_0)$ are only taken with respect to a subset $\Theta \setminus \mathcal{E} \subseteq \Theta$ of the parameter space. However, there is a gap between this property (which is with respect to $\Theta \setminus \mathcal{E}$) and true global identifiability (which is with respect to Θ).

On the other hand, having explicit estimators guarantees us proper global identifiability with respect to the original model family Θ. In fact, the exceptional set $\mathcal{E}$ can typically be characterized explicitly. For instance, in the case of PCFG-IE, the set $\Theta \setminus \mathcal{E}$ contains those $\theta = (\pi, T, O)$ that satisfy full rank conditions:

$$\Theta \setminus \mathcal{E} = \{(\pi, T, O) : \pi \succ 0, T \text{ is invertible}, O \text{ has full column rank}\}. \tag{6}$$

Additionally, the explicit estimators also provides an explicit characterization of the elements in the equivalence class $\mathcal{S}_\Theta(\theta_0)$ for each $\theta_0 \in \Theta \setminus \mathcal{E}$: the set $\mathcal{S}_\Theta(\theta_0)$ contains exactly $k!$ elements corresponding to permutation of the hidden states. Specifically,

$$\mathcal{S}_\Theta((\pi, T, O)) = \{(\Pi^{-1}\pi, \Pi^{-1}T\Pi, O\Pi) : \Pi \text{ is a permutation matrix}. \tag{7}$$

Note that this is shaper than Theorem 1, which only says that the equivalence classes have to be finite.

C Dynamic programs

For a sentence of length L, the number of parse trees is exponential in L. Therefore, dynamic programming is often employed to efficiently compute expectations over the parse trees, the core computation in the E-step of the EM algorithm. In the case of PCFG, this dynamic program is referred to as the CKY algorithm, which runs in $O(L^3k^3)$ time, where k is the number of hidden states. For simple dependency models, a $O(L^3)$ dynamic program was developed by [29]. At a high-level, the states of the dynamic program in both cases are the spans $[i:j]$ of the sentence (and for the PCFG, the these states include the hidden states $z_{[i:j]}$ of the nodes).

In this paper, we need to compute (i) the Jacobian matrix for checking identifiability (Section 4.2) and (ii) the mixing matrix for recovering compound parameters (Section 5.1). Both computations can be performed efficiently with a modified version of the classic dynamic programs, which we will describe in this section.

C.1 Computing the Jacobian matrix

Recall that the j-th row of the Jacobian matrix J is (the transpose of) the gradient of $h_j(\theta) = \mu_j(\theta) - \mu_j(\theta_0)$. Specifically, entry J_{ji} is the derivative of the j-th moment with respect to the i-th parameter:

$$\begin{aligned} J_{ji} &= \frac{\partial h_j(\theta)}{\partial \theta_i} && (8) \\ &= \frac{\partial \mathbb{E}_\theta[\phi_j(\mathbf{x})]}{\partial \theta_i} && (9) \\ &= \sum_{\mathbf{x},z} \frac{\partial p_\theta(\mathbf{x}, z)}{\partial \theta_i} \phi_j(\mathbf{x}). && (10) \end{aligned}$$

We can encode the sum over the exponential set of possible sentences $\mathbf{x}$ and parse trees z using a directed acyclic hypergraph so that each hyperpath through the hypergraph corresponds to a $(\mathbf{x}, z)$ pair. Specifically, a *hypergraph* consists of the following:

- a set of nodes $\mathcal{V}$ with a designated start node $\textsc{Start} \in \mathcal{V}$ and an end node $\textsc{End} \in \mathcal{V}$, and
- a set of hyperedges $\mathcal{E}$ where each hyperedge $e \in \mathcal{E}$ has a source node $e.a \in \mathcal{V}$ and a pair of target nodes $(e.b, e.c) \in \mathcal{V} \times \mathcal{V}$ (we say that e connects $e.a$ to $e.b$ and $e.c$) and an index $e.i \in [n]$ corresponding to a component of the parameter vector $\theta \in \mathbb{R}^n$.

Define a *hyperpath* P to be a subset of the edges $\mathcal{E}$ such that:

- $(\textsc{Start}, a, b) \in P$ for some $a, b \in \mathcal{V}$;
- if $(a, b, c) \in P$ and $b \neq \textsc{End}$, then $(b, d, e) \in P$ for some $d, e \in \mathcal{V}$; and
- if $(a, b, c) \in P$ and $c \neq \textsc{End}$, then $(c, d, e) \in P$ for some $d, e \in \mathcal{V}$.

Each hyperpath P, encoding $(\mathbf{x}, z)$, is associated with a probability equal to the product of all of the parameters on that hyperpath:

$$p_\theta(\mathbf{x}, z) = p_\theta(P) = \prod_{e \in P} \theta_{e.i}. \tag{11}$$

In this way, the hypergraph compactly defines a distribution over exponentially many hyperpaths.

Now, we assume that each moment $\phi_j(\mathbf{x})$ corresponds to a function $f_j : \mathcal{E} \mapsto \mathbb{R}$ mapping each hyperedge e to a real number so that the moment is equal to the product over function values:

$$\phi_j(\mathbf{x}) = \prod_{e \in P} f_j(e), \tag{12}$$

where P is any hyperpath that encodes the sentence $\mathbf{x}$ and some parse tree z (we assume that the product is the same no matter what z is).

Now, let us write out the Jacobian matrix entries in terms of hyperpaths:

$$J_{ji} = \sum_P \sum_{e_0 \in P} \frac{\partial \theta_{e_0.i}}{\partial \theta_i} \prod_{e \in P, e \neq e_0} \theta_{e.i} f_j(e). \tag{13}$$

The sum over hyperpaths P can be computed efficiently as follows. For each hypergraph node a, we compute an inside score $\alpha(a)$, which sums over all possible partial hyperpaths terminating at the target node, and an outside score $\beta(a)$, which sums over all possible partial hyperpaths from the source node:

$$\begin{aligned} \alpha(a) &\overset{\text{def}}{=} \sum_{e \in \mathcal{E} : e.a = a} \theta_{e.i} \alpha(e.b) \alpha(e.c), && (14) \\ \beta(a) &\overset{\text{def}}{=} \sum_{e \in \mathcal{E} : e.b = a} \theta_{e.i} \alpha(e.c) \beta(e.a) \sum_{e \in \mathcal{E} : e.c = a} \theta_{e.i} \alpha(e.b) \beta(e.a). && (15) \end{aligned}$$

The Jacobian entry J_{ji} can be computed as follows:

$$J_{ji} = \sum_{e \in \mathcal{E}} \beta(e.a) \alpha(e.b) \alpha(e.c) \mathbb{I}[i = e.i]. \tag{16}$$