Approximate Homomorphisms and Convergent Representations in Transducers

Santiago Cifuentes[1,2]

[1]Dovetail Research Group
[2]ICC CONICET, Universidad de Buenos Aires

August 2026

Abstract

We study the stability of minimal representations of controlled stochastic processes (in particular, *transducers*) under perturbations. This question is motivated by recent experiments finding predictive-state structure in the latent representations of neural networks. We consider standard, linear and predictive transducers. We introduce notions of approximate homomorphism capturing local structural similarity between them, together with metrics comparing their induced dynamics (which we refer to as *interfaces*), and prove properties such as composability of the approximate homomorphisms. For standard transducers, we show that there exist simple interfaces for which there is no approximate homomorphism between the different implementations of the dynamics. In contrast, for every finite-rank interface $\mathcal{I}$, we prove that all minimal linear transducers implementing interfaces sufficiently close to $\mathcal{I}$ have an approximate homomorphism to the minimal implementation of $\mathcal{I}$, with error linear in the perturbation size. We prove an analogous stability result for predictive transducers under a residual metric using some mild hypothesis regarding the indistinguishability of the belief states. These results identify conditions under which canonical transducer representations are robust to perturbations, while showing that such convergence fails without additional structural restrictions. Under the assumption that these type of abstractions are embedded into the hidden layers of modern AI models, this gives some theoretical support to the hypothesis that their latent representations exhibit structural convergence.

1 Introduction

Since the beginning of neural networks, many AI architectures have included some type of hidden or intermediate layers between the input and output gates in which the models can encode partial results of their computation. In these layers the models usually learn, through training, to represent latent variables and general information useful for their goals [4, 6, 31].

Although the training process is not deterministic and depends on things such as the training algorithm, the initial parameters and the choice of hyperparameters, it has been observed from the beginning of the deep learning revolution that some structure of these hidden layers coincides between different models [33, 37], even when they are implemented in different architectures. There are different ways of measuring this similarity, but for

most of them this "convergent phenomenon" can be found [28]. To name a few of these metrics, similarity can be measured by comparing the distribution of latent vectors inside each layer [9, 29], by comparing functional aspects of the different layers [28], or by finding a linear transformation able to transfer features from one model to another [2, 9, 13, 18]. These experimental results have motivated the recent proposal of the "Platonic Representation Hypothesis" [23]: the idea that "Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces". Although it is unclear to what extent this hypothesis may hold [10, 15, 19], most results suggest that some kind of convergence can sometimes be found in different models trained for a similar task.

Overall, there are three core hypotheses which can help to understand this situation [23]. First, the *simplicity bias* hypothesis states that deep AI architectures trained through stochastic gradient descent have a tendency to converge to structurally simple representations which usually generalize well [5, 25, 51]. Meanwhile, the *capacity hypothesis* states that as models include more parameters they encompass a larger set of possible behaviours, and thus it is more likely for different architectures to have a non-empty intersection regarding the instantiations they allow [23]. Finally, the *multitask hypothesis* states that there are fewer representations capable of performing multiple tasks at the same time [8, 39], and thus as we train models for more complex goals the optimal configurations become sparser.

There exists an ongoing theoretical program trying to give support to these hypotheses. For instance, ideas such as implicit regularization [22] or neural collapse [24], or entire frameworks such as singular learning theory [52] try to explain how modern gradient descent finds robust representations in current deep learning architectures although the number of model parameters allows for overfitting. A different approach relies on the idea of a *world model* [20, 21, 44]. More precisely, different empirical and theoretical results support the idea that modern agents develop an internal mechanism equivalent to a description of the dynamics of the environment surrounding them [11, 38, 43, 46]. If we formalize these structures using some mathematical abstraction, then we can ask the question of whether the set of these abstractions implementing the same dynamics has some shared structure [44]. If the answer is positive, this provides some support for the idea that agents learning from a similar data distribution should share some aspects of their internal representations. See Figure 1 for a diagrammatic sketch of this idea.

In this work we investigate this idea using *transducers* as an abstraction of world models. A transducer is a controlled stochastic system with hidden states, inputs, and outputs, which induces an *interface*: for every finite sequence of interventions, the interface specifies a probability distribution over the corresponding sequence of observations in the real world. These types of structures have been studied recently as tools to formalize world models [7, 44], and different experimental results support the idea that modern AI agents implement this type of structure in their residual stream [46, 47]. Moreover, in [44] it was proven that in some situations the set of transducers implementing a specified behaviour has a unique minimal implementation such that all other implementations can be structurally mapped into the minimal one. From our perspective, this is a positive result encouraging the possibility of convergent structure.

Our goal is to improve this type of result by weakening some of its hypotheses. More precisely, we aim to improve the result by making it robust to noise and approximation. Consider that we have two transducers implementing a *similar* behaviour (measured through some proper metric). Then, is it the case that they share some structure, measured through some other metric? Note that in practical scenarios we expect different models to learn from

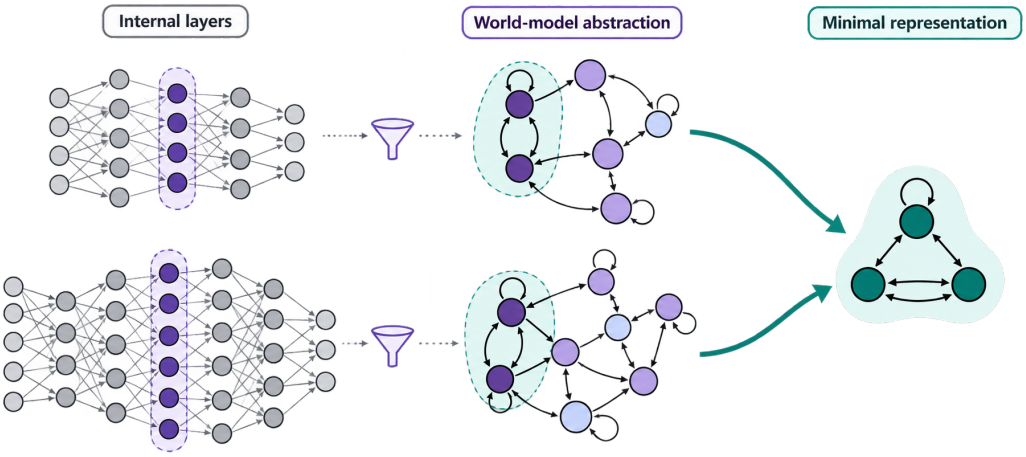


Figure 1: The world model-based approach for understanding convergent structure. We expect that each internal layer of a neural network architecture encodes a world model through some mathematical abstraction (such as a transducer). Then, we search for convergent structure in these abstractions by looking for a minimal model representing the dynamics.

slightly different datasets, and thus we need the convergence of transducers to hold also in the case in which they implement slightly different interfaces.

Our Contributions. We consider three types of transducers: the "standard" ones, linear transducers and predictive transducers, and do the following:

1. We give a notion of approximate homomorphism for all types of transducers that allows to decide when a transducer is ε-similar to another one. Our notion is robust with respect to composition and preserves the dynamics that the transducers represent up to an error that scales as $O(\varepsilon)$ under discounted metrics (i.e. metrics that weight differences in long-horizon predictions using a discount factor that decays exponentially with the number of steps).

2. In the context of standard transducers, we show that there exist interfaces such that the standard transducer implementing these interfaces are structurally far away (using our notion of approximate homomorphism to measure distance). This result shows that the result from [44] regarding the non-existence of a minimal representation for normal transducers cannot be salvaged by introducing an error term in the homomorphism.

3. For linear transducers, we show that for a significant subset of interfaces it is the case that all minimal linear transducers implementing an interface ε-similar (for a sufficiently small $\varepsilon > 0$) to interface $\mathcal{I}$ can be mapped to a common linear transducer introducing some error that scales as $\Gamma_{\mathcal{I}}\varepsilon$, where $\Gamma_{\mathcal{I}}$ represents a constant depending only on $\mathcal{I}$. This shows that the minimality of linear implementations is robust to noise in a small neighbourhood of the interface. This result is obtained by inspecting the canonical construction of the minimal linear transducer, which is obtained by working with the *Hankel matrix* of the interface.

4. Finally, we obtain an analogous result for predictive transducers by leveraging the construction of the minimal predictive implementation based on the notion of a ϵ-machine from computational mechanics [3].

Taken together, these results show that the existence of minimal representations (and thus of potential convergent structure) is robust to error for the family of linear and predictive transducers. Thus, we provide more theoretical support for the Platonic Representation Hypothesis under the hypothesis that world models show up inside the internal structure of modern AI architectures in the form of these types of abstractions. In the case of predictive transducers, this was partially observed empirically [46, 47]. Meanwhile, although there is no previous experiment finding linear transducers within the hidden layers of modern models, we suspect that these architectures should leverage the fact that any representation they contain is embedded in a linear space, and thus it is natural for them to prefer linear transducers over predictive ones. We believe that a fruitful direction for future work would be to reproduce the set-up of [47] but looking for a minimal linear transducer (or a functionally equivalent mechanism) inside the residual stream.

The remainder of the paper is organized as follows. Section 2 reviews standard, linear, and predictive transducers together with their exact notions of reduction and minimality. Section 3 introduces approximate homomorphisms, metrics on interfaces, and the composability and continuity results. Section 4 studies approximate common minima for nearby interfaces, giving respectively the negative result for unrestricted transducers and the positive results for linear and predictive transducers. Finally, in Section 5 we conclude the paper by summarizing our results, discussing their limitations and describing future lines of research. All proofs are deferred to the Appendix to improve readability.

Related work. Finite-state transducers have a long history in automata theory, beginning with the Mealy and Moore machines [34, 36]. In their deterministic form, they describe systems whose internal state is updated in response to an input while producing an output. Weighted and probabilistic variants replace deterministic transitions by numerical weights or stochastic kernels, and have been extensively studied in formal language theory [35]. The stochastic transducers considered here are closely related to controlled Markov models, and in particular they resemble Markov Decision Processes (MDPs) [40].

To identify convergent structure we use the notion of homomorphism, which corresponds to a map from one transducer to another that preserves local structure. This type of "coarse-graining" operations have a long history in the different abstractions we mentioned before. For Markov chains, classical lumpability identifies states whose transition probabilities agree after aggregation [26]. Probabilistic bisimulation gives a related behavioural equivalence for labelled probabilistic transition systems [30]. In the MDP literature, [17] develops exact state equivalences and model minimization based on bisimulation, while [41] formulates MDP and semi-MDP homomorphisms as maps that preserve rewards and aggregate transition probabilities. These notions have subsequently been organized into broader taxonomies of state abstraction [32].

We will study notions of homomorphism that allow for some error, and thus we refer to them as approximate homomorphisms. In the context of MDPs, such approximate reductions have already been considered [1, 42, 50], and our definitions as well as our robustness results (regarding composition and preservation of the interface up to discounted metrics) have analogues in the literature. In probabilistic transition systems, the notion of approximate bisimulation has a long history [14, 16] and remains an active area of research [27, 49].

The notion of the Hankel matrix of a process was introduced in the context of weighted automata to construct minimal linear implementations [45]. In that setting, the minimal realizations obtained are unique up to an invertible linear change of coordinates (i.e. a base change). Regarding predictive transducers, we use tools from computational mechanics [12, 48] to obtain minimal representations. In particular, the extension of computational mechanics to input-output processes [3] can be applied almost directly to our context.

2 Types of transducers

In this section we describe the different types of transducers that we will consider in this paper.

2.1 "Standard" Transducers

We use *transducers* to model world models.

Definition 1. *A **transducer** is given by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \kappa, p)$ where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions (or inputs), $\mathcal{O}$ is a set of reactions (or outputs), κ is a Markov kernel[1] of the form $\{\kappa_\tau(s', o|s, a) : s, s' \in \mathcal{S}, a \in \mathcal{A}, o \in \mathcal{O}, \tau \in \mathbb{N}\}$ and $p \in \Delta(\mathcal{S})$ is an initial distribution over the set of states.*

Transducers represent a world model indicating, for every possible sequence of actions $a_1 \dots a_n$, a distribution on the reaction $o_1 \dots o_n$ of the environment. They use (hidden) states to keep track of the previous events, and the function κ describes the relation between actions and outputs and how the state is updated. This function can depend on the timestep $\tau \in \mathbb{N}$, but for simplicity we will assume that $\kappa_\tau = \kappa_0$ for every τ (this corresponds to assuming *stationary* dynamics). Also, we may omit κ in the notation and simply talk about the probabilities of the events. For example, we write $\Pr_T(o|s, a)$ to denote the value $\sum_{s' \in \mathcal{S}} \kappa(s', o|s, a)$, or similarly $\Pr_T(s'|s, a, o)$ to denote $\kappa(s', o|s, a) / \Pr_T(o|s, a)$ whenever the denominator is positive. We will assume for simplicity that $\mathcal{S}, \mathcal{A}$ and $\mathcal{O}$ are countable.

A *trace* over T is a finite or infinite sequence of outputs. As mentioned, any transducer defines a probability for each trace *conditioned* on each sequence of actions. We refer to such a description (i.e. a list of probabilities $\Pr(o_1 \dots o_n|\mathbf{a})$ for every finite sequence $o_1 \dots o_n \in \mathcal{O}^n$ and infinite sequence $\mathbf{a} \in \mathcal{A}^\omega$) as an **interface** $\mathcal{I}$. We will only be interested in anticipation-free interfaces, i.e. those that satisfy $\Pr(o_1 \dots o_n|\mathbf{a}) = \Pr(o_1 \dots o_n|a_1 \dots a_n)$. Anticipation-free interfaces coincide exactly with interfaces "implementable" by transducers [44][Lemma 4].

More precisely, given $o_1 \dots o_n$ and $a_1 \dots a_n$ the (conditioned) probability that the transducer T induces can be computed as

$$\Pr_T(o_1 \dots o_n|a_1 \dots a_n) = \sum_{s_0 \dots s_n \in \mathcal{S}^{n+1}} p(s_0) \prod_{t=1}^{n} \kappa(s_t, o_t|s_{t-1}, a_t) \tag{1}$$

Whenever $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{O}| < \infty$ this computation can be simplified: if $M_{a,o} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is given by $M_{a,o}(s, s') = \kappa(s', o|s, a)$, then

$$\Pr_T(o_1 \dots o_n|a_1 \dots a_n) = p M_{a_1,o_1} \dots M_{a_n,o_n} \mathbf{1}$$

[1] In this context, a Markov kernel is simply a set of conditional distributions.

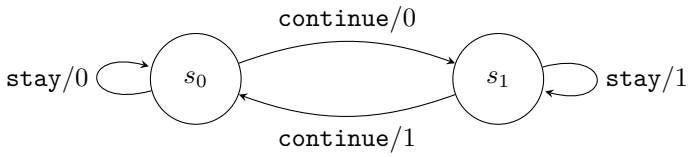


Figure 2: Example of a "deterministic" transducer. Each edge contains an action a followed by an output o. An edge from s to s' with label a/o indicates that $\kappa(s', o|s, a) = 1$.

Example 1. Figure 2 shows a transducer with deterministic dynamics (i.e. for every $s \in \mathcal{S}, a \in \mathcal{A}$ there is some $s' \in \mathcal{S}$ and $o \in \mathcal{O}$ such that $\kappa(s', o|s, a) = 1$). The states are $\mathcal{S} = \{s_0, s_1\}$, the actions $\mathcal{A} = \{\texttt{continue}, \texttt{stay}\}$, and the possible outputs $\mathcal{O} = \{0, 1\}$. The initial distribution is concentrated in state s_0. It represents a system that outputs $010101\ldots$ indefinitely as long as the action `continue` is chosen at each step. If `stay` is employed instead, the dynamics are "frozen" for one step.

We will assume that all the states of a transducer are reachable from some state with initial positive probability. Namely, for every state s there must exist a state s_0 such that $p(s_0) > 0$ and a sequence of actions $a_1 \ldots a_k$ such that $\Pr_T(s|s_0, a_1 \ldots a_k) > 0$. Also, we will sometimes use $\Sigma = \mathcal{A} \times \mathcal{O}$.

There is a well-defined notion of **homomorphism** for these objects, which allows us to coarse-grain states as well as input and output symbols.

Definition 2. *Given two transducers $T_1 = (\mathcal{S}_1, \mathcal{A}_1, \mathcal{O}_1, \kappa_1, p_1)$ and $T_2 = (\mathcal{S}_2, \mathcal{A}_2, \mathcal{O}_2, \kappa_2, p_2)$, a **homomorphism** is given by three mappings $\langle \phi : \mathcal{S}_1 \to \mathcal{S}_2, f : \mathcal{A}_1 \to \mathcal{A}_2, g : \mathcal{O}_1 \to \mathcal{O}_2 \rangle$ satisfying*

$$\kappa_2(s_2, o_2|\phi(s_1), f(a_1)) = \sum_{\substack{s' \in \phi^{-1}(s_2) \\ o' \in g^{-1}(o_2)}} \kappa_1(s', o'|s_1, a_1), \tag{2}$$

$$p_2(s_2) = \sum_{s_1 \in \phi^{-1}(s_2)} p_1(s_1), \tag{3}$$

for every $s_1 \in \mathcal{S}_1$, $a_1 \in \mathcal{A}_1$, $s_2 \in \mathcal{S}_2$, and $o_2 \in \mathcal{O}_2$.

Condition (2) says that the joint one-step distribution on the next state and output is preserved after applying the coarse-grainings ϕ and g (and translating actions through f). Condition (3) makes sure that the initial distributions are equivalent up to ϕ.

If we require $\mathcal{O}_1 = \mathcal{O}_2$ and $\mathcal{A}_1 = \mathcal{A}_2$ then both transducers have the same "type". Moreover, if also $f = g = \texttt{id}$ and ϕ is surjective we say that the homomorphism is a *reduction*.

Example 2. Consider the transducer from Figure 3. There is a reduction from this transducer to the one from Figure 2: define ϕ as $\phi(s_0) = \phi(s_2) = s_0$ and $\phi(s_1) = \phi(s_3) = s_1$, while taking $f = g = \texttt{id}$. In some sense, the transducer from Figure 3 implements the interface in an "inefficient" manner.

Reductions can be composed. Thus, after fixing an interface $\mathcal{I}$ we can look at the set of transducers implementing $\mathcal{I}$, and if we quotient them properly (identifying transducers

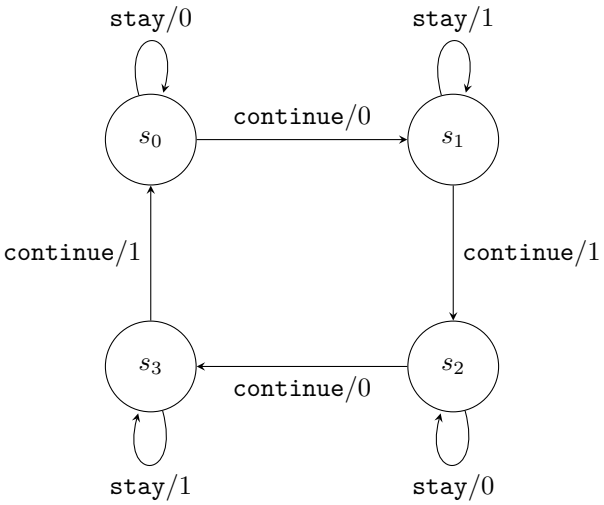


Figure 3: Example of a "deterministic" transducer that implements the same interface as the one from Figure 2. An edge from s to s' with label a/o indicates that $\kappa(s', o|s, a) = 1$.

T_1 and T_2 such that there are reductions both from T_1 to T_2 and from T_2 to T_1) then the reduction relation gives the set a *poset* structure.

In [44] some properties of these posets are proven, and in particular the fact that in general they need not have a unique minimum. This situation can be salvaged in at least two ways. First, if we consider linear transducers (which allow for "negative" probabilities), then uniqueness of the minimum can be proven [44][Theorem 2]. Second, we can restrict attention to the subposet of *predictive* transducers (intuitively, those whose state transitions are deterministic given the last state, action and output): in that case, there is a unique minimum, and it coincides with the ϵ-machine from computational mechanics [3] representing the dynamics [44][Theorem 3].

Before proceeding, we note that the definition of homomorphism we introduced is not exactly the same as the one from [44]. In Appendix A.1 we compare them and show nonetheless that they coincide when we restrict to reductions. Later we will see that our proposal is easier to extend to the approximate setting. In particular, Condition (2) states that κ_2 must be equal to the pushforward of κ_1 through ϕ and g. Thus, we can introduce an error term in the homomorphism by comparing κ_2 to this pushforward using any distance between distributions.

2.2 Linear transducers

We will also consider *linear transducers*: a model of a transducer in which the states are embedded in a vector space. From now on we define, for every interface $\mathcal{I}$, its associated formal series: for every $w \in \Sigma^* = (\mathcal{A} \times \mathcal{O})^*$ let

$$F_{\mathcal{I}}(w) = F_{\mathcal{I}}((a_1, o_1) \dots (a_n, o_n)) = \Pr_{\mathcal{I}}(o_1 \cdots o_n \mid a_1 \cdots a_n), \qquad F_{\mathcal{I}}(\epsilon) = 1,$$

where ϵ denotes the empty word.

Definition 3. *A **linear transducer** over $(\mathcal{A}, \mathcal{O})$ is a tuple $G = (V, \xi, \lambda, \{M_\sigma\}_{\sigma\in\Sigma})$, where V is a real vector space, $\xi \in V$ is an initial vector, $\lambda \in V^*$ is a linear functional, and each $M_\sigma : V \to V$ is a linear map. For a word $w = \sigma_1 \cdots \sigma_n$ we define M_w recursively by*

$$M_\epsilon = \mathrm{id}_V, \qquad M_{w\sigma} = M_\sigma M_w.$$

The linear transducer generates the formal series

$$F_G(w) = \lambda(M_w \xi).$$

We say that G implements an interface $\mathcal{I}$ if $F_G = F_\mathcal{I}$. We will assume without loss of generality that $V = \mathrm{span}\{M_w\xi : w \in \Sigma^*\}$, i.e. that the whole space V is "used" by the transducer[2]*.*

The size of a linear transducer is measured by the dimension of the linear space needed to represent the series, which can be infinite.

We now introduce the concept of the *Hankel matrix* of the interface.

Definition 4. *The **Hankel matrix** of $\mathcal{I}$ is the infinite matrix*

$$H_\mathcal{I}(u, v) = F_\mathcal{I}(uv), \qquad u, v \in \Sigma^*.$$

For each prefix $u \in \Sigma^*$ define the row $h_\mathcal{I}(u) : \Sigma^* \to \mathbb{R}$ such that $h_\mathcal{I}(u)(v) = F_\mathcal{I}(uv)$. Then, the dimension of $\mathcal{I}$ is

$$\dim(\mathcal{I}) = \mathrm{rank}(H_\mathcal{I}) = \dim \mathrm{span}\{h_\mathcal{I}(u) : u \in \Sigma^*\}.$$

If this rank is finite, we call $\mathcal{I}$ a finite-rank interface.

A linear transducer implementing an interface can be obtained from its Hankel matrix. Let

$$V_\mathcal{I} = \mathrm{span}\{h_\mathcal{I}(u) : u \in \Sigma^*\}.$$

Pick $\xi_\mathcal{I} = h_\mathcal{I}(\epsilon)$ and let $\lambda_\mathcal{I} : V_\mathcal{I} \to \mathbb{R}$ be evaluation at the empty suffix as

$$\lambda_\mathcal{I}(r) = r(\epsilon).$$

For every $\sigma \in \Sigma$, define the shift operator $R^\mathcal{I}_\sigma : V_\mathcal{I} \to V_\mathcal{I}$ by

$$R^\mathcal{I}_\sigma h_\mathcal{I}(u) = h_\mathcal{I}(u\sigma),$$

and extend linearly. This is well-defined: if $\sum_i \alpha_i h_\mathcal{I}(u_i) = 0$, then for every suffix v,

$$\sum_i \alpha_i h_\mathcal{I}(u_i\sigma)(v) = \sum_i \alpha_i F_\mathcal{I}(u_i\sigma v) = \sum_i \alpha_i h_\mathcal{I}(u_i)(\sigma v) = 0.$$

Thus, it follows that $G_\mathcal{I} = (V_\mathcal{I}, \xi_\mathcal{I}, \lambda_\mathcal{I}, \{R^\mathcal{I}_\sigma\}_{\sigma\in\Sigma})$ is a linear transducer, and $\lambda_\mathcal{I}(R^\mathcal{I}_w \xi_\mathcal{I}) = F_\mathcal{I}(w)$ for every word w.

This representation is minimal: if $G = (V, \xi, \lambda, \{M_\sigma\})$ implements $\mathcal{I}$, then

$$h_\mathcal{I}(u)(v) = F_\mathcal{I}(uv) = \lambda(M_v M_u \xi).$$

[2] We add this hypothesis to improve the clarity of our exposition. All results still hold when removing this condition.

for every word u, v. Hence all Hankel rows are obtained from vectors $M_u\xi \in V$, so

$$\operatorname{rank}(H_{\mathcal{I}}) \leq \dim V.$$

Note that minimal linear transducers are unique up to invertible linear changes of coordinates.

In the context of linear transducers we will use *linear reductions* to formalize the idea of homomorphisms between models.

Definition 5. *Let $G = (V, \xi, \lambda, \{M_\sigma\}_{\sigma\in\Sigma})$ and $G' = (W, \xi', \lambda', \{N_\sigma\}_{\sigma\in\Sigma})$ be linear transducers over the same input and output alphabets. A **linear reduction** from G to G' is a surjective linear map $L : V \to W$ satisfying*

1. *$L\xi = \xi'$.*
2. *$LM_\sigma = N_\sigma L$ for every $\sigma \in \Sigma$.*
3. *$\lambda' L = \lambda$.*

The first condition preserves the initial vector, the second says that L translates the internal dynamics, and the third preserves the "reading" of the vectors. These conditions imply that $\lambda'(N_w\xi') = \lambda(M_w\xi)$ for every word w.

By the construction above, it can be proven that any linear transducer G implementing $\mathcal{I}$ can be reduced to $G_{\mathcal{I}}$.

Lemma 1. *Let $G = (V, \xi, \lambda, \{M_\sigma\})$ be a linear transducer implementing an interface $\mathcal{I}$. Then there is a linear reduction $\rho_G : G \to G_{\mathcal{I}}$ given by*

$$\rho_G(M_w\xi) = h_{\mathcal{I}}(w).$$

Example 3. The deterministic transducer T from Figure 2 admits a simple linear representation. Let

$$V = \mathbb{R}^2, \qquad \xi = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \lambda(x_0, x_1) = x_0 + x_1,$$

where the two standard basis vectors represent the states s_0 and s_1. Consider the transition maps

$$M_{\mathsf{continue},0} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad M_{\mathsf{continue},1} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

and

$$M_{\mathsf{stay},0} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad M_{\mathsf{stay},1} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then, the linear transducer $G = (V, \xi, \lambda, (M_\sigma)_{\sigma\in\Sigma})$ implements the same interface as T.

In general, every standard transducer T can be transformed into a linear transducer whose underlying space has dimension equal to the number of states of T.

2.3 Predictive transducers

We finally consider predictive transducers, which correspond to transducers whose internal states do not contain predictive information that is unavailable from the observable input–output history. From now on, we say that a history $h = (a_1, o_1) \dots (a_n, o_n) \in \Sigma^*$ is *admissible* for an interface $\mathcal{I}$ if $\Pr_{\mathcal{I}}(o_1 \cdots o_n \mid a_1 \cdots a_n) > 0$. The empty history is always admissible. For any admissible history $h = (a_1, o_1) \dots (a_n, o_n)$ of an interface $\mathcal{I}$, we consider the *residual interface* $\mathcal{I}^h$ as the interface obtained by conditioning on h: for every $u \in \mathcal{A}^m$ and $v \in \mathcal{O}^m$,

$$\Pr_{\mathcal{I}^h}(v \mid u) := \frac{\Pr_{\mathcal{I}}(o_1 \cdots o_n v \mid a_1 \cdots a_n u)}{\Pr_{\mathcal{I}}(o_1 \cdots o_n \mid a_1 \cdots a_n)}. \tag{4}$$

For a transducer $T = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \kappa, p)$ and a state $s \in \mathcal{S}$, let $\mathcal{I}_{T,s}$ denote the interface generated by the same kernel κ with initial distribution δ_s (i.e. when all probability mass is concentrated on s). For an admissible history $h = (a_1, o_1) \dots (a_n, o_n)$, also write

$$q_T(s \mid h) := \Pr_T(S_n = s \mid o_{1:n}, a_{1:n})$$

for the posterior distribution over the internal state at step n after observing h.

Definition 6. *Let T be a transducer. We say that T is **predictive** if, for every admissible history h and every state s such that $q_T(s \mid h) > 0$,*

$$\mathcal{I}_{T,s} = \mathcal{I}_T^h. \tag{5}$$

Equivalently, conditional on the observable history, knowing the current internal state does not change the expected distribution for future events.

For this class of transducers there is always a minimal implementation of each interface, and it can be constructed explicitly. To do this, identify histories that make exactly the same predictions: for admissible histories h and h', define the *predictive equivalence relation* as

$$h \sim_{\mathcal{I}} h' \quad \Longleftrightarrow \quad \mathcal{I}^h = \mathcal{I}^{h'}.$$

Denote the equivalence class of h by $[h]_{\mathcal{I}}$ and let

$$\mathcal{S}_\epsilon(\mathcal{I}) := \{[h]_{\mathcal{I}} : h \text{ is admissible for } \mathcal{I}\}.$$

The transitions are defined in the expected way in the next definition. This construction corresponds to the notion of an ϵ-machine from computational mechanics [3].

Definition 7. *Let $\mathcal{I}$ be an interface. Its ϵ-**transducer** is given by*

$$E(\mathcal{I}) = (\mathcal{S}_\epsilon(\mathcal{I}), \mathcal{A}, \mathcal{O}, \kappa_\epsilon, \delta_{[\epsilon]_{\mathcal{I}}}),$$

where, for every admissible history h, action a, and output o, we set

$$\kappa_\epsilon(s', o \mid [h]_{\mathcal{I}}, a) := \mu_\epsilon(o \mid [h]_{\mathcal{I}}, a)\mathbf{1}\{s' = \delta_\epsilon([h]_{\mathcal{I}}, a, o)\}. \tag{6}$$

where $\mu_\epsilon(o \mid [h]_{\mathcal{I}}, a) = \Pr_{\mathcal{I}^h}(o \mid a)$ and $\delta_\epsilon([h]_{\mathcal{I}}, a, o) = [h(a, o)]_{\mathcal{I}}$.

Observe that this transducer evolves *deterministically*: for every state s, input a and output o there is a unique next possible state s'. This ensures that Eq. (5) is satisfied.

The next proposition states that this implementation is the minimal one among the predictive ones.

Proposition 1. *The transducer $E(\mathcal{I})$ implements $\mathcal{I}$ and is predictive. Moreover, if $T = (\mathcal{S}, \mathcal{A}, \mathcal{O}, \kappa, p)$ is any predictive transducer implementing $\mathcal{I}$, then there is a reduction from T to $E(\mathcal{I})$.*

Example 4. The transducer from Figure 2 is predictive. Moreover, it is also the minimal predictive transducer for that interface.

See Figure 4 for a diagram showcasing the structure of the poset of standard, linear and predictive transducers. As already mentioned, due to Lemma 1 and Proposition 1 the poset for linear and predictive transducers each has a minimum for every interface. Meanwhile, for the case of standard transducers there are interfaces for which there is no unique minimum.

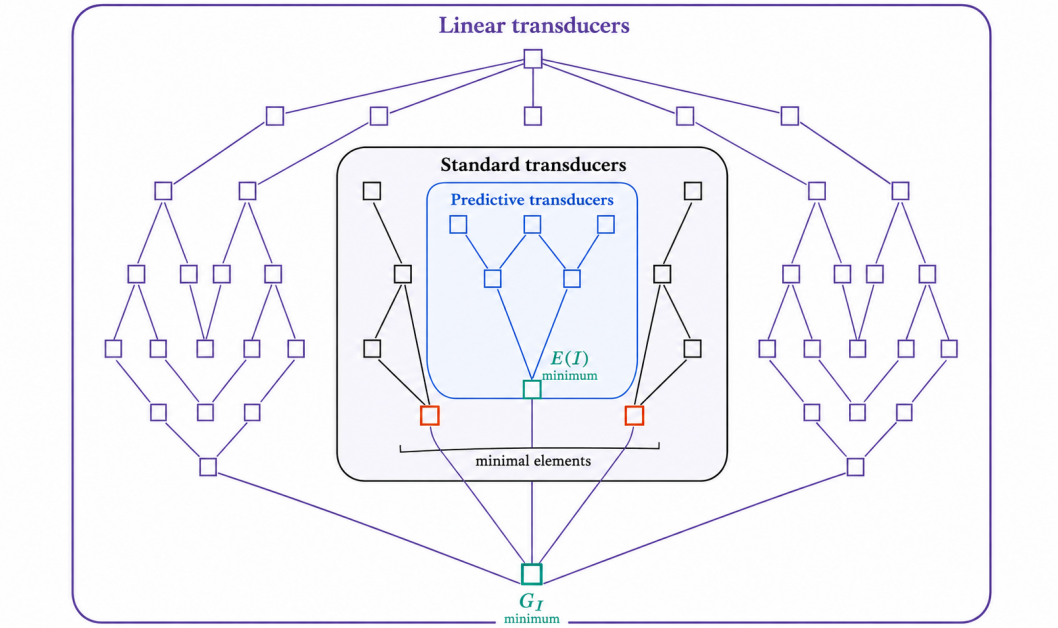


Figure 4: A diagram of the lattice of transducers for standard, linear and predictive implementations. Due to Lemma 1 and Proposition 1 the lattice of linear and predictive transducers has a unique minimum, while the one of standard transducers can have more than one minimal element.

3 Approximate homomorphisms and the space of interfaces

In this section we provide approximate variants of the notions of homomorphisms introduced in the previous section, and prove some basic properties.

3.1 The case of "standard" transducers

The type of coarse-grainings that Definition 2 allows is exact in a strong structural sense. It says that the whole one-step mechanism of T_2 is obtained by pushing forward the one-step mechanism of T_1 along the maps ϕ, f, and g. Thus, if two states of T_1 are identified by ϕ, they must have exactly the same coarse-grained output law and exactly the same coarse-grained transition law.

For real world models obtained through learning or other iterative procedures we don't expect them to be structurally identical. Thus, the purpose of approximate homomorphisms is to introduce some degree of error in this notion. We keep the maps ϕ, f, and g; but we now allow the push-forward dynamics to differ by some $\varepsilon > 0$.

Definition 8. *Given two transducers $T_1 = (\mathcal{S}_1, \mathcal{A}_1, \mathcal{O}_1, \kappa_1, p_1)$ and $T_2 = (\mathcal{S}_2, \mathcal{A}_2, \mathcal{O}_2, \kappa_2, p_2)$, a **$\varepsilon$-homomorphism** is given by three mappings $\langle \phi : \mathcal{S}_1 \to \mathcal{S}_2, f : \mathcal{A}_1 \to \mathcal{A}_2, g : \mathcal{O}_1 \to \mathcal{O}_2 \rangle$ satisfying*

$$\|(\phi \times g)_* \kappa_1(\cdot, \cdot | s_1, a_1) - \kappa_2(\cdot, \cdot | \phi(s_1), f(a_1))\|_{\mathrm{TV}} \leq \varepsilon \tag{7}$$

$$\|\phi_* p_1 - p_2\|_{\mathrm{TV}} \leq \varepsilon, \tag{8}$$

for every $s_1 \in \mathcal{S}_1$ and $a_1 \in \mathcal{A}_1$[3]*.*

The choice of total variation is not completely arbitrary: we will see that due to its properties (which are enumerated in the Appendix A.2) approximate homomorphisms are composable.

Example 5. Consider the actionless transducers from Figure 5 with $\varepsilon \in (0, 1]$. Each edge has a label (o, p) indicating the probability p of transitioning using that edge and outputting o in the process. There is a ε-reduction from the transducer on the left to the one on the right: take $\phi(s_0) = t_0$ and $\phi(s_1) = \phi(s_2) = t_1$. Meanwhile, there is no 0-reduction (i.e. exact reduction) between them.

This notion of approximate homomorphism ensures each state $s \in \mathcal{S}_1$ gets mapped to a state whose one-step dynamics are similar after coarse-graining. Thus, if we look at *approximate reductions* (enforcing that $\mathcal{A}_1 = \mathcal{A}_2$, $\mathcal{O}_1 = \mathcal{O}_2$, $f = g = \mathtt{id}$, and ϕ is surjective), one transducer T_1 can be approximately reduced to another one T_2 only if their states are *locally* similar. Does this imply that the interfaces they induce are also similar? We recall that for exact homomorphisms this is the case.

Observation 1. *If there is a reduction from T_1 to T_2 then $\mathcal{I}_{T_1} = \mathcal{I}_{T_2}$ [44][Lemma 5].*

To approach this question in the approximate setting we need a way to compare different interfaces, i.e. a metric over this space.

Observe that an interface $\mathcal{I}$ is given essentially by a map $D_{\mathcal{I}} : \mathcal{A}^* \to \Delta(\mathcal{O}^*)$ such that $D_{\mathcal{I}}(a_1 \dots a_n)$ represents the distribution $\Pr_{\mathcal{I}}(\cdot | a_1 \dots a_n)$ which has support over $\mathcal{O}^n$. Then, to define a metric for interfaces we can pick any metric for distributions and then aggregate it over all the possible action sequences in $\mathcal{A}^*$. For instance, we can consider total variation to compare the distributions and aggregate them with the supremum, obtaining

$$d_\infty(\mathcal{I}_1, \mathcal{I}_2) = \sup_{\mathbf{a} \in \mathcal{A}^*} ||D_{\mathcal{I}_1}(\mathbf{a}) - D_{\mathcal{I}_2}(\mathbf{a})||_{\mathrm{TV}} \tag{9}$$

[3]Here $(\phi \times g)_*$ and ϕ_* denote the push-forwards of the distributions. See Appendix A.2 for a precise definition.

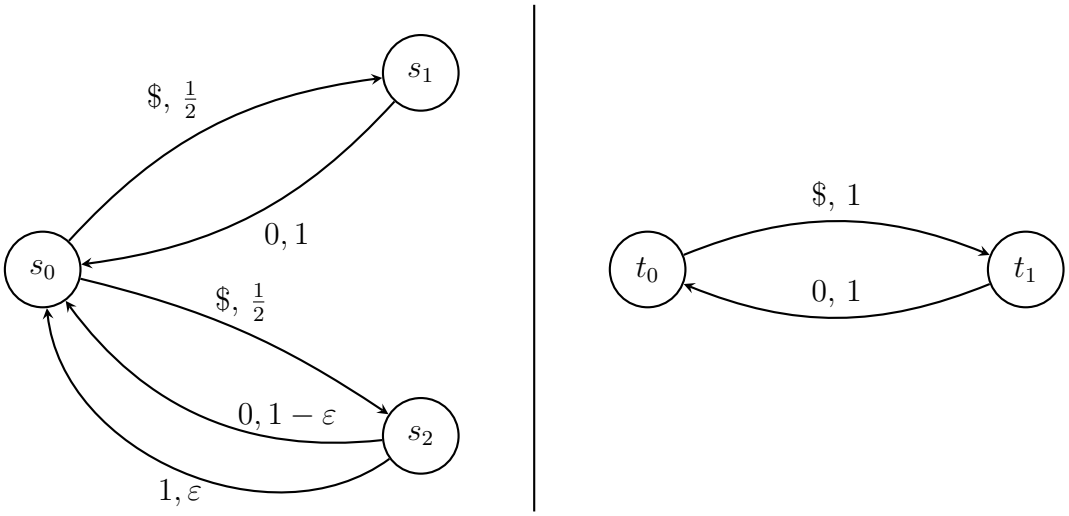


Figure 5: Two actionless transducers (or rather, transducers with a single action a), with outputs $\mathcal{O} = \{0, 1, \$\}$. An edge from s to s' with label o, p indicates that $\kappa(s', o|s, a) = p$.

We could also weight each sequence of actions according to its length, reflecting the choice to place less weight on long-horizon discrepancies. Thus, we can consider

$$d_\gamma(\mathcal{I}_1, \mathcal{I}_2) = \sum_{n=0}^{\infty} \gamma^n \sup_{\mathbf{a} \in \mathcal{A}^n} ||D_{\mathcal{I}_1}(\mathbf{a}) - D_{\mathcal{I}_2}(\mathbf{a})||_{\mathrm{TV}} \tag{10}$$

for some $\gamma \in (0, 1)$. Note that the distances in Eqs. (9) and (10) are indeed well-defined metrics over the set of interfaces.

Since each transducer T induces an interface $\mathcal{I}_T$ through Eq. (1), any metric between interfaces can be seen as a pseudometric[4] between transducers as

$$d(T_1, T_2) = d(\mathcal{I}_{T_1}, \mathcal{I}_{T_2}).$$

Are these metrics "continuous" with respect to the notion of reduction? Namely, is there a metric d and a function $f : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ with $f(x) \underset{x \to 0}{\to} 0$ such that, if there is an ε-reduction from T_1 to T_2, then $d(\mathcal{I}_{T_1}, \mathcal{I}_{T_2}) \leq f(\varepsilon)$? We first observe that this is not the case for the supremum distance in Eq. (9).

Example 6. Pick d_∞ as in Eq. (9), and consider the transducers from Example 5. Then, if T_1 is the transducer on the left and T_2 the one on the right, it can be seen that $d(\mathcal{I}_{T_1}, \mathcal{I}_{T_2}) = 1$ for every $\varepsilon > 0$.

Intuitively, the supremum distance is not controlled by the approximate homomorphism notion because the error bound applies only to the one-step dynamics. Thus, the interfaces implemented by the two transducers at long horizons (i.e. the distribution $\Pr(\cdot|\mathbf{a})$ for $\mathbf{a} \in \mathcal{A}^n$ with $n \to \infty$) can be arbitrarily far away in metrics such as total variation.

[4] It is a pseudo metric because different transducers implementing the same interface are at distance 0.

Nonetheless, this observation suggests that the discounted metrics from Eq. (10) might be preserved by the approximate homomorphism notion, and indeed this is the case.

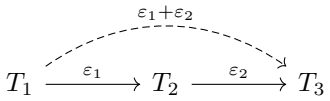
Theorem 1. *Suppose there is an ε-reduction from T_1 to T_2. Then, if d_γ is the distance from Eq. (10), it holds that*

$$d_\gamma(\mathcal{I}_{T_1}, \mathcal{I}_{T_2}) \leq \frac{\varepsilon}{(1-\gamma)^2}. \tag{11}$$

We write $T_1 \overset{\varepsilon}{\to} T_2$ to indicate that there is an ε-homomorphism from T_1 to T_2. As already noted, exact homomorphisms can be composed, and thus the reduction relation is transitive. For approximate homomorphisms we can prove the following additive version.

Proposition 2. *If $T_1 \overset{\varepsilon_1}{\to} T_2$ and $T_2 \overset{\varepsilon_2}{\to} T_3$, then $T_1 \overset{\varepsilon_1+\varepsilon_2}{\to} T_3$.*

This proposition states the existence of the dashed arrow in the following diagram:



Theorem 1 and Proposition 2 suggest that this notion of approximate homomorphism is natural and algebraically convenient. We recall that composability can be shown because we use total variation to compare the one-step dynamics: a different choice of distance to compare the distributions may not preserve this property.

3.2 The case of linear transducers

To introduce an approximation error in the exact linear reduction we will equip each state vector space with a norm, which we will use to measure the distance between different vectors (mainly, between the vector obtained through the reduction and the vectors from the transducer itself). Throughout this subsection, we assume that the output alphabet O is finite. If V is a normed vector space, we denote its dual norm by $\|\cdot\|_{V^*}$. Given a linear operator C, we write $||C||_{a\to b}$ to denote the norm $\sup_{x:||x||_a=1} ||Cx||_b$.

An arbitrary linear transducer does not necessarily implement an interface. In particular, there are some transducers for which the norm of the state vector tends to infinity as the transducer reads symbols. Such a behaviour troubles our notion of approximate reduction, since a small margin of error in the one-step dynamics can be amplified arbitrarily in the subsequent steps. Thus, to rule out this situation, we introduce the notion of *contractive* transducer.

Definition 9. *A linear transducer $G = (V, \xi, \lambda, \{M_{a,o}\}_{(a,o)\in A\times O})$ is* contractive *if*

$$\|\xi\|_V \leq 1, \qquad \|\lambda\|_{V^*} \leq 1, \qquad \|M_{a,o}x\|_V \leq \|x\|_V,$$

for every $a \in A$, $o \in O$, and $x \in V$.

The last condition ensures that after applying an evolution operator the norm of the vector state does not increase. Note that the canonical representation given by the Hankel matrix satisfies this definition. Indeed, on the Hankel row space $V_{\mathcal{I}}$, the prediction norm

$$\|r\|_{\text{pred}} = \sup_{v\in\Sigma^*} |r(v)|$$

makes every shift $R^{\mathcal{I}}_{a,o}$ nonexpansive, because

$$\|R^{\mathcal{I}}_{a,o} r\|_{\mathrm{pred}} = \sup_{v \in \Sigma^*} |r((a,o)v)| \leq \|r\|_{\mathrm{pred}}.$$

Moreover, $\|\xi_{\mathcal{I}}\|_{\mathrm{pred}} = 1$, while $\|\lambda_{\mathcal{I}}\|_{V^*_{\mathcal{I}}} \leq 1$ because $\lambda_{\mathcal{I}}(r) = r(\epsilon)$.

We now introduce our notion of approximate linear reduction.

Definition 10. *Let $G = (V, \xi, \lambda, \{M_{a,o}\}_{a,o})$ and $G' = (W, \xi', \lambda', \{N_{a,o}\}_{a,o})$ be contractive linear transducers over the same input and output alphabets, and let $\varepsilon \geq 0$. A bounded surjective linear map $L : V \to W$ is a* ε-linear reduction *from G to G' if*

$$\begin{aligned} \|L\xi - \xi'\|_W &\leq \varepsilon, \\ \|LM_{a,o}x - N_{a,o}Lx\|_W &\leq \varepsilon \|x\|_V, \\ \|\lambda' L - \lambda\|_{V^*} &\leq \varepsilon, \end{aligned}$$

for every $(a,o) \in A \times O$ and $x \in V$. We write $G \xrightarrow{\varepsilon} G'$ when such a map exists.

When G and G' are contractive, the conditions for $\varepsilon = 0$ are precisely the equations defining an exact linear reduction. This definition extends approximate reductions between standard transducers.

Proposition 3. *Let T_1 and T_2 be finite standard transducers over the same alphabets, and suppose that a surjective state map $\phi : \mathcal{S}_1 \to \mathcal{S}_2$ is an ε-reduction. Then, if G_1 and G_2 are the linear implementations corresponding to T_1 and T_2 equipped with their ℓ_1 norms, there exists a 2ε-linear reduction from G_1 to G_2 induced by ϕ.*

Approximate linear reductions can be composed in the same way as the standard approximate reductions. From now on, for a bounded linear map L, write $c(L) = \max\{1, \|L\|\}$. The following holds.

Proposition 4. *Suppose that $G_0 \xrightarrow{\varepsilon_1} G_1$ through the linear map L and $G_1 \xrightarrow{\varepsilon_2} G_2$ through the linear map K. Then KL is a $c(K)\varepsilon_1 + c(L)\varepsilon_2$ linear reduction from G_0 to G_2.*

We next compare the interfaces implemented by approximately reduced linear transducers. In the case of standard transducers we could prove in Theorem 1 that the discounted metrics were preserved after an approximate reduction. For linear transducers we obtain a similar result, but with a weaker bound.

Theorem 2. *Let G and G' be contractive linear transducers implementing interfaces $\mathcal{I}_G$ and $\mathcal{I}_{G'}$. If $G \xrightarrow{\varepsilon} G'$, then, for every $n \geq 1$,*

$$\sup_{\boldsymbol{a} \in \mathcal{A}^n} \|D_{\mathcal{I}_G}(\boldsymbol{a}) - D_{\mathcal{I}_{G'}}(\boldsymbol{a})\|_{\mathrm{TV}} \leq \min\left\{1, \frac{n+2}{2}|O|^n \varepsilon\right\}. \tag{12}$$

Consequently, for every $\gamma \in (0,1)$,

$$d_\gamma(\mathcal{I}_G, \mathcal{I}_{G'}) \leq \sum_{n \geq 0} \gamma^n \min\left\{1, \frac{n+2}{2}|O|^n \varepsilon\right\}, \tag{13}$$

which converges to 0 as $\varepsilon \to 0$.

Note that this implies that for small enough ε both transducers implement a similar interface. The bound is somewhat weaker when compared to the one from Theorem 1 because the notion of approximate reduction for standard transducers is stronger with respect to the one-step equivalence of the dynamics. For instance, Eq. (7) requires that the overall error (i.e. total variation) is bounded, while in Definition 10 we bound each error independently. This is the reason why a term $|\mathcal{O}|$ shows up in the bound. We could fix this by changing the definition of approximate linear reduction, but it would require us to also modify the notion of contractive transducer. Moreover, the required change gives a notion of contractive transducer which does not include the canonical Hankel representations, which we want to use in later proofs.

Nonetheless, we want to highlight the fact that there are many other valid choices regarding these definitions. In our case, we wanted to prioritize the fact that our abstractions should extend the notion of approximate reduction for standard transducers (proven in Proposition 3), should allow to represent the canonical Hankel constructions and should satisfy the simple and basic properties already seen for standard transducers (composability in Proposition 4 and continuity with regard to the discounted metrics in Theorem 2).

4 Approximate reductions between implementations of similar interfaces

In this section we will study the set of transducers which implement a given interface $\mathcal{I}$, and we will try to relate them through approximate homomorphisms. Moreover, we will look at the set of transducers implementing a *similar* interface (using one of the distances for interfaces mentioned previously). Ideally, we would like for this set of transducers to share some property, since that would indicate an *emergent* property related to the representation of the interfaces.

The following definition formalizes this set.

Definition 11. *Let $\mathcal{I}$ be an interface, $\varepsilon \geq 0$ and d some metric over the set of interfaces. We define*

$$\mathcal{L}_{\mathcal{I}}^{\varepsilon,d} = \{T : T \text{ is a transducer and } d(\mathcal{I}_T, \mathcal{I}) \leq \varepsilon\}.$$

as the set of transducers that ε-approximate $\mathcal{I}$.

For every set $\mathcal{L}_{\mathcal{I}}^{\varepsilon,d}$ we would like to understand whether there is some $T \in \mathcal{L}_{\mathcal{I}}^{\varepsilon,d}$ such that, for every other $T' \in \mathcal{L}_{\mathcal{I}}^{\varepsilon,d}$, it holds that $T' \xrightarrow{\delta} T$ for some small δ, ideally scaling as $\delta = O(\varepsilon)$. We call such a transducer a δ-minima of $\mathcal{L}_{\mathcal{I}}^{\varepsilon,d}$. We define

$$\delta_\varepsilon(\mathcal{I}) = \inf\left\{\delta \in \mathbb{R}_{\geq 0} : \mathcal{L}_{\mathcal{I}}^{\varepsilon,d} \text{ has a } \delta\text{-minima}\right\}.$$

With this notation, our goal is to find bounds for $\delta_\varepsilon(\mathcal{I})$ in terms of ε. Is there a subset of interfaces which is well-behaved in this sense? Does it matter which type of transducers we consider? See Figure 6 for a sketch of the type of behaviour that we aim for.

In the next subsections we will consider the set $\mathcal{L}_{\mathcal{I}}^{\varepsilon,d}$ restricted to different types of transducers. To avoid cluttering the notation we won't add any more indices to this symbol, but rather take the convention that in each respective subsection this set is restricted to the set of transducers studied in the corresponding subsection.

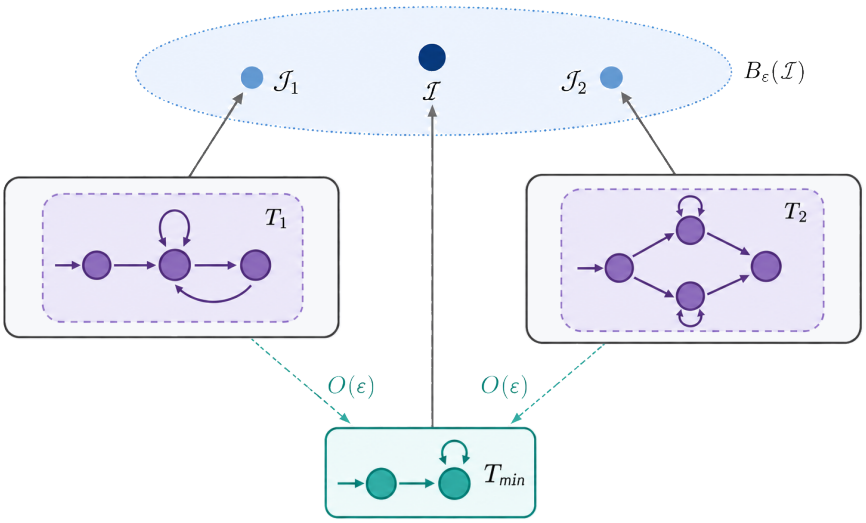


Figure 6: Schematic description of the type of convergence result we would like to prove. After fixing an interface of interest $\mathcal{I}$, we look at all interfaces ε-close to $\mathcal{I}$ in some distance. For each of these interfaces (such as $\mathcal{J}_1$ and $\mathcal{J}_2$) there are many transducers implementing the dynamics (respectively, T_1 and T_2). We say that there is a $O(\varepsilon)$-minimum if there is some transducer T_{min} implementing an interface from $B_\varepsilon(\mathcal{I})$ such that for any transducer T implementing an interface in $B_\varepsilon(\mathcal{I})$ it holds that $T \overset{O(\varepsilon)}{\to} T_{min}$. In the diagram this is represented by the transducer T_{min} implementing the interface $\mathcal{I}$, and there are $O(\varepsilon)$-reductions from both T_1 and T_2.

4.1 Non-existence of δ-minima for standard transducers

As mentioned before, there exist an $\mathcal{I}$ such that $\mathcal{L}_{\mathcal{I}}^{0,d}$ does not have a 0-minimum when considering only standard transducers[5]. Can this situation be avoided using δ reductions? Note that as δ increases we allow more reductions (in the limit, taking $\delta = 1$ allows every possible reduction), and thus it should make it simpler for convergent structures to arise.

In the next proposition we show that this is not the case.

Proposition 5. *There exists an interface $\mathcal{I}$ such that, for every $\delta < 1$, the set $\mathcal{L}_{\mathcal{I}}^{0,d}$ does not have a δ-minima.*

Even though $\mathcal{L}_{\mathcal{I}}^{0,d}$ does not have a common representation, it might be the case that when looking at approximate implementations of $\mathcal{I}$ there is some convergent structure. Again, the answer is negative.

Proposition 6. *Let $\mathcal{I}$ be the interface from Proposition 5. Then, for every $\varepsilon \geq 0$, every transducer C, and every $\delta < 1/2$, it is not true that T' δ-reduces to C for every $T' \in \mathcal{L}_{\mathcal{I}}^{\varepsilon,d_\infty}$. In particular, $\mathcal{L}_{\mathcal{I}}^{\varepsilon,d_\infty}$ does not have a δ-minima.*

These two results show that, in the case of standard transducers, there are interfaces for which no convergent structure exists between the different implementations of the interface,

[5]Note that if $\varepsilon = 0$ the choice of distance is irrelevant.

at least when we formalize this structure through our local notion of approximate homomorphism. This result is robust even when nearby interfaces are considered. Moreover, the counterexample is simple (it is a low-dimensional finite-rank interface), and the result can be proven for other distances (such as the discounted one d_γ). Thus, we don't believe that there is a reasonable restricted set of interfaces for which we could bound $\delta_\varepsilon(\mathcal{I})$ by $O(\varepsilon)$. We remark that these results are an extension of the ones from [44] in the context of approximate homomorphisms and approximate implementations of interfaces.

4.2 Existence of δ-minima for linear transducers

We now show a positive result for linear transducers. We will show that for all finite-rank interfaces $\mathcal{I}$ all nearby interfaces have a canonical representation which is similar to the one from $\mathcal{I}$. This is intuitive: note that the canonical construction is induced by the rows of the Hankel matrix. If an interface is slightly perturbed then the Hankel matrix is slightly perturbed as well. Thus, we can map the new Hankel matrix to the original one identifying each row with the corresponding one from the original matrix.

To obtain the strongest result possible we will give a norm to the canonical representation which dominates the predictive one.

Definition 12. *Let $\mathcal{I}$ be an interface and let $V_\mathcal{I} = \operatorname{span}\{h_\mathcal{I}(w) : w \in \Sigma^*\}$ be its Hankel row space. For $r \in V_\mathcal{I}$, define the atomic norm as*

$$\|r\|_{\mathrm{at},\mathcal{I}} := \inf\left\{\sum_{i=1}^{m} |\alpha_i| : r = \sum_{i=1}^{m} \alpha_i h_\mathcal{I}(w_i)\right\}.$$

We write $G^{\mathrm{at}}_\mathcal{I}$ for the canonical Hankel implementation equipped with this norm.

We are using the word *atom* to refer to each row of the Hankel matrix. We formally show that $||\cdot||_{\mathrm{at}}$ is a norm.

Lemma 2. *Let $\mathcal{I}$ be an interface and equip its Hankel row space $V_\mathcal{I}$ with the atomic norm. Then $\|\cdot\|_{\mathrm{at},\mathcal{I}}$ is a norm and, for every $r \in V_\mathcal{I}$,*

$$\|r\|_{\mathrm{pred}} \le \|r\|_{\mathrm{at},\mathcal{I}}.$$

Moreover,

$$\|\xi_\mathcal{I}\|_{\mathrm{at},\mathcal{I}} = 1, \qquad \|\lambda_\mathcal{I}\|_{(V_\mathcal{I},\|\cdot\|_{\mathrm{at},\mathcal{I}})^*} = 1,$$

and every shift $R^\mathcal{I}_\sigma$ is nonexpansive. Consequently, $G^{\mathrm{at}}_\mathcal{I}$ is a contractive linear transducer.

This norm measures what is the best way to write r as a linear sum of the rows of the Hankel matrix, where we weight each sum with the sum of the absolute values of its coefficients.

Our notion of linear reduction requires the mapping to be surjective. Thus, to ensure this property we will look at invertible minors of the Hankel matrix. Let $\mathcal{I}$ be a finite-rank interface with $d = \operatorname{rank}(H_\mathcal{I})$. We can choose prefixes $p_1, \dots, p_d$ and suffixes $q_1, \dots, q_d$ such that

$$C_\mathcal{I} := \big(F_\mathcal{I}(p_i q_j)\big)_{i,j=1}^{d} \tag{14}$$

is invertible and $q_1 = \epsilon$. Put

$$B_{\mathcal{I}} = \begin{pmatrix} h_{\mathcal{I}}(p_1) \\ \vdots \\ h_{\mathcal{I}}(p_d) \end{pmatrix}, \qquad \Gamma_{\mathcal{I}} := \|C_{\mathcal{I}}^{-1}\|_{\infty\to 1}.$$

For any interface $\mathcal{J}$, define analogously $C_{\mathcal{J}} := \big(F_{\mathcal{J}}(p_i q_j)\big)_{i,j=1}^{d}$ and $\mathrm{ev}_{\mathcal{J}} : V_{\mathcal{J}} \to \mathbb{R}^d$ given by

$$\mathrm{ev}_{\mathcal{J}}(r) := \big(r(q_1), \ldots, r(q_d)\big),$$

. Finally, let $\Pi_{\mathcal{J}\to\mathcal{I}} : V_{\mathcal{J}} \longrightarrow V_{\mathcal{I}}$ be the linear map

$$\Pi_{\mathcal{J}\to\mathcal{I}}(r) := \mathrm{ev}_{\mathcal{J}}(r) C_{\mathcal{I}}^{-1} B_{\mathcal{I}}.$$

The mapping $\Pi_{\mathcal{J}\to\mathcal{I}}$ translates $r \in V_{\mathcal{J}}$ into a vector from $V_{\mathcal{I}}$ by first evaluating the suffixes $\{q_i\}_{1\le i\le d}$, then doing a change of coordinates using $C_{\mathcal{I}}$ and finally projecting the result into the rows from $V_{\mathcal{I}}$ indexed by $\{p_i\}_{1\le i\le d}$.

The following lemma shows that this mapping commutes with the shift operators up to a small error with respect to the $||\cdot||_{\mathrm{at}}$ norm if the interfaces are close. Moreover, whenever $C_{\mathcal{J}}$ is invertible the mapping is surjective.

Lemma 3. *If $d_\infty(\mathcal{I}, \mathcal{J}) \le \varepsilon$, then, for every $w \in \Sigma^*$,*

$$\|\Pi_{\mathcal{J}\to\mathcal{I}} h_{\mathcal{J}}(w) - h_{\mathcal{I}}(w)\|_{\mathrm{at},\mathcal{I}} \le \Gamma_{\mathcal{I}}\varepsilon.$$

Moreover, if $C_{\mathcal{J}}$ is invertible, then $\Pi_{\mathcal{J}\to\mathcal{I}}$ is surjective.

It is a well-known fact that if a finite matrix M is invertible, then adding a small amount of noise to M keeps it invertible. We apply this observation to $C_{\mathcal{I}}$ to guarantee that $C_{\mathcal{J}}$ remains invertible.

Theorem 3. *Let $\mathcal{I}$ be a finite-rank interface and choose the minor $C_{\mathcal{I}}$ as in (14), with $q_1 = \epsilon$. Then, there exists $\overline{\varepsilon}(\mathcal{I}) > 0$ such that, for every interface $\mathcal{J}$ satisfying*

$$d_\infty(\mathcal{I}, \mathcal{J}) \le \varepsilon < \overline{\varepsilon}(\mathcal{I}),$$

the map $\Pi_{\mathcal{J}\to\mathcal{I}} : G_{\mathcal{J}}^{\mathrm{at}} \to G_{\mathcal{I}}^{\mathrm{at}}$ is a $2\Gamma_{\mathcal{I}}\varepsilon$-linear reduction.

From this theorem we get as an immediate corollary the existence of δ-minima for a small enough neighbourhood of every finite-rank interface.

Corollary 1. *For every finite-rank interface $\mathcal{I}$, its canonical realization $G_{\mathcal{I}}$, equipped with the atomic norm, is a $2\Gamma_{\mathcal{I}}\varepsilon$-minimum of $\mathcal{L}_{\mathcal{I}}^{\varepsilon,d_\infty}$ for every $0 \le \varepsilon < \overline{\varepsilon}(\mathcal{I})$, when restricting $\mathcal{L}_{\mathcal{I}}^{\varepsilon,d_\infty}$ to contain only the canonical linear realizations with the atomic norm.*

Note that our main theorem has to bound ε to ensure that the reduction is surjective. Even though this requirement is reasonable (otherwise, we could reduce small transducers into subcomponents of bigger ones), in many applications it might make sense to ignore this restriction. That's why we phrased Theorem 3 in an independent way.

The results in this section are in some sense satisfactory: we observed that convergent structure (i.e. a δ-minima for $\delta = O(\varepsilon)$) exists for every finite rank interface in a neighbourhood of the interface. Observe that in our statements we have to pick a norm and a

distance in a somewhat arbitrary way. However, a similar result can be proven using the predictive norm. Conceptually, we believe that these result indicate that the convergent structure exists at the level of linear transducers even in the presence of perturbations in the implementations.

We remark that in Corollary 1 we restrict the lattice to the minimal linear implementations. If we don't do this, we still can prove the existence of a reduction from any linear transducer G in the set by composing the map ρ_G from Lemma 1 with the one from Theorem 3 whenever ρ_G is bounded. Then, using Proposition 4 we would obtain an error that depends on $\|\rho_G\|$. However, there is no uniform bound on this value when using the atomic norm. There are other choices of norms which can solve this problem but they seem quite unnatural, and therefore we prefer to keep the corollary as stated, applying only to the lattice of "optimal" implementations.

4.3 Existence of δ-minima for predictive transducers under a specific metric

As already mentioned, restricting to predictive transducers restores a canonical minimum for each fixed interface, which we denote by $\mathsf{E}(\mathcal{I})$ (see Proposition 1). This exact statement is not stable under the supremum metric.

Proposition 7. *There is an interface $\mathcal{I}$ and a sequence of interfaces $(\mathcal{J}_n)_{n\geq 1}$ such that*

$$d_\infty(\mathcal{I}, \mathcal{J}_n) = 2^{-n} \longrightarrow 0,$$

but every approximate reduction $\mathsf{E}(\mathcal{J}_n) \overset{\delta}{\to} \mathsf{E}(\mathcal{I})$ satisfies $\delta \geq 1/2$. Moreover, if a transducer C receives δ-reductions from both $\mathsf{E}(\mathcal{I})$ and $\mathsf{E}(\mathcal{J}_n)$, then $\delta \geq 1/4$.

This counterexample also applies to the discounted metric, and we believe it highlights a limitation of predictive transducers. More precisely, if there is a history h such that the interface $\mathcal{I}$ conditioned on h behaves in an extremely different way from the interface $\mathcal{J}$ conditioned on h, then the local structure related to both predictive states will be different, even if h is highly unlikely (and thus, it is ignored by most reasonable metrics). We now show that, as one would expect, this problem can be avoided by using precisely a notion of distance between interfaces that values every possible conditioning independently of its probability.

For a positive probability history h, let $\mathcal{I}^h$ denote the residual interface after conditioning on h (see Eq. (4) for the exact definition). Write $\mathrm{supp}(\mathcal{I})$ for the set of positive probability histories and define

$$d_{\mathrm{res}}(\mathcal{I}, \mathcal{J}) = \begin{cases} \sup\limits_{h \in \mathrm{supp}(\mathcal{I})} d_\infty(\mathcal{I}^h, \mathcal{J}^h), & \mathrm{supp}(\mathcal{I}) = \mathrm{supp}(\mathcal{J}), \\ 1, & \text{otherwise.} \end{cases}$$

This metric compares the predictive laws after every history, including the ones with low probability. Let $[h]_{\mathcal{I}}$ denote the state of $\mathsf{E}(\mathcal{I})$ reached after h, and put

$$\Delta_{\mathcal{I}} = \inf \left\{ d_\infty(\mathcal{I}^h, \mathcal{I}^u) : h, u \in \mathrm{supp}(\mathcal{I}), [h]_{\mathcal{I}} \neq [u]_{\mathcal{I}} \right\}.$$

We use the convention $\Delta_{\mathcal{I}} = +\infty$ if there is only one predictive state. If $\mathsf{E}(\mathcal{I})$ is finite, then $\Delta_{\mathcal{I}} > 0$. Intuitively, $\Delta_{\mathcal{I}}$ is a lower bound on the difference between the interfaces induced by each of the predictive states.

If there is an interface $\mathcal{J}$ such that $d_{\text{res}}(\mathcal{I}, \mathcal{J})$ is much smaller than $\Delta_{\mathcal{I}}$ then the predictive transducers implementing $\mathcal{I}$ and $\mathcal{J}$ will be similar. We formalize this in our last theorem.

Theorem 4. *Let $\mathcal{I}$ and $\mathcal{J}$ be interfaces over the same input and output alphabets. Suppose that $\Delta_{\mathcal{I}} > 0$ and that*

$$d_{\text{res}}(\mathcal{I}, \mathcal{J}) \leq \varepsilon < \min\left\{1, \frac{\Delta_{\mathcal{I}}}{2}\right\}.$$

Then the assignment

$$\phi([h]_{\mathcal{J}}) := [h]_{\mathcal{I}},$$

is a well-defined surjective state map and determines an ε-reduction $E(\mathcal{J}) \longrightarrow E(\mathcal{I})$. Consequently, every predictive transducer implementing $\mathcal{J}$ admits an ε-reduction to $E(\mathcal{I})$.

The intuition for the condition $\varepsilon < \frac{\Delta_{\mathcal{I}}}{2}$ comes from the triangle inequality: if we want to prove that ϕ is well defined, we need to ensure that if $h_1 \sim h_2$ in $\mathcal{J}$ then $h_1 \sim h_2$ in $\mathcal{I}$. This is equivalent to verifying that $\mathcal{I}^{h_1} = \mathcal{I}^{h_2}$, and by triangle inequality we see that

$$\begin{aligned} d_\infty(\mathcal{I}^{h_1}, \mathcal{I}^{h_2}) &\leq d_\infty(\mathcal{I}^{h_1}, \mathcal{J}^{h_1}) + d_\infty(\mathcal{J}^{h_1}, \mathcal{J}^{h_2}) + d_\infty(\mathcal{J}^{h_2}, \mathcal{I}^{h_2}) \\ &\leq 2\varepsilon < \Delta_{\mathcal{I}}. \end{aligned}$$

Then, by definition of $\Delta_{\mathcal{I}}$ it must be the case that $\mathcal{I}^{h_1} = \mathcal{I}^{h_2}$.

As a corollary, we obtain the existence of δ-minimums for the set of predictive transducers.

Corollary 2. *Let $\mathcal{I}$ be an interface such that $\Delta_{\mathcal{I}} > 0$. Then, for every $0 < \varepsilon < \min\left\{1, \frac{\Delta_{\mathcal{I}}}{2}\right\}$ and every predictive transducer T implementing an interface $\mathcal{J}$ with $d_{\text{res}}(\mathcal{I}, \mathcal{J}) \leq \varepsilon$ there is a ε-reduction from T to $\mathsf{E}(\mathcal{I})$.*

This is a positive result regarding convergent structure, but its application is restricted to scenarios where the residual distance makes sense. In the following we describe an example of a situation in which two different stochastic systems can induce interfaces which are close according to d_{res}.

Example 7. Let $\mathcal{S} = \mathcal{O} = \{0, 1, 2\}$ and $\mathcal{A} = \{\star\}$ (i.e. the dynamics are actionless because there is a single action). and consider the transition matrix

$$P = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}.$$

Let state i output i before making a transition. Thus, the transition kernel is

$$\kappa(j, o \mid i) = P(i, j)\mathbf{1}_{\{o=i\}}.$$

Consider two transducers T and T_ε using the kernel κ, but with different initial distributions given by $p = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ for T and $p_\varepsilon = \left(\frac{1}{3} + \varepsilon, \frac{1}{3} - \varepsilon, \frac{1}{3}\right)$ for T_ε. See Figure 7 for a visual description.

It can be seen that $d_{\text{res}}(\mathcal{I}_T, \mathcal{I}_{T_\varepsilon}) \leq \varepsilon$: note that conditioning on any non-empty history h the interfaces $\mathcal{I}_T^h$ and $\mathcal{I}_{T_\varepsilon}^h$ are equal, and moreover $d_\infty(\mathcal{I}_T, \mathcal{I}_{T_\varepsilon}) \leq \varepsilon$. Moreover, it can be seen that $\Delta_{\mathcal{I}_T} = \frac{1}{6}$. Thus, due to Theorem 4 we conclude that if $\varepsilon < \frac{1}{12}$ it holds that the minimal predictive implementation of $\mathcal{I}_T$ and of $\mathcal{I}_{T_\varepsilon}$ are ε-close through an approximate reduction.

More generally, different processes with the same transition dynamics but different initial distributions can be close in the d_{res} metric whenever the effect of conditioning ensures that the current state for both processes is the same.

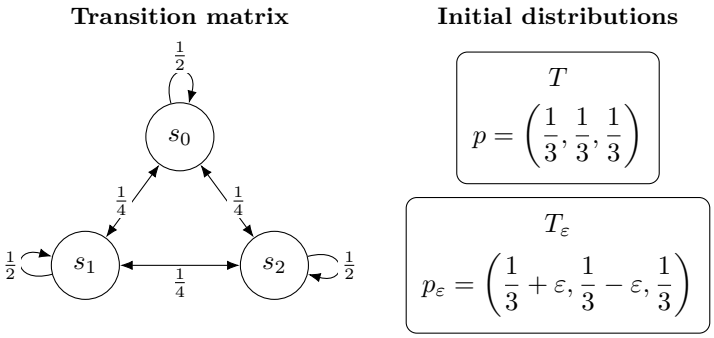


Figure 7: We consider two transducers T and T_ε with the same transition kernel described by the Markov process on the left but different initial distributions given on the right side. It can be seen that for small enough $\varepsilon > 0$ the induced interfaces satisfy $d_{\text{res}}(\mathcal{I}_T, \mathcal{I}_{T_\varepsilon}) \leq \varepsilon < \Delta_{\mathcal{I}}/2$.

5 Conclusion

Summary. In this work we looked for theoretical evidence supporting the empirical observation that different neural network models, sometimes even supported on different architectures, tend to converge to similar representations in their internal layers. To investigate this idea we proposed the approach described in Figure 1: we assumed that the reasoning inside the internal layers can be represented through some abstraction, and then we tried to prove some convergence at the level of these objects. In our case, we considered transducers to capture these world models, and to find convergent structure we looked for homomorphisms between them. Previous work had already proven that for the case of linear and predictive transducers there always exists a minimal transducer implementing a given dynamics, and that for all other non-minimal transducers there is a homomorphism to this minimal one [44]. In this paper we improved this result by showing that the existence of such an homomorphism remains even when we consider transducers that do not implement *exactly* the same dynamics.

To do this, we first introduced a notion of approximate homomorphism for standard transducers (Definition 8) as well as for the linear ones (Definition 9). We showed that, although there are many ways to define such a family of homomorphisms, the ones proposed here have good algebraic properties: they preserve the dynamics under discounted metrics (Theorems 1 and 2), are composable (Propositions 2 and 4), and the linear approximate homomorphism is a direct extension of the standard one (Proposition 3).

With these tools developed, we looked for convergence theorems in the approximate setting: we looked for conditions under which all transducers implementing similar dynamics share some common structure, which we aimed to capture through approximate homomorphisms (see Figure 6 for a visual sketch of the idea). In particular, we showed that (1) For standard transducers, this type of convergence seems to not be possible, even when considering simple dynamics (Propositions 5 and 6), (2) For linear transducers, simple enough dynamics (more technically, finite-rank interfaces) always admit a type of convergence between all the minimal linear implementations of ε-close dynamics (Theorem 3), and, finally, that (3) For predictive transducer, there is a metric such that all transducer implementing dynamics close enough according to this metric will share structural properties between them (Theorem 4).

We believe these are positive theoretical results regarding the existence of convergent structure, in the context of both both linear and predictive transducers. The former family seems to be the natural model for capturing latent representation in modern neural networks, considering especially that the model parameters live in a vector space. The latter one, although less natural, still has been seen to show up inside the residual stream of transformers [46, 47], and thus understanding these representational properties might shed light into the behaviour of modern LLMs.

Limitations. Through the development of this work we discovered that there are many ways to formalize approximate convergence in the context of transducers. Although the proposals here satisfy good properties and extend previous ones [41] there is still room for developing a more general theory of approximate homomorphisms. Moreover, in the context of linear transducers we had to equip the underlying vector space with a norm to measure distance between vectors, thus introducing another "parameter" to our theory. Although many of our main theorems can be proven for other choices of norms and distances (such as Theorem 3), it would be great to have a more robust understanding of the precise hypothesis required to conclude structural convergence in the approximate setting.

Future work. We describe some future lines of work starting from the developments in this paper.

- **Experimental validation**: these results give predictions on the structural convergence of deep neural networks under the hypothesis that they use transducers in their latent space. In particular, for models trained on similar data it must be the case that their internal representation can be translated with a linear map (such type of translation scheme is usually referred to as "stitching", and has been studied in the literature [2, 9]). To validate these hypotheses, it would be interesting to train modern models with data generated from specific linear transducers and then see whether these transducers can be found in the learned representations. To do this, one could reproduce the setting from [46, 47].

- **Poset structure**: the original goal of this project was to study the poset of approximate homomorphism between world models. More precisely, we would like to understand how this poset looks like when we order it through the relation induced by the existence of an approximate homomorphism. A central and simple question is: under which hypothesis can we guarantee that this poset has cut-points, in the sense of intermediate models M such that, for any other model M', it holds that M has an homomorphism to M' or the other way around. Given the introduced notions of approximate homomorphisms, this question can now be approached in the context of transducers.

- **Improved abstractions**: Our abstractions are still limited and do not represent the myriad of forms on which "abstraction" and "reasoning" can occur inside modern AI models. Two recognizable improvement would be (1) introducing some non-linearity in the notion of linear transducers, to model the effect of the activation functions between layers, and (2) introducing a global error inside the notion of approximate homomorphism to allow homomorphisms that preserve the local structure of most of the states but fail completely in a small subset (such a notion would capture more faithfully what happens during model stitching when the target network is a bigger model than the source network).

6 Acknowledgements

This work was funded by the Advanced Research + Invention Agency (ARIA) through project code MSAI-SE01-P005. We would like to thank the Dovetail Research team[6] for comments and suggestions on the draft of this paper, and we are especially grateful to Alex Altair, Alfred Harwood, Jose Faustino and Neal Batra for fruitful discussions.

AI disclosure: We used ChatGPT 5.5 and 5.6 for proofreading, creating diagrams, writing down simple proofs (such as the proof from Lemma 2) and quickly exploring variants of the results (such as checking whether the proof of Theorem 3 holds for other choices of distances and norms). All content created by AI was revised and rewritten to improve readability and clarity of exposition.

A Appendix

A.1 Comparison of notions of exact homomorphisms

We describe the original formulation of homomorphism and see how it differs from ours.

Definition 13 (Homomorphism from [44])**.** *Given two transducers $T_1 = (\mathcal{S}_1, \mathcal{A}_1, \mathcal{O}_1, \kappa_1, p_1)$ and $T_2 = (\mathcal{S}_2, \mathcal{A}_2, \mathcal{O}_2, \kappa_2, p_2)$, a homomorphism is given by three mappings $\langle \phi : \mathcal{S}_1 \to \mathcal{S}_2, f : \mathcal{A}_1 \to \mathcal{A}_2, g : \mathcal{O}_1 \to \mathcal{O}_2 \rangle$ satisfying (3), the condition*

$$\Pr_{T_2}(o_2|\phi(s_1), f(a_1)) = \sum_{o_1 \in g^{-1}(o_2)} \Pr_{T_1}(o_1|s_1, a_1). \tag{15}$$

for every $s_1 \in \mathcal{S}_1, a_1 \in \mathcal{A}_1$ and $o_2 \in \mathcal{O}_2$, and whenever $\Pr_{T_1}(o_1|s_1, a_1) > 0$ the condition

$$\Pr_{T_2}(s_2|\phi(s_1), f(a_1), g(o_1)) = \sum_{s' \in \phi^{-1}(s_2)} \Pr_{T_1}(s'|s_1, a_1, o_1) \tag{16}$$

for every $o_1 \in \mathcal{O}_1$ and $s_2 \in \mathcal{S}_2$.

We now show that this definition and ours coincide when g is injective.

Proposition 8. *Definitions 2 and 13 coincide if g is injective.*

Proof. Let $T_1 = (\mathcal{S}_1, \mathcal{A}_1, \mathcal{O}_1, \kappa_1, p_1)$ and $T_2 = (\mathcal{S}_2, \mathcal{A}_2, \mathcal{O}_2, \kappa_2, p_2)$ be transducers, and let $\langle \phi, f, g \rangle$ be maps $\phi : \mathcal{S}_1 \to \mathcal{S}_2$, $f : \mathcal{A}_1 \to \mathcal{A}_2$ and $g : \mathcal{O}_1 \to \mathcal{O}_2$. We will prove that this tuple satisfies Definition 2 if and only if it satisfies Definition 13.

Assume first that the joint-kernel condition (2) holds. Summing both sides over $s_2 \in \mathcal{S}_2$ gives

$$\begin{aligned}
\Pr_{T_2}(o_2|\phi(s_1), f(a_1)) &= \sum_{s_2 \in \mathcal{S}_2} \kappa_2(s_2, o_2|\phi(s_1), f(a_1)) \\
&= \sum_{s_2 \in \mathcal{S}_2} \sum_{\substack{s' \in \phi^{-1}(s_2) \\ o' \in g^{-1}(o_2)}} \kappa_1(s', o'|s_1, a_1) \\
&= \sum_{o' \in g^{-1}(o_2)} \Pr_{T_1}(o'|s_1, a_1),
\end{aligned}$$

[6] https://dovetailresearch.org/.

which is exactly (15). Now fix $o_1 \in \mathcal{O}_1$ such that $\Pr_{T_1}(o_1|s_1, a_1) > 0$. Since g is injective, $g^{-1}(g(o_1)) = \{o_1\}$. Hence

$$\begin{aligned}\Pr_{T_2}(s_2|\phi(s_1), f(a_1), g(o_1)) &= \frac{\kappa_2(s_2, g(o_1)|\phi(s_1), f(a_1))}{\Pr_{T_2}(g(o_1)|\phi(s_1), f(a_1))} \\ &= \frac{\sum_{s' \in \phi^{-1}(s_2)} \kappa_1(s', o_1|s_1, a_1)}{\Pr_{T_1}(o_1|s_1, a_1)} \\ &= \sum_{s' \in \phi^{-1}(s_2)} \Pr_{T_1}(s'|s_1, a_1, o_1).\end{aligned}$$

Thus the original conditional formulation follows. The initial condition is the same in both definitions.

Conversely, assume Definition 13. We prove (2). If $o_2 \notin g(\mathcal{O}_1)$, then the right-hand side of (15) is zero. Therefore $\Pr_{T_2}(o_2|\phi(s_1), f(a_1)) = 0$, and then $\kappa_2(s_2, o_2|\phi(s_1), f(a_1)) = 0$ for every s_2.

It remains to consider $o_2 \in g(\mathcal{O}_1)$. By injectivity there is a unique $o_1 \in \mathcal{O}_1$ such that $g(o_1) = o_2$. If $\Pr_{T_1}(o_1|s_1, a_1) > 0$, multiplying (16) by (15) gives

$$\begin{aligned}\kappa_2(s_2, o_2|\phi(s_1), f(a_1)) &= \Pr_{T_2}(o_2|\phi(s_1), f(a_1)) \Pr_{T_2}(s_2|\phi(s_1), f(a_1), o_2) \\ &= \Pr_{T_1}(o_1|s_1, a_1) \sum_{s' \in \phi^{-1}(s_2)} \Pr_{T_1}(s'|s_1, a_1, o_1) \\ &= \sum_{s' \in \phi^{-1}(s_2)} \kappa_1(s', o_1|s_1, a_1) \\ &= \sum_{\substack{s' \in \phi^{-1}(s_2) \\ o' \in g^{-1}(o_2)}} \kappa_1(s', o'|s_1, a_1).\end{aligned}$$

If $\Pr_{T_1}(o_1|s_1, a_1) = 0$, then (15) gives $\Pr_{T_2}(o_2|\phi(s_1), f(a_1)) = 0$. Again nonnegativity forces both sides of (2) to be zero. The kernel condition therefore holds in all cases, and the initial condition is shared by the two definitions. □

When g is not injective, the two formulations need not agree. In the original definition, the only outputs that can be coarse-grained are those with the same output laws for each state. Meanwhile, our formulation only requires equality after averaging over the whole fiber $g^{-1}(o)$ and $\phi^{-1}(s)$. As already mentioned, since we focus on reductions this distinction is irrelevant.

A.2 Total variation

We collect here some relevant facts about total variation. All probability spaces in the sequel are finite or countable.

Definition 14 (Total variation)**.** *Let $\mu, \nu \in \Delta(X)$. Their total variation distance is*

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{A \subseteq X} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{x \in X} |\mu(x) - \nu(x)|.$$

If $h : X \to Y$ is a map and $\mu \in \Delta(X)$, we write $h_*\mu \in \Delta(Y)$ for the push-forward distribution,

$$(h_*\mu)(y) = \sum_{x \in h^{-1}(y)} \mu(x).$$

If $K : X \to \Delta(Y)$ is a Markov kernel and $\mu \in \Delta(X)$, we write $\mu K \in \Delta(Y)$ for the distribution

$$(\mu K)(y) = \sum_{x \in X} \mu(x) K(y|x).$$

Lemma 4. *Let $\mu, \nu \in \Delta(X)$, let $h : X \to Y$ and $r : Y \to Z$ be maps, and let $K, L : X \to \Delta(Y)$ be Markov kernels. Then, the following properties hold:*

1. ***Convexity.*** *If $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$, and $\mu_i, \nu_i \in \Delta(X)$, then*

$$\left\| \sum_i \lambda_i \mu_i - \sum_i \lambda_i \nu_i \right\|_{\mathrm{TV}} \leq \sum_i \lambda_i \|\mu_i - \nu_i\|_{\mathrm{TV}}.$$

2. ***Push-forwards compose.***

$$(r \circ h)_* \mu = r_*(h_*\mu).$$

3. ***Push-forward contraction.***

$$\|h_*\mu - h_*\nu\|_{\mathrm{TV}} \leq \|\mu - \nu\|_{\mathrm{TV}}.$$

   In particular, marginalization contracts total-variation.

4. ***Kernel contraction.***

$$\|\mu K - \nu K\|_{\mathrm{TV}} \leq \|\mu - \nu\|_{\mathrm{TV}}.$$

5. ***Kernel perturbation bound.*** *If*

$$\sup_{x \in X} \|K(\cdot|x) - L(\cdot|x)\|_{\mathrm{TV}} \leq \varepsilon,$$

   then

$$\|\mu K - \nu L\|_{\mathrm{TV}} \leq \|\mu - \nu\|_{\mathrm{TV}} + \varepsilon.$$

Proof. Convexity follows directly from the ℓ^1 expression for total variation and the triangle inequality:

$$\begin{aligned} \left\| \sum_i \lambda_i \mu_i - \sum_i \lambda_i \nu_i \right\|_{TV} &= \frac{1}{2} \sum_x \left| \sum_i \lambda_i (\mu_i(x) - \nu_i(x)) \right| \\ &\leq \sum_i \lambda_i \frac{1}{2} \sum_x |\mu_i(x) - \nu_i(x)| \\ &= \sum_i \lambda_i \|\mu_i - \nu_i\|_{\mathrm{TV}}. \end{aligned}$$

The composition identity is immediate from the definition of push-forward. For push-forward contraction, use the supremum characterization:

$$\|h_*\mu - h_*\nu\|_{\mathrm{TV}} = \sup_{B\subseteq Y} |\mu(h^{-1}(B)) - \nu(h^{-1}(B))| \le \sup_{A\subseteq X} |\mu(A) - \nu(A)| = \|\mu - \nu\|_{\mathrm{TV}}.$$

Kernel contraction follows from the ℓ^1 expression:

$$\begin{aligned}\|\mu K - \nu K\|_{\mathrm{TV}} &= \frac{1}{2}\sum_y \left|\sum_x (\mu(x) - \nu(x))K(y|x)\right| \\ &\le \frac{1}{2}\sum_x |\mu(x) - \nu(x)| \sum_y K(y|x) = \|\mu - \nu\|_{\mathrm{TV}}.\end{aligned}$$

Finally,

$$\|\mu K - \nu L\|_{\mathrm{TV}} \le \|\mu K - \nu K\|_{\mathrm{TV}} + \|\nu K - \nu L\|_{\mathrm{TV}}.$$

The first term is bounded by $\|\mu - \nu\|_{\mathrm{TV}}$ by kernel contraction, and the second by

$$\sum_x \nu(x)\|K(\cdot|x) - L(\cdot|x)\|_{\mathrm{TV}} \le \varepsilon$$

by convexity. This proves the perturbation bound. □

A.3 Deferred proofs

Proof of Lemma 1. We first check that ρ_G is well-defined. Suppose that $\sum_i \alpha_i M_{w_i}\xi = 0$. Then, for every suffix $v \in \Sigma^*$,

$$\sum_i \alpha_i h_{\mathcal{I}}(w_i)(v) = \sum_i \alpha_i F_{\mathcal{I}}(w_i v) = \sum_i \alpha_i \lambda(M_v M_{w_i}\xi) = \lambda\left(M_v \sum_i \alpha_i M_{w_i}\xi\right) = 0.$$

Hence $\sum_i \alpha_i h_{\mathcal{I}}(w_i) = 0$, so the assignment is well-defined. Linearity is immediate from the definition. It is surjective because $V_{\mathcal{I}}$ is spanned by the rows $h_{\mathcal{I}}(w)$. The identities $\rho_G\xi = \xi_{\mathcal{I}}$, $\rho_G M_\sigma = R^{\mathcal{I}}_\sigma \rho_G$ and $\lambda_{\mathcal{I}}\rho_G = \lambda$ follow directly. □

Proof of Proposition 1. Starting from $[\epsilon]_{\mathcal{I}}$, Eq. (6) reproduces the conditional output probabilities of $\mathcal{I}$ after every admissible history. Hence $E(\mathcal{I})$ implements $\mathcal{I}$. Its state after observing h is $[h]_{\mathcal{I}}$, so the residual future law is determined by the current state. Thus, it is predictive.

Now let T be a predictive transducer implementing $\mathcal{I}$. For each reachable $s \in \mathcal{S}$, choose an admissible history h such that $q_T(s \mid h) > 0$ and define

$$\phi_T(s) = [h]_{\mathcal{I}}. \tag{17}$$

This is well defined: if both h and h' are compatible with s, predictivity gives $\mathcal{I}^h = \mathcal{I}_{T,s} = \mathcal{I}^{h'}$. It is surjective because every admissible history has at least one state in the support of its posterior. Also, every state with positive initial probability is compatible with the empty history, so $(\phi_T)_* p = \delta_{[\epsilon]_{\mathcal{I}}}$.

Fix s and choose a compatible history h. Predictivity gives $\Pr_T(o \mid s, a) = \Pr_{\mathcal{I}^h}(o \mid a)$. Moreover, whenever $\kappa(s', o \mid s, a) > 0$, the state s' is compatible with the extended history

$h(a,o)$, and therefore $\phi_T(s') = [h(a,o)]_{\mathcal{I}}$. Thus all the probability mass associated with output o is pushed forward to the unique state prescribed by Eq. (6), and

$$(\phi_T \times \mathtt{id})_* \kappa(\cdot,\cdot \mid s,a) = \kappa_\epsilon(\cdot,\cdot \mid \phi_T(s),a).$$

Hence ϕ_T is a reduction. □

Proof of Theorem 1. Let $\phi : \mathcal{S}_1 \to \mathcal{S}_2$ be the map of the ε-reduction. Since a reduction preserves the input and output alphabets, let's write the common alphabets of both transducers as $\mathcal{A}$ and $\mathcal{O}$. Fix a section $r : \mathcal{S}_2 \to \mathcal{S}_1$ of ϕ, so that $\phi(r(u)) = u$ for every $u \in \mathcal{S}_2$.

For $\mathbf{a} \in \mathcal{A}^n$, define two probability distributions $P_{\mathbf{a}}$ and $Q_{\mathbf{a}}$ on $\mathcal{O}^n \times \mathcal{S}_2$ by

$$P_{\mathbf{a}}(\mathbf{o},u) = \sum_{s \in \phi^{-1}(u)} \Pr{}_{T_1}\big(O_{1:n} = \mathbf{o},\, S_n^1 = s \mid A_{1:n} = \mathbf{a}\big),$$
$$Q_{\mathbf{a}}(\mathbf{o},u) = \Pr{}_{T_2}\big(O_{1:n} = \mathbf{o},\, S_n^2 = u \mid A_{1:n} = \mathbf{a}\big),$$

where S_n^i is a random variable denoting the state of the transducer T_i at step n, $O_{1:n}$ denotes the first n observed outputs and $A_{1:n}$ the first n inputs. Thus, $P_{\mathbf{a}}$ is the joint law of the output prefix and the coarse-grained state $\phi(S_n^1)$ under T_1 assuming inputs $\mathbf{a}$, whereas $Q_{\mathbf{a}}$ is the corresponding joint law under T_2. We prove by induction on n that, for every $\mathbf{a} \in \mathcal{A}^n$,

$$\|P_{\mathbf{a}} - Q_{\mathbf{a}}\|_{\mathrm{TV}} \le (n+1)\varepsilon. \tag{18}$$

This is intuitive: initially the two distribution differ by at most ε because of the error in the initial distribution, and after each step this error increases by at most ε because the one-step transitions between T_1 and T_2 differ locally (i.e. when comparing $s \in \mathcal{S}_1$ with $\phi(s)$) by at most ε

For $n = 0$, the output prefix is empty and

$$P_\epsilon(\epsilon,u) = (\phi_* p_1)(u), \qquad Q_\epsilon(\epsilon,u) = p_2(u).$$

Consequently, the initial-distribution condition in the definition of an ε-reduction gives

$$\|P_\epsilon - Q_\epsilon\|_{\mathrm{TV}} = \|\phi_* p_1 - p_2\|_{\mathrm{TV}} \le \varepsilon.$$

Now fix $\mathbf{a} \in \mathcal{A}^n$, a next action $b \in \mathcal{A}$, and $(\mathbf{o},u) \in \mathcal{O}^n \times \mathcal{S}_2$. Define a probability distribution $\lambda_{\mathbf{a},\mathbf{o},u}$ on $\phi^{-1}(u)$ as follows. If $P_{\mathbf{a}}(\mathbf{o},u) > 0$, let

$$\lambda_{\mathbf{a},\mathbf{o},u}(s) = \frac{\Pr{}_{T_1}\big(O_{1:n} = \mathbf{o},\, S_n^1 = s \mid \mathcal{A}_{1:n} = \mathbf{a}\big)}{P_{\mathbf{a}}(\mathbf{o},u)}.$$

If $P_{\mathbf{a}}(\mathbf{o},u) = 0$, set arbitrarily $\lambda_{\mathbf{a},\mathbf{o},u} = \delta_{r(u)}$. Then, $\lambda_{\mathbf{a},\mathbf{o},u}(s)$ represents the probability for the state of transducer T_1 to be s at step n conditioned on T_1 being at a state in $\phi^{-1}(u)$.

For $\mathbf{o} \in \mathcal{O}^n$, let

$$j_{\mathbf{o}} : \mathcal{S}_2 \times \mathcal{O} \longrightarrow \mathcal{O}^{n+1} \times \mathcal{S}_2, \qquad j_{\mathbf{o}}(u',o) = (\mathbf{o}o, u'),$$

where $\mathbf{o}o$ denotes concatenation. Consider the Markov kernels $\widehat{K}_{1,\mathbf{a},b}$ and $\widehat{K}_{2,b}$ from $\mathcal{O}^n \times \mathcal{S}_2$ to $\mathcal{O}^{n+1} \times \mathcal{S}_2$ given by

$$\widehat{K}_{1,\mathbf{a},b}\big(\cdot \mid \mathbf{o},u\big) = (j_{\mathbf{o}})_* \left[\sum_{s \in \phi^{-1}(u)} \lambda_{\mathbf{a},\mathbf{o},u}(s)\, (\phi \times \mathrm{id}_{\mathcal{O}})_* \kappa_1(\cdot,\cdot \mid s,b) \right],$$
$$\widehat{K}_{2,b}\big(\cdot \mid \mathbf{o},u\big) = (j_{\mathbf{o}})_* \kappa_2(\cdot,\cdot \mid u,b).$$

The first kernel performs one step of T_1, coarse-grains the next state through ϕ, and retains the already observed output prefix. The second kernel performs the corresponding operation for T_2.

For every $(\mathbf{o}, u)$, push-forward contraction, convexity of total variation, and the one-step condition of the ε-reduction give

$$\begin{aligned}
&\left\| \widehat{K}_{1,\mathbf{a},b}(\cdot \mid \mathbf{o}, u) - \widehat{K}_{2,b}(\cdot \mid \mathbf{o}, u) \right\|_{\mathrm{TV}} \\
&\quad \le \left\| \sum_{s \in \phi^{-1}(u)} \lambda_{\mathbf{a},\mathbf{o},u}(s) (\phi \times \mathrm{id}_{\mathcal{O}})_* \kappa_1(\cdot, \cdot \mid s, b) - \kappa_2(\cdot, \cdot \mid u, b) \right\|_{\mathrm{TV}} \\
&\quad \le \sum_{s \in \phi^{-1}(u)} \lambda_{\mathbf{a},\mathbf{o},u}(s) \left\| (\phi \times \mathrm{id}_{\mathcal{O}})_* \kappa_1(\cdot, \cdot \mid s, b) - \kappa_2(\cdot, \cdot \mid \phi(s), b) \right\|_{\mathrm{TV}} \\
&\quad \le \varepsilon. \qquad (19)
\end{aligned}$$

Moreover, by construction of the conditional distributions $\lambda_{\mathbf{a},\mathbf{o},u}$ it follows that

$$P_{\mathbf{a}b} = P_{\mathbf{a}} \widehat{K}_{1,\mathbf{a},b}, \qquad Q_{\mathbf{a}b} = Q_{\mathbf{a}} \widehat{K}_{2,b}.$$

Thus, applying the kernel perturbation bound from Lemma 4 and then the induction hypothesis yields

$$\begin{aligned}
\| P_{\mathbf{a}b} - Q_{\mathbf{a}b} \|_{\mathrm{TV}} &\le \| P_{\mathbf{a}} - Q_{\mathbf{a}} \|_{\mathrm{TV}} + \sup_{(\mathbf{o},u)} \left\| \widehat{K}_{1,\mathbf{a},b}(\cdot \mid \mathbf{o}, u) - \widehat{K}_{2,b}(\cdot \mid \mathbf{o}, u) \right\|_{\mathrm{TV}} \\
&\le (n+1)\varepsilon + \varepsilon = (n+2)\varepsilon.
\end{aligned}$$

This completes the induction.

The output distribution $D_{\mathcal{I}_{T_1}}(\mathbf{a})$ is the marginal of $P_{\mathbf{a}}$ on $\mathcal{O}^n$, and $D_{\mathcal{I}_{T_2}}(\mathbf{a})$ is the corresponding marginal of $Q_{\mathbf{a}}$. Since marginalization contracts total variation,

$$\left\| D_{\mathcal{I}_{T_1}}(\mathbf{a}) - D_{\mathcal{I}_{T_2}}(\mathbf{a}) \right\|_{\mathrm{TV}} \le (n+1)\varepsilon$$

for every $\mathbf{a} \in \mathcal{A}^n$. Therefore

$$\begin{aligned}
d_\gamma(\mathcal{I}_{T_1}, \mathcal{I}_{T_2}) &= \sum_{n=0}^{\infty} \gamma^n \sup_{\mathbf{a} \in \mathcal{A}^n} \left\| D_{\mathcal{I}_{T_1}}(\mathbf{a}) - D_{\mathcal{I}_{T_2}}(\mathbf{a}) \right\|_{\mathrm{TV}} \\
&\le \varepsilon \sum_{n=0}^{\infty} (n+1)\gamma^n = \frac{\varepsilon}{(1-\gamma)^2}.
\end{aligned}$$

□

Proof of Proposition 2. Let $\langle \phi_1, f_1, g_1 \rangle$ be the ε_1-homomorphism from T_1 to T_2 and $\langle \phi_2, f_2, g_2 \rangle$ the ε_2-homomorphism from T_2 to T_3. Define the homomorphism $\langle \phi = \phi_2 \circ \phi_1, f = f_2 \circ f_1, g = g_2 \circ g_1 \rangle$ from T_1 to T_3. We will prove that this is a $(\varepsilon_1 + \varepsilon_2)$-homomorphism.

Fix $s \in \mathcal{S}_1$ and $a \in \mathcal{A}_1$. By the composition rule for push-forwards, the triangle inequality, and contraction of total variation under push-forwards,

$$\begin{aligned}
&\| (\phi \times g)_* \kappa_1(\cdot, \cdot \mid s, a) - \kappa_3(\cdot, \cdot \mid \phi(s), f(a)) \|_{\mathrm{TV}} \\
&\quad \le \| (\phi_2 \times g_2)_* \left((\phi_1 \times g_1)_* \kappa_1(\cdot, \cdot \mid s, a) - \kappa_2(\cdot, \cdot \mid \phi_1(s), f_1(a)) \right) \|_{\mathrm{TV}} \\
&\qquad + \| (\phi_2 \times g_2)_* \kappa_2(\cdot, \cdot \mid \phi_1(s), f_1(a)) - \kappa_3(\cdot, \cdot \mid \phi_2(\phi_1(s)), f_2(f_1(a))) \|_{\mathrm{TV}} \\
&\quad \le \varepsilon_1 + \varepsilon_2.
\end{aligned}$$

The initial distributions satisfy

$$\begin{aligned}\|\phi_* p_1 - p_3\|_{\mathrm{TV}} &\le \|(\phi_2)_*((\phi_1)_* p_1) - (\phi_2)_* p_2\|_{\mathrm{TV}} + \|(\phi_2)_* p_2 - p_3\|_{\mathrm{TV}} \\ &\le \|(\phi_1)_* p_1 - p_2\|_{\mathrm{TV}} + \|(\phi_2)_* p_2 - p_3\|_{\mathrm{TV}} \\ &\le \varepsilon_1 + \varepsilon_2.\end{aligned}$$

Thus $\langle \phi_2 \circ \phi_1, f_2 \circ f_1, g_2 \circ g_1 \rangle$ is an $(\varepsilon_1 + \varepsilon_2)$-homomorphism.

If both original maps are reductions, then the action and output maps are identities and ϕ_1, ϕ_2 are surjective. Hence $\phi_2 \circ \phi_1$ is surjective, so the composition is an $(\varepsilon_1 + \varepsilon_2)$-reduction. □

Proof of Proposition 3. For $i \in \{1, 2\}$, let $V_i = \mathbb{R}^{S_i}$ with the ℓ_1 norm, let $\xi_i = p_i$, and define

$$\lambda_i(x) = \sum_{s \in S_i} x_s, \qquad M^i_{a,o} e_s = \sum_{t \in S_i} \kappa_i(t, o \mid s, a) e_t.$$

Each $M^i_{a,o}$ is nonnegative and column-substochastic, and hence is an ℓ_1 contraction. Moreover, $\|\xi_i\|_1 = \|\lambda_i\|_{V_i^*} = 1$, so these linearizations are contractive.

Define $L_\phi e_s = e_{\phi(s)}$ and extend linearly. Since ϕ is surjective, so is L_ϕ, and $\|L_\phi\|_{1\to 1} = 1$. The initial-state condition of the ordinary reduction gives

$$\|L_\phi \xi_1 - \xi_2\|_1 = 2\|\phi_* p_1 - p_2\|_{\mathrm{TV}} \le 2\varepsilon.$$

For a fixed $s \in \mathcal{S}_1$ and $a \in \mathcal{A}$, the vector $(L_\phi M^1_{a,o} - M^2_{a,o} L_\phi) e_s$ is the o-component of the difference between the two joint laws on $\mathcal{S}_2 \times \mathcal{O}$ appearing in the definition of an ordinary ε-reduction. Namely,

$$\begin{aligned}(L_\phi M^1_{a,o} - M^2_{a,o} L_\phi) e_s &= \sum_{t_1 \in \mathcal{S}_1} \kappa_1(t_1, o|s, a) e_{\phi(t_1)} - \sum_{t_2 \in \mathcal{S}_2} \kappa_2(t_2, o|\phi(s), a) e_{t_2} \\ &= \sum_{t_2 \in \mathcal{S}_2} \left[\sum_{t_1 \in \phi^{-1}(t_2)} \kappa_1(t_1, o|s, a) - \kappa_2(t_2, o|\phi(s), a) \right] e_{t_2}\end{aligned}$$

Consequently,

$$\|(L_\phi M^1_{a,o} - M^2_{a,o} L_\phi) e_s\|_1 \le 2\varepsilon.$$

Taking the maximum over the columns gives

$$\|L_\phi M^1_{a,o} - M^2_{a,o} L_\phi\|_{1\to 1} \le 2\varepsilon.$$

Finally, $\lambda_2 L_\phi = \lambda_1$. Thus L_ϕ is a 2ε-linear reduction. The factor 2 shows up because of the normalization $\frac{1}{2}$ in total variation. □

Proof of Proposition 4. Write $G_i = (V_i, \xi_i, \lambda_i, \{M^i_{a,o}\}_{a,o})$. Since L and K are bounded and surjective, so is KL. For the initial vectors,

$$\begin{aligned}\|KL\xi_0 - \xi_2\| &\le \|K\| \, \|L\xi_0 - \xi_1\| + \|K\xi_1 - \xi_2\| \\ &\le \|K\|\varepsilon_1 + \varepsilon_2.\end{aligned}$$

For each symbol (a, o),

$$\begin{aligned}KLM^0_{a,o} - M^2_{a,o} KL = K(LM^0_{a,o} - M^1_{a,o} L) \\ + (KM^1_{a,o} - M^2_{a,o} K)L,\end{aligned}$$

and hence the operator norm of this difference is at most $\|K\|\varepsilon_1 + \|L\|\varepsilon_2$. Finally,

$$\lambda_2 KL - \lambda_0 = (\lambda_2 K - \lambda_1)L + (\lambda_1 L - \lambda_0),$$

whose dual norm is at most $\|L\|\varepsilon_2 + \varepsilon_1$. All three quantities are bounded by $c(K)\varepsilon_1 + c(L)\varepsilon_2$. □

Proof of Theorem 2. Let $L : V \to W$ be the ε-linear reduction. For $a = a_1 \cdots a_n \in \mathcal{A}^n$ and $o = o_1 \cdots o_n \in \mathcal{O}^n$, write $M_{a,o} = M_{a_n,o_n} \cdots M_{a_1,o_1}$ and $N_{a,o} = N_{a_n,o_n} \cdots N_{a_1,o_1}$. Symbol-wise contractivity implies $\|M_{a,o}\xi\| \leq 1$. We claim that

$$\|LM_{a,o}\xi - N_{a,o}\xi'\| \leq (n+1)\varepsilon. \tag{20}$$

For $n = 0$, this is the initial-vector condition. If the claim holds at length n and $\sigma = (a_{n+1}, o_{n+1})$, then

$$\begin{aligned}
&\|LM_\sigma M_{a,o}\xi - N_\sigma N_{a,o}\xi'\| \\
&\quad \leq \|(LM_\sigma - N_\sigma L)M_{a,o}\xi\| + \|N_\sigma(LM_{a,o}\xi - N_{a,o}\xi')\| \\
&\quad \leq \varepsilon + (n+1)\varepsilon.
\end{aligned}$$

This proves (20) by induction.

For every output word $o \in \mathcal{O}^n$,

$$\begin{aligned}
&|\lambda(M_{a,o}\xi) - \lambda'(N_{a,o}\xi')| \\
&\quad \leq |(\lambda - \lambda' L)(M_{a,o}\xi)| + |\lambda'(LM_{a,o}\xi - N_{a,o}\xi')| \\
&\quad \leq (n+2)\varepsilon.
\end{aligned}$$

Summing over the $|\mathcal{O}|^n$ output words and dividing by two gives the second quantity in the minimum in (12), the bound by one holds because both sides are probability distributions.

Equation (13) follows by summing the finite horizon bounds. For each fixed n, the remaining summand tends to zero with ε and is bounded by γ^n. Since $\sum_n \gamma^n < \infty$, we conclude that the right hand side converges to 0. □

Proof of Proposition 5. Fix $\mathcal{A} = \{\ell, g\}$ and $\mathcal{O} = \{\$, 0, 1\}$. Define $\mathcal{I}$ as follows: in the first step, the output is the symbol \$ with probability 1. Then, a fair coin is thrown, and the output is always 0 or 1 for all the next steps, depending on this coin. Thus, for every $a_1 \cdots a_k \in \mathcal{A}^k$ with $k \geq 2$, we have

$$\Pr_{\mathcal{I}}(\$ \mid a_1) = 1$$

and

$$\Pr_{\mathcal{I}}(\$0^{k-1} \mid a_1 \cdots a_k) = \Pr_{\mathcal{I}}(\$1^{k-1} \mid a_1 \cdots a_k) = \frac{1}{2},$$

Note that the interface is *independent* of the actions taken.

Consider the three-state transducer W with states $S_W = \{r, t_0, t_1\}$ and initial distribution centred at r and with kernel

$$\kappa_W(t_0, \$ \mid r, a) = \frac{1}{2}, \qquad \kappa_W(t_1, \$ \mid r, a) = \frac{1}{2},$$

$$\kappa_W(t_0, 0 \mid t_0, a) = 1, \qquad \kappa_W(t_1, 1 \mid t_1, a) = 1.$$

for every $a \in \mathcal{A}$. Clearly W implements $\mathcal{I}$, and it can be proven that there is no transducer with less than 3 states implementing this interface.

Now, let's define another implementation U. Its states are $S_U = \{r, c_{00}, c_{01}, c_{10}, c_{11}, z_0, z_1\}$ with initial distribution centered at r. We describe the kernel by steps. First, we state that

$$\kappa_U(c_{ij}, \$ \mid r, a) = \frac{1}{4} \qquad (i, j \in \{0, 1\}).$$

for every $a \in \mathcal{A}$. Namely, in the first step the transducer transitions with uniform probability to any of the states c_{ij}.

From c_{ij}, the action ℓ reads the first coordinate, while action g reads the second one. More precisely, we have

$$\kappa_U(z_i, i \mid c_{ij}, \ell) = 1, \qquad \kappa_U(z_j, j \mid c_{ij}, g) = 1.$$

Finally, for every $a \in \mathcal{A}$ we set

$$\kappa_U(z_0, 0 \mid z_0, a) = 1, \qquad \kappa_U(z_1, 1 \mid z_1, a) = 1.$$

It can be checked that the transducer U also implements $\mathcal{I}$: after the first output \$, the pair (i, j) is uniformly chosen; and whichever coordinate is read by the second action the final result is a fair bit. Afterwards, the machine moves to z_0 or z_1, where the same bit is repeated forever.

Now suppose, towards a contradiction, that some $T \in \mathcal{L}_{\mathcal{I}}^{0,d}$ receives a δ-reduction from every element of $\mathcal{L}_{\mathcal{I}}^{0,d}$, with $\delta < 1$. Since $W \in \mathcal{L}_{\mathcal{I}}^{0,d}$, there is a surjective state map $S_W \to S_T$. Hence $|S_T| \leq |S_W| = 3$, and the fact that any implementation of $\mathcal{I}$ must have at least three states implies that $|S_T| = 3$.

The three states of T can be labelled $x_\$, x_0, x_1$ depending on which node from W is the one mapped to them through the δ-reduction. Note that it must be the case that

$$\Pr{}_T(\star | x_\star, a) = 1$$

for every $\star \in \mathcal{O}$. To see this, first note that there must be some state which outputs \$ with probability one. Otherwise, it would be impossible for T to implement $\mathcal{I}$ exactly. With the same reasoning we can see that there must be some state that always outputs 0 and another one that always outputs 1. Then, we conclude that there is only one possibility for the reduction from W to T considering that $\delta < 1$.

Since $U \in \mathcal{L}_{\mathcal{I}}^{0,d}$, there is a δ-reduction $\psi : U \xrightarrow{\delta} T$. Consider the state $c_{01} \in S_U$. There are three possible images, and we go through them one by one.

If $\psi(c_{01}) = x_\$$ we reach an absurd, since the distributions between those states are at distance 1: c_{01} assigns 0 probability to outputting \$. If $\psi(c_{01}) = x_0$, then, under action g, the state c_{01} outputs 1 with probability one, while x_0 outputs 0 with probability one. Thus ψ is not a proper δ-reduction with $\delta < 1$. If $\psi(c_{01}) = x_1$, we can argue in the same way.

Therefore, no such transducer $T \in \mathcal{L}_{\mathcal{I}}^{0,d}$ exists. □

Proof of Proposition 6. Let W and U be the two exact implementations of $\mathcal{I}$ constructed in the proof of Proposition 5. Since $\mathcal{I}_W = \mathcal{I}_U = \mathcal{I}$, we have $W, U \in \mathcal{L}_{\mathcal{I}}^{\varepsilon, d_\infty}$ for every $\varepsilon \geq 0$. Thus it is enough to prove the following claim: if W δ-reduces to C and U δ-reduces to C, then $\delta \geq \frac{1}{2}$.

Let $\varphi : S_W \to S_C$ be the state map of a δ-reduction from W to C. Since ordinary reductions are surjective on states, we have $|S_C| \leq |S_W| = 3$.

First suppose that $|S_C| \leq 2$. The three states r, t_0, t_1 of W have one-step output marginals $\delta_\$$, δ_0 and δ_1 respectively, under every action. Since there are at most two states in C, two of r, t_0, t_1 must have the same image $x \in S_C$. Hence, for two distinct outputs $o \neq o'$, the output marginal $\mathrm{out}_C(x, a)$ is within total variation distance δ of both δ_o and $\delta_{o'}$. Marginalization cannot increase total variation, so

$$1 = \|\delta_o - \delta_{o'}\|_{\mathrm{TV}} \leq \|\delta_o - \mathrm{out}_C(x, a)\|_{\mathrm{TV}} + \|\mathrm{out}_C(x, a) - \delta_{o'}\|_{\mathrm{TV}} \leq 2\delta.$$

Thus $\delta \geq 1/2$.

It remains to consider the case $|S_C| = 3$, where φ is bijective. Write

$$x_\$ = \varphi(r), \qquad x_0 = \varphi(t_0), \qquad x_1 = \varphi(t_1).$$

as before. The reduction $W \overset{\delta}{\to} C$ implies that, for every action a,

$$\|\mathrm{out}_C(x_\$, a) - \delta_\$\|_{\mathrm{TV}} \leq \delta,$$

$$\|\mathrm{out}_C(x_0, a) - \delta_0\|_{\mathrm{TV}} \leq \delta, \qquad \|\mathrm{out}_C(x_1, a) - \delta_1\|_{\mathrm{TV}} \leq \delta.$$

Let $\psi : S_U \to S_C$ be the map of a δ-reduction from U to C. Consider the state $c_{01} \in S_U$. There are three possibilities.

If $\psi(c_{01}) = x_\$$, then under action ℓ, the state c_{01} outputs 0 with probability one. The reduction $U \overset{\delta}{\to} C$ gives

$$\|\delta_0 - \mathrm{out}_C(x_\$, \ell)\|_{\mathrm{TV}} \leq \delta.$$

Together with the estimate coming from φ, we have

$$\|\mathrm{out}_C(x_\$, \ell) - \delta_\$\|_{\mathrm{TV}} \leq \delta,$$

and then

$$1 = \|\delta_0 - \delta_\$\|_{\mathrm{TV}} \leq 2\delta.$$

If $\psi(c_{01}) = x_0$, then under action g, the state c_{01} outputs 1 with probability one. Hence

$$\|\delta_1 - \mathrm{out}_C(x_0, g)\|_{\mathrm{TV}} \leq \delta.$$

But x_0 is δ-close to a 0-state, so

$$\|\mathrm{out}_C(x_0, g) - \delta_0\|_{\mathrm{TV}} \leq \delta.$$

Therefore

$$1 = \|\delta_1 - \delta_0\|_{\mathrm{TV}} \leq 2\delta.$$

The last case can be treated in the same way. □

Proof of Lemma 2. For every $w, v \in \Sigma^*$ it is the case that $0 \leq h_{\mathcal{I}}(w)(v) = F_{\mathcal{I}}(wv) \leq 1$. Hence, if $r = \sum_{i=1}^m \alpha_i h_{\mathcal{I}}(w_i)$, then

$$\begin{aligned}
\|r\|_{\mathrm{pred}} &= \sup_{v \in \Sigma^*} \left| \sum_{i=1}^m \alpha_i h_{\mathcal{I}}(w_i)(v) \right| \\
&\leq \sum_{i=1}^m |\alpha_i| \sup_{v \in \Sigma^*} |F_{\mathcal{I}}(w_i v)| \\
&\leq \sum_{i=1}^m |\alpha_i|.
\end{aligned}$$

Taking the infimum over all atomic decompositions of r gives $\|r\|_{\mathrm{pred}} \le \|r\|_{\mathrm{at},\mathcal{I}}$.

It is straightforward to prove that $\|\lambda r\|_{\mathrm{at},\mathcal{I}} = |\lambda|\|r\|_{\mathrm{at},\mathcal{I}}$, and the triangle inequality is also easy to prove by concatenating atomic decompositions of the two summands. Finally, if $\|r\|_{\mathrm{at},\mathcal{I}} = 0$, the preceding inequality implies that $\|r\|_{\mathrm{pred}} = 0$. Thus $r(v) = 0$ for every $v \in \Sigma^*$, and hence $r = 0$. Therefore $\|\cdot\|_{\mathrm{at},\mathcal{I}}$ is a norm.

We now check that $\|\xi_{\mathcal{I}}\|_{\mathrm{at},\mathcal{I}} = 1$. Since $\xi_{\mathcal{I}} = h_{\mathcal{I}}(\epsilon)$ is itself an atom, $\|\xi_{\mathcal{I}}\|_{\mathrm{at},\mathcal{I}} \le 1$. On the other hand, we have $\|\xi_{\mathcal{I}}\|_{\mathrm{pred}} \ge |\xi_{\mathcal{I}}(\epsilon)| = F_{\mathcal{I}}(\epsilon) = 1$. Thus, using the previous shown relation between the atomic and predictive norms we conclude that $\|\xi_{\mathcal{I}}\|_{\mathrm{at},\mathcal{I}} = 1$.

For the readout functional, recall that $\lambda_{\mathcal{I}}(r) = r(\epsilon)$. Thus

$$|\lambda_{\mathcal{I}}(r)| = |r(\epsilon)| \le \|r\|_{\mathrm{pred}} \le \|r\|_{\mathrm{at},\mathcal{I}},$$

and consequently $\|\lambda_{\mathcal{I}}\|_{(V_{\mathcal{I}},\|\cdot\|_{\mathrm{at},\mathcal{I}})^*} \le 1$. Since $\lambda_{\mathcal{I}}(\xi_{\mathcal{I}}) = F_{\mathcal{I}}(\epsilon) = 1$ and $\|\xi_{\mathcal{I}}\|_{\mathrm{at},\mathcal{I}} = 1$ the reverse inequality also holds, and we obtain $\|\lambda_{\mathcal{I}}\|_{(V_{\mathcal{I}},\|\cdot\|_{\mathrm{at},\mathcal{I}})^*} = 1$.

Finally, we prove that the shift operators are nonexpansive in the atomic norm. For any atomic decomposition $r = \sum_{i=1}^m \alpha_i h_{\mathcal{I}}(w_i)$, the definition of the shift gives

$$R_\sigma^{\mathcal{I}} r = \sum_{i=1}^m \alpha_i h_{\mathcal{I}}(w_i\sigma).$$

Every $h_{\mathcal{I}}(w_i\sigma)$ is again a Hankel-row atom, and hence

$$\|R_\sigma^{\mathcal{I}} r\|_{\mathrm{at},\mathcal{I}} \le \sum_{i=1}^m |\alpha_i|.$$

Taking the infimum over all atomic decompositions of r proves $\|R_\sigma^{\mathcal{I}} r\|_{\mathrm{at},\mathcal{I}} \le \|r\|_{\mathrm{at},\mathcal{I}}$.

Using all the results, we can conclude that $G_{\mathcal{I}}^{\mathrm{at}}$ is a contractive transducer. □

Proof of Lemma 3. Because the vectors $h_{\mathcal{I}}(p_1), \dots, h_{\mathcal{I}}(p_d)$ form a basis of $V_{\mathcal{I}}$, every Hankel row of $\mathcal{I}$ is reconstructed from its values on the selected suffixes. Namely,

$$h_{\mathcal{I}}(w) = \mathrm{ev}_{\mathcal{I}}(h_{\mathcal{I}}(w)) C_{\mathcal{I}}^{-1} B_{\mathcal{I}}.$$

Consequently,

$$\Pi_{\mathcal{J}\to\mathcal{I}} h_{\mathcal{J}}(w) - h_{\mathcal{I}}(w) = e_w C_{\mathcal{I}}^{-1} B_{\mathcal{I}}$$

where $e_w = \mathrm{ev}_{\mathcal{J}}(h_{\mathcal{J}}(w)) - \mathrm{ev}_{\mathcal{I}}(h_{\mathcal{I}}(w))$. The condition $d_\infty(\mathcal{I},\mathcal{J}) \le \varepsilon$ implies $\|e_w\|_\infty \le \varepsilon$. Since the rows of $B_{\mathcal{I}}$ are atoms for $\|\cdot\|_{\mathrm{at},\mathcal{I}}$,

$$\begin{aligned}\|e_w C_{\mathcal{I}}^{-1} B_{\mathcal{I}}\|_{\mathrm{at},\mathcal{I}} &\le \|e_w C_{\mathcal{I}}^{-1}\|_1 \\ &\le \Gamma_{\mathcal{I}}\varepsilon.\end{aligned}$$

The images of $h_{\mathcal{J}}(p_1), \dots, h_{\mathcal{J}}(p_d)$ have coordinate matrix $C_{\mathcal{J}} C_{\mathcal{I}}^{-1}$ in the basis given by the rows of $B_{\mathcal{I}}$. If $C_{\mathcal{J}}$ is invertible, these images span $V_{\mathcal{I}}$, proving surjectivity. □

Proof of Theorem 3. Choose $\bar{\varepsilon}(\mathcal{I}) > 0$ so that $d\bar{\varepsilon}(\mathcal{I}) < \sigma_{\min}(C_{\mathcal{I}})$, where $\sigma_{\min}(C_{\mathcal{I}})$ denotes the minimal singular value of $C_{\mathcal{I}}$. Then, using standard perturbation arguments we may conclude that $C_{\mathcal{J}}$ is invertible, and Lemma 3 shows that $\Pi_{\mathcal{J}\to\mathcal{I}}$ is surjective. More precisely, note that, because $d_\infty(\mathcal{I},\mathcal{J}) \le \varepsilon$,

$$\|C_{\mathcal{J}} - C_{\mathcal{I}}\|_2 \le \|C_{\mathcal{J}} - C_{\mathcal{I}}\|_F \le d\varepsilon < \sigma_{\min}(C_{\mathcal{I}}),$$

and therefore, for every $x \neq 0$,

$$\begin{aligned}\|C_{\mathcal{J}}x\|_2 &\geq \|C_{\mathcal{I}}x\|_2 - \|(C_{\mathcal{J}} - C_{\mathcal{I}})x\|_2 \\ &\geq \sigma_{\min}(C_{\mathcal{I}})\|x\|_2 - \|C_{\mathcal{J}} - C_{\mathcal{I}}\|_2\|x\|_2 \\ &\geq (\sigma_{\min}(C_{\mathcal{I}}) - d\varepsilon)\|x\|_2 > 0.\end{aligned}$$

Applying Lemma 3 to the empty word gives

$$\|\Pi_{\mathcal{J}\to\mathcal{I}}\xi_{\mathcal{J}} - \xi_{\mathcal{I}}\|_{\mathrm{at},\mathcal{I}} \leq \Gamma_{\mathcal{I}}\varepsilon.$$

For $\sigma \in \Sigma$ and $w \in \Sigma^*$, let $E_w = \Pi_{\mathcal{J}\to\mathcal{I}}h_{\mathcal{J}}(w) - h_{\mathcal{I}}(w)$. Since $R^{\mathcal{I}}_\sigma h_{\mathcal{I}}(w) = h_{\mathcal{I}}(w\sigma)$, we have

$$\begin{aligned}&(\Pi_{\mathcal{J}\to\mathcal{I}}R^{\mathcal{J}}_\sigma - R^{\mathcal{I}}_\sigma\Pi_{\mathcal{J}\to\mathcal{I}})h_{\mathcal{J}}(w) \\ &\quad = E_{w\sigma} - R^{\mathcal{I}}_\sigma E_w.\end{aligned}$$

By Lemma 2, the shifts of $G_{\mathcal{I}}$ are non-expansive with respect to the atomic norm, so

$$\|(\Pi_{\mathcal{J}\to\mathcal{I}}R^{\mathcal{J}}_\sigma - R^{\mathcal{I}}_\sigma\Pi_{\mathcal{J}\to\mathcal{I}})h_{\mathcal{J}}(w)\|_{\mathrm{at},\mathcal{I}} \leq 2\Gamma_{\mathcal{I}}\varepsilon. \tag{21}$$

If $r = \sum_i \alpha_i h_{\mathcal{J}}(w_i)$, linearity and (21) give

$$\begin{aligned}&\|(\Pi_{\mathcal{J}\to\mathcal{I}}R^{\mathcal{J}}_\sigma - R^{\mathcal{I}}_\sigma\Pi_{\mathcal{J}\to\mathcal{I}})r\|_{\mathrm{at},\mathcal{I}} \\ &\quad \leq 2\Gamma_{\mathcal{I}}\varepsilon\sum_i |\alpha_i|.\end{aligned}$$

Taking the infimum over all atomic decompositions of r proves the required bound.

The choice $q_1 = \epsilon$ makes the readout exact. Indeed, the first column of $C_{\mathcal{I}}$ is $B_{\mathcal{I}}(\epsilon)$, and hence

$$\begin{aligned}\lambda_{\mathcal{I}}(\Pi_{\mathcal{J}\to\mathcal{I}}r) &= (\Pi_{\mathcal{J}\to\mathcal{I}}r)(\epsilon) \\ &= \mathrm{ev}_{\mathcal{J}}(r)C_{\mathcal{I}}^{-1}B_{\mathcal{I}}(\epsilon) \\ &= \mathrm{ev}_{\mathcal{J}}(r)e_1 = r(\epsilon) = \lambda_{\mathcal{J}}(r).\end{aligned}$$

Thus all three defects are bounded by $2\Gamma_{\mathcal{I}}\varepsilon$.

Finally, we show that $\Pi_{\mathcal{J}\to\mathcal{I}}$ is bounded. For every $w \in \Sigma^*$,

$$\|\Pi_{\mathcal{J}\to\mathcal{I}}h_{\mathcal{J}}(w)\|_{\mathrm{at},\mathcal{I}} \leq \|h_{\mathcal{I}}(w)\|_{\mathrm{at},\mathcal{I}} + \|\Pi_{\mathcal{J}\to\mathcal{I}}h_{\mathcal{J}}(w) - h_{\mathcal{I}}(w)\|_{\mathrm{at},\mathcal{I}} \leq 1 + \Gamma_{\mathcal{I}}\varepsilon.$$

Consequently, if $r = \sum_{i=1}^m \alpha_i h_{\mathcal{J}}(w_i)$, then

$$\|\Pi_{\mathcal{J}\to\mathcal{I}}r\|_{\mathrm{at},\mathcal{I}} \leq (1 + \Gamma_{\mathcal{I}}\varepsilon)\sum_{i=1}^m |\alpha_i|.$$

Taking the infimum over all atomic decompositions of r we conclude that $\|\Pi_{J\to I}\| \leq 1 + \Gamma_I\varepsilon$. □

Proof of Proposition 7. Let $\mathcal{O} = \{0,1\}$, $\mathcal{A} = \{\star\}$ (i.e. the dynamics are actionless) and let $\mathcal{I}$ be the process of independent fair bits. Write $D^{(m)}_{\mathcal{K}}$ for the length-m output law of an interface $\mathcal{K}$. For $n \geq 1$, define $\mathcal{J}_n$ as follows. Its first n outputs are independent fair bits. If

these outputs are 0^n, then every later output is 0; otherwise, all later outputs continue to be independent fair bits.

The length-m distributions agree for $m \le n$. For $m > n$ they differ only on strings beginning with 0^n, and a direct calculation gives

$$\left\| D_{\mathcal{I}}^{(m)} - D_{\mathcal{J}_n}^{(m)} \right\|_{\mathrm{TV}} = 2^{-n} - 2^{-m}.$$

Taking the supremum over m gives $d_\infty(\mathcal{I}, \mathcal{J}_n) = 2^{-n}$.

The transducer $\mathsf{E}(\mathcal{I})$ has one state, whose output law is $\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. In $\mathsf{E}(\mathcal{J}_n)$, the state reached after 0^n outputs 0 deterministically. Any state map to the one-state target must send this state to the unique state of $\mathsf{E}(\mathcal{I})$. Marginalizing the joint one-step kernels to outputs therefore gives

$$\delta \ge \left\| \delta_0 - \left(\tfrac{1}{2}\delta_0 + \tfrac{1}{2}\delta_1\right) \right\|_{\mathrm{TV}} = \frac{1}{2}.$$

For the last claim, surjectivity of a reduction from the one-state transducer $\mathsf{E}(\mathcal{I})$ forces C to have one state. Let ν be its output law. The two reductions imply

$$\left\| \nu - \left(\tfrac{1}{2}\delta_0 + \tfrac{1}{2}\delta_1\right) \right\|_{\mathrm{TV}} \le \delta, \qquad \|\nu - \delta_0\|_{\mathrm{TV}} \le \delta.$$

Then, the triangle inequality gives $1/2 \le 2\delta$. □

Proof of Theorem 4. Since $d_{\mathrm{res}}(\mathcal{I}, \mathcal{J}) \le \varepsilon < 1$, the definition of d_{res} implies that $\mathrm{supp}(\mathcal{I}) = \mathrm{supp}(\mathcal{J})$. Moreover, for every h in this common support we have $d_\infty(\mathcal{I}^h, \mathcal{J}^h) \le \varepsilon$.

We first prove that ϕ is well defined. Suppose that $[h]_{\mathcal{J}} = [u]_{\mathcal{J}}$. By definition of predictive equivalence, $\mathcal{J}^h = \mathcal{J}^u$. Therefore, by the triangle inequality,

$$\begin{aligned} d_\infty(\mathcal{I}^h, \mathcal{I}^u) &\le d_\infty(\mathcal{I}^h, \mathcal{J}^h) + d_\infty(\mathcal{J}^h, \mathcal{J}^u) + d_\infty(\mathcal{J}^u, \mathcal{I}^u) \\ &\le 2\varepsilon < \Delta_{\mathcal{I}}. \end{aligned}$$

If $[h]_{\mathcal{I}} \ne [u]_{\mathcal{I}}$, the definition of $\Delta_{\mathcal{I}}$ would instead give $d_\infty(\mathcal{I}^h, \mathcal{I}^u) \ge \Delta_{\mathcal{I}}$, which is a contradiction. Hence $[h]_{\mathcal{I}} = [u]_{\mathcal{I}}$, proving that ϕ is well defined.

The map is surjective. Indeed, every state of $E(\mathcal{I})$ is of the form $[h]_{\mathcal{I}}$ for some $h \in \mathrm{supp}(\mathcal{I})$. Since the interfaces have the same support, $h \in \mathrm{supp}(\mathcal{J})$, and therefore

$$[h]_{\mathcal{I}} = \phi([h]_{\mathcal{J}}).$$

It remains to verify the approximate one-step condition. For an interface K, write

$$\mu_K^h(o \mid a) = \Pr_{K^h}(o \mid a)$$

for the one-step output law after h. Fix $h \in \mathrm{supp}(\mathcal{I}) = \mathrm{supp}(\mathcal{J})$ and $a \in \mathcal{A}$. For every o having positive conditional probability, the canonical transducers move respectively to $[h(a,o)]_{\mathcal{J}}$ and $[h(a,o)]_{\mathcal{I}}$. By definition of ϕ, it holds that $\phi([h(a,o)]_{\mathcal{J}}) = [h(a,o)]_{\mathcal{I}}$. Furthermore, equality of supports implies that $\mu_{\mathcal{I}}^h(o \mid a) > 0$ if and only if $\mu_{\mathcal{J}}^h(o \mid a) > 0$. Consequently, after pushing the kernel of $E(\mathcal{J})$ forward through ϕ, both joint kernels place their mass corresponding to o on the same pair $\big([h(a,o)]_{\mathcal{I}}, o\big)$. It follows that

$$\begin{aligned} &\|(\phi \times \mathrm{id}_{\mathcal{O}})_* \kappa_{\epsilon,\mathcal{J}}(\cdot,\cdot \mid [h]_{\mathcal{J}}, a) - \kappa_{\epsilon,\mathcal{I}}(\cdot,\cdot \mid [h]_{\mathcal{I}}, a)\|_{\mathrm{TV}} \\ &\quad = \left\| \mu_{\mathcal{J}}^h(\cdot \mid a) - \mu_{\mathcal{I}}^h(\cdot \mid a) \right\|_{\mathrm{TV}} \\ &\quad \le d_\infty(\mathcal{J}^h, \mathcal{I}^h) \le \varepsilon. \end{aligned}$$

The initial state is preserved exactly:

$$\phi_* \delta_{[\epsilon]_{\mathcal{J}}} = \delta_{[\epsilon]_{\mathcal{I}}}.$$

Thus ϕ determines an ε-reduction $E(\mathcal{J}) \to E(\mathcal{I})$. □

Proof of Corollary 2. Let T be a predictive transducer from the statement. By Proposition 1 there is an exact reduction from T to $\mathsf{E}(\mathcal{J})$. Then, by Theorem 4 there is a ε-reduction from $E(\mathcal{J})$ to $E(\mathcal{I})$. Composing them we get the desired result, using Proposition 2 to bound the error of the composition. □

References

[1] David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *International Conference on Machine Learning*, pages 2915–2923. PMLR, 2016.

[2] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

[3] Nix Barnett and James P Crutchfield. Computational mechanics of input–output processes: Structured transformations and the ϵ-transducer. *Journal of Statistical Physics*, 161(2):404–451, 2015.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[5] Yakir Berchenko. Simplicity bias in overparameterized machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11052–11060, 2024.

[6] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

[7] Alexander Boyd, Franz Nowak, David Hyland, Manuel Baltieri, and Fernando E. Rosas. From monoliths to modules: Decomposing transducers for efficient world modelling, 2025.

[8] Rosa Cao and Daniel Yamins. Explanatory models in neuroscience: Part 2–constraint-based intelligibility. *arXiv preprint arXiv:2104.01489*, 2021.

[9] Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. Transferring linear features across language models with model stitching. *Advances in Neural Information Processing Systems*, 38:48531–48563, 2026.

[10] Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. *arXiv preprint arXiv:2411.05561*, 2024.

[11] Santiago Cifuentes. General agents contain world models, even under partial observability and stochasticity. *arXiv preprint arXiv:2602.03146*, 2026.

[12] James P Crutchfield. Inferring the dynamic, quantifying physical complexity. In *Measures of Complexity and Chaos*, pages 327–338. Springer, 1989.

[13] Adrián Csiszárik, Péter Kőrösi-Szabó, Akos Matszangosz, Gergely Papp, and Dániel Varga. Similarity and matching of neural network representations. *Advances in Neural Information Processing Systems*, 34:5656–5668, 2021.

[14] Josée Desharnais, Vineet Gupta, Radha Jagadeesan, and Prakash Panangaden. Metrics for labelled markov processes. *Theoretical computer science*, 318(3):323–354, 2004.

[15] Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in neural information processing systems*, 34:1556–1568, 2021.

[16] Norman Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*, pages 162–169. AUAI Press, 2004.

[17] Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003.

[18] Matt Gorbett and Suman Jana. Characterizing linear alignment across language models. *arXiv preprint arXiv:2603.18908*, 2026.

[19] Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

[20] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

[21] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[22] Wei Hu. Understanding surprising generalization phenomena in deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22669–22669, 2024.

[23] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

[24] Jiachen Jiang, Jinxin Zhou, Peng Wang, Qing Qu, Dustin Mixon, Chong You, and Zhihui Zhu. Generalized neural collapse for a large number of classes. *arXiv preprint arXiv:2310.05351*, 2023.

[25] Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

[26] John G Kemeny, J Laurie Snell, et al. *Finite markov chains*, volume 26. van Nostrand Princeton, NJ, 1969.

[27] Stefan Kiefer and Qiyi Tang. Approximate bisimulation minimisation. *arXiv preprint arXiv:2110.00326*, 2021.

[28] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

[29] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

[30] Kim Guldstrand Larsen and Arne Skou. Bisimulation through probabilistic testing. *Information and Computation*, 94(1):1–28, 1991.

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[32] Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. *AI&M*, 1(2):3, 2006.

[33] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015.

[34] George H Mealy. A method for synthesizing sequential circuits. *The Bell System Technical Journal*, 34(5):1045–1079, 1955.

[35] Mehryar Mohri. Finite-state transducers in language and speech processing. *Computational linguistics*, 23(2):269–311, 1997.

[36] Edward F Moore et al. Gedanken-experiments on sequential machines. *Automata studies*, 34(129-153):129–153, 1956.

[37] Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

[38] Aran Nayebi. What capable agents must know: Selection theorems for robust decision-making under uncertainty. *arXiv preprint arXiv:2603.02491*, 2026.

[39] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020.

[40] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[41] Balaraman Ravindran and Andrew G. Barto. SMDP homomorphisms: An algebraic approach to abstraction in semi-Markov decision processes. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 1011–1016, 2003.

[42] Balaraman Ravindran and Andrew G. Barto. Approximate homomorphisms: A framework for non-exact minimization in Markov Decision Processes. Manuscript, 2004.

[43] Jonathan Richens, David Abel, Alexis Bellot, and Tom Everitt. General agents contain world models, 2025. Accepted at ICML 2025.

[44] Fernando Rosas, Alexander Boyd, and Manuel Baltieri. Ai in a vat: Fundamental limits of efficient world modelling for agent sandboxing and interpretability. *arXiv preprint arXiv:2504.04608*, 2025.

[45] Marcel Paul Schützenberger. On the definition of a family of automata. *Inf. Control.*, 4(2-3):245–270, 1961.

[46] Adam Shai, Loren Amdahl-Culleton, Casper L Christensen, Henry R Bigelow, Fernando E Rosas, Alexander B Boyd, Eric A Alt, Kyle J Ray, and Paul M Riechers. Transformers learn factored representations. *arXiv preprint arXiv:2602.02385*, 2026.

[47] Adam S Shai, Sarah E Marzen, Lucas Teixeira, Alexander G Oldenziel, and Paul M Riechers. Transformers represent belief state geometry in their residual stream. *Advances in Neural Information Processing Systems*, 37:75012–75034, 2024.

[48] Cosma Rohilla Shalizi and James P Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of statistical physics*, 104(3):817–879, 2001.

[49] Timm Spork, Christel Baier, Joost-Pieter Katoen, Jakob Piribauer, and Tim Quatmann. A spectrum of approximate probabilistic bisimulations. In *35th International Conference on Concurrency Theory (CONCUR 2024)*, volume 311 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 37:1–37:19. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024.

[50] Jonathan J. Taylor, Doina Precup, and Prakash Panangaden. Bounding performance loss in approximate MDP homomorphisms. In *Advances in Neural Information Processing Systems 21*, pages 1649–1656, 2008.

[51] Guillermo Valle-Perez, Chico Q Camargo, and Ard A Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. *arXiv preprint arXiv:1805.08522*, 2018.

[52] Sumio Watanabe. *Mathematical theory of Bayesian statistics*. CRC press, 2018.