Stochastic Token Maps: Monad Liftings, Markov Categories, Convex Algebras, and Coalgebras over Token Space

Wuming Pan
College of Computer Science, Sichuan University
Chengdu, P.R. China, 610065
panwuming@scu.edu.cn

Revised preprint (V1.1), November 22, 2026

Abstract

Token maps, the morphisms of Token Space, are deterministic functions that preserve every structural record of a heap. Sampling a token from the output distribution of a language model is therefore not a Token map. This paper studies what a *stochastic* Token map should be. The answer depends on a choice, namely a lifting of the finitely supported distribution monad $\mathcal{D}$ along the forgetful functor from Token classes to sets. We describe three conservative liftings, called Dirac, independent and coupling, and show that the coupling lifting is the least lifting whose relations are convex. For each of the three the Kleisli category is a Markov category whose deterministic morphisms are exactly the Token maps, and the function objects of Token Space survive as stochastic function objects. Stochastic Token maps between causal sequence classes are the non-anticipating kernels, so positionwise sampling becomes a morphism. A kernel can preserve every Token locally without being a mixture of Token maps. The Eilenberg–Moore category of the coupling lifting consists of convex Token classes, which contain non-free objects such as Euclidean representation spaces. Final coalgebras lift from sets to Token classes, which gives a behavioural invariance principle for next-token predictors and a joint-sampling statement for generation. The object-level coupling construction is classical; the paper distinguishes the classical ingredients and lists open problems.

Keywords: Token Space; stochastic Token map; monad lifting; Markov category; convex algebra; Eilenberg–Moore category; coalgebra; language model.

1 Introduction

Token Space [1] describes AI computations with a record language. A *Token class* $T = \langle B, E, \Upsilon \rangle$ consists of a carrier B, a fixed set E of core symbols and a heap Υ of finite words over $B \sqcup E$, the Tokens. A *Token map* is a function between carriers that sends every Token of the source heap to a Token of the target heap. The resulting category TS_E is a quasitopos, and network layers, architectures and parameters are expressed inside it.

Token maps are deterministic and all-or-nothing. The framework says so explicitly: sampling a token from $\Delta(V)$ "is stochastic and is not a deterministic Token map," and randomness must be supplied as data or modelled by "a stochastic extension of the category" [1]. This draft develops such an extension.

A first attempt is to replace a function $f : B_1 \to B_2$ by a kernel $f : B_1 \to \mathcal{D}(B_2)$, where $\mathcal{D}(B_2)$ is the set of finitely supported probability distributions. This leaves the main question open:

what does it mean for a kernel to preserve a Token? Consider a Token (r, x, y) and a target heap of Tokens (r, u, v). The outputs at x and y are now random. One may require every pair of possible outputs to be related, or only that the two outputs can be sampled jointly so that they are related, or that the kernel is an average of deterministic Token maps. These requirements differ, and they give different categories.

We organise the possibilities with four standard categorical tools, each of which contributes a different part of the picture (Table 1).

Tool	Category	What it contributes
Monad lifting (Section 3)	a monad $\mathcal{D}^L$ on TS_E over $\mathcal{D}$	Fixes the meaning of "preserving a Token under randomness" by a heap on $\mathcal{D}(B)$.
Markov category (Sections 4, 5)	$\mathsf{Kl}(\mathcal{D}^L)$	Stochastic Token maps with composition, tensor, copying and discarding. The deterministic morphisms are the Token maps.
Convex algebras (Section 6)	$\mathsf{EM}(\mathcal{D}^{\mathrm{c}}) \cong \mathsf{CTS}_E$	Affine Token maps between convex Token classes. Contains non-free objects such as $\mathbb{R}^d$ and supports expectations.
Coalgebras (Section 7)	$\overline{H}$-coalgebras in TS_E	Next-token predictors, prefix trees and generation, with a coinductive invariance principle.

Table 1: The four perspectives studied in this paper.

Results. The statements proved below are the following.

(1) A lifting of a monad on **Set** to TS_E is a choice of heaps satisfying three conditions (Lemma 3.2). For every monad M the coupling construction M^{c} is a lifting (Theorem 3.5). For $\mathcal{D}$ there are three conservative liftings $\mathcal{D}^{\delta} \leq \mathcal{D}^{\times} \leq \mathcal{D}^{\mathrm{c}}$, and $\mathcal{D}^{\mathrm{c}}$ is the least lifting with convex relations (Proposition 3.7).

(2) A kernel between causal sequence classes is a stochastic Token map for the coupling lifting exactly when it is non-anticipating (Theorem 4.3). Positionwise sampling is such a map, whereas for the independent lifting only deterministic maps remain.

(3) Mixtures of Token maps are stochastic Token maps, but not conversely: there are stochastic Token maps between classes that admit no Token map at all (Proposition 4.6, Examples 4.7 and 4.8).

(4) For each of the three liftings, $\mathsf{Kl}(\mathcal{D}^L)$ is a Markov category whose deterministic subcategory is TS_E (Theorem 5.2), and the exponentials of TS_E give stochastic function objects whose global elements are the stochastic Token maps (Proposition 5.3).

(5) $\mathsf{EM}(\mathcal{D}^{\mathrm{c}})$ is the category of convex Token classes and affine Token maps (Theorem 6.2); expectations of Token maps into a convex Token class are Token maps (Proposition 6.4).

(6) Final coalgebras lift along $\mathsf{TS}_E \to$ **Set** (Theorem 7.1). For next-token predictors this gives an explicit behavioural heap (Proposition 7.2) and a joint-sampling property of generation (Proposition 7.4).

What is known. Most ingredients are classical. The relational and topological background of Token Space is surveyed in [2]; the present paper does not claim those ingredients as new. Lifting a functor to relations by couplings goes back to Barr [8, 25]. For a finite relational structure $\mathbf{A}$, the object $\mathcal{D}^{\mathrm{c}}(\mathbf{A})$ restricted to rational distributions is the structure $\mathrm{LP}(\mathbf{A})$ of constraint satisfaction, and rational-valued stochastic Token maps into it, together with their

local coupling witnesses, describe feasible solutions of the basic linear programming relaxation [24, 9]. Kleisli categories of commutative affine monads are the standard source of Markov categories [13]. Codensity liftings are due to Katsumata and Sato [21], classical barycentric and monadic accounts of convexity to Stone and Świrszcz [29, 31], and final coalgebras in fibrations to Hermida and Jacobs [18]. The contribution here is the assembly of these tools over TS_E, the comparison of liftings, and the reading of sampling, causal structure, attention and generation in this language.

Scope and verification. Proofs are included, and finite examples were checked by exact-arithmetic scripts (Appendix A). The checks do not replace the general proofs, and this preprint has not undergone independent peer review. The literature comparison is not an exhaustive priority claim. We work in ordinary set theory with the axiom of choice; the general-monad coupling proof uses a section of a surjection. Except for the explicitly countably supported monad in Remark 7.5, all probability distributions in this paper have finite support.

2 Preliminaries

2.1 Token classes and shapes

Fix a set E, the *core*, and put $F_E(B) = (B \sqcup E)^*$ and $F_E(f) = (f \sqcup \mathrm{id}_E)^*$. A *Token class* over E is $T = \langle B, E, \Upsilon\rangle$ with $\Upsilon \subseteq F_E(B)$. A *Token map* $T_1 \to T_2$ is a function $f : B_1 \to B_2$ with $F_E(f)(\Upsilon_1) \subseteq \Upsilon_2$. These form the category TS_E [1]. We write $U : \mathsf{TS}_E \to \mathbf{Set}$ for the faithful functor $U(T) = B$, $U(f) = f$.

A *shape* is a finite word s over $E \sqcup \{\square\}$; its *arity* $|s|$ is the number of blanks, and $s[b_1, \dots, b_n]$ denotes the word obtained by filling the blanks from left to right. Let Sh_E be the set of shapes. A heap is the same thing as a family of relations

$$R^T_s = \{(b_1, \dots, b_n) \in B^n : s[b_1, \dots, b_n] \in \Upsilon\}, \qquad s \in \mathrm{Sh}_E,\ n = |s|,$$

and a function f is a Token map iff $f^n(R^{T_1}_s) \subseteq R^{T_2}_s$ for every s. For $n = 0$ the relation is either $\varnothing$ or the one-element set B^0, a *nullary fact*, and the condition is an implication between facts. We write $\pi_i : R^T_s \to B$ for the i-th coordinate.

For a function $g : A \to B$ and heaps Υ_A on A, Υ_B on B, put

$$g^*\Upsilon_B = F_E(g)^{-1}(\Upsilon_B), \qquad g_*\Upsilon_A = F_E(g)(\Upsilon_A).$$

Lemma 2.1. *A function $g : A \to B$ is a Token map $\langle A, E, \Upsilon_A\rangle \to \langle B, E, \Upsilon_B\rangle$ iff $\Upsilon_A \subseteq g^*\Upsilon_B$, iff $g_*\Upsilon_A \subseteq \Upsilon_B$. The heaps on a fixed carrier form a complete lattice under inclusion, and g^* preserves arbitrary intersections. Hence U is a topological functor in the sense of [5].*

Proof. The equivalences restate the definition. Preimages preserve intersections. For a family of functions $g_j : A \to B_j$ into Token classes, $\bigcap_j g_j^*\Upsilon_{B_j}$ is the initial heap. □

Throughout this paper $\otimes$ denotes the categorical product of Token classes and its induced Kleisli tensor. This differs from the concatenation tensor over coproduct carriers studied in [3].

A Token map g is *induced* when $\Upsilon_A = g^*\Upsilon_B$. We use the following constructions of TS_E [1]. The terminal class is $1_E = \langle\{*\}, E, F_E(\{*\})\rangle$. The product $T_A \otimes T_B$ has carrier $A \times B$ and relations

$$R^{T_A \otimes T_B}_s \cong R^{T_A}_s \times R^{T_B}_s, \qquad \big((a_i, b_i)\big)_i \longmapsto \big((a_i)_i, (b_i)_i\big). \tag{1}$$

1id1:core:20261122-0005@v1 1id1.net

The exponential $T_B^{T_A}$ has carrier B^A, the set of all functions, and $(u_1,\dots,u_n) \in R_s^{T_B^{T_A}}$ iff $(u_1(a_1),\dots,u_n(a_n)) \in R_s^{T_B}$ for every $(a_1,\dots,a_n) \in R_s^{T_A}$. Thus U preserves finite products and exponentials strictly. The class with carrier B and empty heap is denoted $D_E(B)$, and the class with the full heap $C_E(B)$.

2.2 The distribution monad

For a set X let $\mathcal{D}(X)$ be the set of functions $p : X \to [0,1]$ with finite support $\operatorname{supp} p$ and $\sum_x p(x) = 1$. For $f : X \to Y$ put $\mathcal{D}(f)(p)(y) = \sum_{f(x)=y} p(x)$. With the Dirac distributions $\delta_X : X \to \mathcal{D}(X)$ and the averaging maps $\mu_X : \mathcal{D}\mathcal{D}(X) \to \mathcal{D}(X)$, $\mu(P) = \sum_q P(q)\, q$, this is a monad on **Set**. For finite V, $\mathcal{D}(V)$ is the simplex written $\Delta(V)$ in [1]. The product distribution $p \otimes q \in \mathcal{D}(X \times Y)$ defines natural maps $\nabla_{X,Y} : \mathcal{D}(X) \times \mathcal{D}(Y) \to \mathcal{D}(X \times Y)$.

A morphism $A \to B$ of the Kleisli category $\mathsf{Kl}(\mathcal{D})$ is a *kernel* $f : A \to \mathcal{D}(B)$; the composite of f with $g : B \to \mathcal{D}(C)$ is $(g \bullet f)(a) = \sum_b f(a)(b)\, g(b)$.

For a general monad (M, η, μ) on **Set** we use two standard notions. M is *commutative* when it carries natural maps $\nabla_{X,Y} : MX \times MY \to M(X \times Y)$ making it a symmetric monoidal monad for the cartesian structure [22, 23], and *affine* when $M1 \cong 1$. The monad $\mathcal{D}$ is commutative and affine. The associated *strength* is $t_{X,Y} = \nabla_{X,Y} \circ (\eta_X \times \mathrm{id}) : X \times MY \to M(X \times Y)$.

2.3 Markov categories

A *Markov category* [13, 10] is a symmetric monoidal category in which every object X carries a commutative comonoid $(\mathrm{copy}_X : X \to X \otimes X,\ \mathrm{del}_X : X \to I)$ compatible with the tensor product, and del is natural, so that the unit I is terminal. A morphism f is *deterministic* when $\mathrm{copy} \circ f = (f \otimes f) \circ \mathrm{copy}$. Deterministic morphisms form a cartesian monoidal subcategory. If M is a commutative affine monad on a category with finite products, then $\mathsf{Kl}(M)$ is a Markov category, with tensor given by the product on objects and comonoids inherited along the canonical functor [13, Corollary 3.2].

3 Liftings of a monad to Token Space

3.1 Liftings as choices of heaps

Definition 3.1. Let (M, η, μ) be a monad on **Set**. A *lifting* of M to TS_E is a monad $(\overline{M}, \overline{\eta}, \overline{\mu})$ on TS_E with $U\overline{M} = MU$, $U\overline{\eta} = \eta U$ and $U\overline{\mu} = \mu U$.

Lemma 3.2 (Lifting criterion)**.** *Liftings of M correspond bijectively to assignments L of a heap $L(T) \subseteq F_E(MB)$ to every Token class $T = \langle B, E, \Upsilon\rangle$ such that, writing $M^L T = \langle MB, E, L(T)\rangle$,*

(L1) *if $f : T_1 \to T_2$ is a Token map, then $Mf : M^L T_1 \to M^L T_2$ is a Token map;*

(L2) *$\eta_B : T \to M^L T$ is a Token map;*

(L3) *$\mu_B : M^L M^L T \to M^L T$ is a Token map.*

Proof. A lifting determines L. Conversely (L1) makes M^L a functor with $UM^L = MU$, and (L2), (L3) supply the components of $\overline{\eta}$ and $\overline{\mu}$. Naturality and the monad equations hold in **Set**, and U is faithful, so they hold in TS_E. □

We identify a lifting with L and write $R_s^{L(T)}$ for the relations of $L(T)$. Liftings are ordered by $L \le L'$ iff $L(T) \subseteq L'(T)$ for all T.

1id1:core:20261122-0005@v1 1id1.net

Definition 3.3. A lifting L is *conservative* when every unit $\eta_B : T \to M^L T$ is induced, that is, $\eta_B^* L(T) = \Upsilon$.

Conservativity says that the lifted class introduces no new Tokens among the deterministic points. The assignment $L(T) = F_E(MB)$ is always a lifting, and for $M = \mathcal{D}$ it is not conservative. This is the reason for the condition.

3.2 Three liftings of the distribution monad

Fix $T = \langle B, E, \Upsilon \rangle$ and a shape s of arity n. For a monad M define the *marginal map*

$$\mathrm{m}_s^T = \langle M\pi_1, \ldots, M\pi_n \rangle : M(R_s^T) \longrightarrow (MB)^n.$$

For $M = \mathcal{D}$, the tuple $\mathrm{m}_s^T(\omega)$ lists the marginals of a distribution ω on R_s^T, and ω is called a *coupling* of these marginals supported in R_s^T.

Definition 3.4. The *unit lifting* M^η and the *coupling lifting* M^{c} of a monad M are given by

$$R_s^{M^\eta T} = \eta_B^n(R_s^T), \qquad R_s^{M^{\mathrm{c}} T} = \mathrm{m}_s^T\big(M(R_s^T)\big).$$

For $M = \mathcal{D}$ we write $\mathcal{D}^\delta = \mathcal{D}^\eta$ and call it the *Dirac lifting*. The *independent lifting* $\mathcal{D}^\times$ is given by

$$R_s^{\mathcal{D}^\times T} = \{(p_1, \ldots, p_n) \in \mathcal{D}(B)^n : \ \operatorname{supp} p_1 \times \cdots \times \operatorname{supp} p_n \subseteq R_s^T\}.$$

For a nullary shape all three constructions for $\mathcal{D}$ keep the fact unchanged: $R_s^{\mathcal{D}^L T}$ is nonempty iff R_s^T is. Informally, a tuple of distributions is a Token of $\mathcal{D}^\delta T$ when it is a Token of T viewed as Dirac distributions; of $\mathcal{D}^\times T$ when *every* choice of outcomes is a Token of T; and of $\mathcal{D}^{\mathrm{c}} T$ when the outcomes can be sampled *jointly* so that the result is always a Token of T.

Theorem 3.5 (Coupling lifting)**.** *For every monad M on* **Set***, the assignments M^η and M^{c} are liftings of M to TS_E.*

Proof. We verify the conditions of Lemma 3.2 shape by shape.

Unit lifting. (L1) follows from $Mf \circ \eta = \eta \circ f$, (L2) holds by definition, and (L3) from $\mu_B \circ \eta_{MB} = \mathrm{id}$.

Coupling lifting, (L1). A Token map $f : T_1 \to T_2$ restricts to $f_s : R_s^{T_1} \to R_s^{T_2}$ with $\pi_i \circ f_s = f \circ \pi_i$. Hence $(Mf)^n \circ \mathrm{m}_s^{T_1} = \mathrm{m}_s^{T_2} \circ M(f_s)$, and the image of $\mathrm{m}_s^{T_1}$ is sent into the image of $\mathrm{m}_s^{T_2}$.

(L2). For $x \in R_s^T$ we have $M\pi_i(\eta(x)) = \eta(x_i)$, so $\eta_B^n(x) = \mathrm{m}_s^T(\eta(x))$.

(L3). Put $W = R_s^{M^{\mathrm{c}} T} \subseteq (MB)^n$ with coordinates $\rho_i : W \to MB$. The map $\mathrm{m}_s^T : M(R_s^T) \to W$ is surjective; choose a section σ, so $M\pi_i \circ \sigma = \rho_i$. A Token of $M^{\mathrm{c}} M^{\mathrm{c}} T$ of shape s has the form $(M\rho_i(\Omega))_i$ with $\Omega \in M(W)$. Put $\omega = \mu(M\sigma(\Omega)) \in M(R_s^T)$. By naturality of μ,

$$M\pi_i(\omega) = \mu_B\big(M(M\pi_i \circ \sigma)(\Omega)\big) = \mu_B\big(M\rho_i(\Omega)\big).$$

Thus μ_B^n sends the given Token to $\mathrm{m}_s^T(\omega) \in W$. □

For $M = \mathcal{D}$ the section σ is needed only on the finite support of Ω, and the proof reads as follows: a mixture of tuples that each have a coupling in R_s^T is coupled by the corresponding mixture of couplings.

Proposition 3.6 (Independent lifting)**.** *$\mathcal{D}^\times$ is a lifting of $\mathcal{D}$.*

1id1:core:20261122-0005@v1 1id1.net

Proof. (L1) holds because $\operatorname{supp} \mathcal{D}f(p) = f(\operatorname{supp} p)$, and (L2) because $\operatorname{supp} \delta_x = \{x\}$. For (L3) let $(P_1, \dots, P_n)$ be a Token of $\mathcal{D}^\times \mathcal{D}^\times T$ and $x_i \in \operatorname{supp} \mu(P_i)$. Since $\operatorname{supp} \mu(P_i) = \bigcup_{q \in \operatorname{supp} P_i} \operatorname{supp} q$, there are $q_i \in \operatorname{supp} P_i$ with $x_i \in \operatorname{supp} q_i$. The tuple $(q_i)_i$ is a Token of $\mathcal{D}^\times T$, so $(x_i)_i \in R_s^T$. □

Proposition 3.7 (Comparison)**.** (a) *$\mathcal{D}^\delta \le \mathcal{D}^\times \le \mathcal{D}^{\mathrm{c}}$, and all three are conservative.*

(b) *$R_s^{\mathcal{D}^{\mathrm{c}} T}$ is the convex hull of $\delta_B^n(R_s^T)$ in $\mathcal{D}(B)^n$, where mixtures are taken componentwise.*

(c) *M^η is the least lifting of M. $\mathcal{D}^{\mathrm{c}}$ is the least lifting L of $\mathcal{D}$ for which every $R_s^{L(T)}$ is convex.*

(d) *Both inequalities in* (a) *are strict.*

Proof. (a) If $x \in R_s^T$ then $\prod_i \operatorname{supp} \delta_{x_i} = \{x\} \subseteq R_s^T$. If $\prod_i \operatorname{supp} p_i \subseteq R_s^T$ and $n \ge 1$, the product distribution is a coupling supported in R_s^T; nullary shapes were treated above. Conservativity of $\mathcal{D}^{\mathrm{c}}$: a coupling whose marginals are $\delta_{x_1}, \dots, \delta_{x_n}$ is δ_x, so $x \in R_s^T$. The other two cases are immediate.

(b) A coupling $\omega = \sum_x \omega(x) \delta_x$ has marginals $\sum_x \omega(x) \delta_{x_i}$, so $\mathrm{m}_s^T(\omega) = \sum_x \omega(x)\, \delta_B^n(x)$, and every finite convex combination of points $\delta_B^n(x)$, $x \in R_s^T$, arises in this way.

(c) Every lifting contains M^η by (L2). A lifting of $\mathcal{D}$ with convex relations contains the convex hulls of the relations of $\mathcal{D}^\delta$, which is $\mathcal{D}^{\mathrm{c}}$ by (b); and $\mathcal{D}^{\mathrm{c}}$ has convex relations.

(d) These examples work for every core, including $E = \varnothing$. For $T = \langle \{0,1\}, E, \{(0),(1)\} \rangle$ the relation of shape $\square$ in $\mathcal{D}^\delta T$ consists of the two Dirac distributions and that in $\mathcal{D}^\times T$ of all distributions. For $T = \langle \{0,1\}, E, \{(0,0),(1,1)\} \rangle$ the relation of shape $\square\square$ in $\mathcal{D}^\times T$ is $\{(\delta_0, \delta_0), (\delta_1, \delta_1)\}$, while that in $\mathcal{D}^{\mathrm{c}} T$ is $\{(p,p) : p \in \mathcal{D}\{0,1\}\}$. □

The second example in (d) shows the role of repeated carrier elements. An equality Token (x, x) is preserved by every kernel under the coupling lifting, because a distribution can be coupled with itself along the diagonal. Under the independent lifting two independent samples would have to agree, which forces a Dirac distribution.

3.3 Codensity liftings

A second family of liftings is defined by testing against Token maps into a fixed class. Let (Ω, a) be an Eilenberg–Moore algebra of M and $\mathcal{O} = \langle \Omega, E, \Upsilon_\Omega \rangle$ any Token class on its carrier. For a function $k : B \to \Omega$ write $k^\dagger = a \circ Mk : MB \to \Omega$.

Proposition 3.8 (Codensity lifting)**.** *The assignment*

$$M^{\mathcal{O}}(T) = \bigcap_{k \in \mathsf{TS}_E(T, \mathcal{O})} (k^\dagger)^* \Upsilon_\Omega$$

is a lifting of M. If $M = \mathcal{D}$ and $a : \mathcal{D}^{\mathrm{c}}\mathcal{O} \to \mathcal{O}$ is a Token map, then $\mathcal{D}^{\mathrm{c}} \le \mathcal{D}^{\mathcal{O}}$.

Proof. (L1): for Token maps $f : T_1 \to T_2$ and $k : T_2 \to \mathcal{O}$ we have $k^\dagger \circ Mf = (k \circ f)^\dagger$, and $k \circ f$ is one of the maps tested for T_1. (L2): $k^\dagger \circ \eta = k$. (L3): by definition every $k^\dagger : M^{\mathcal{O}} T \to \mathcal{O}$ is a Token map, so a Token w of $M^{\mathcal{O}} M^{\mathcal{O}} T$ satisfies $F_E((k^\dagger)^\dagger)(w) \in \Upsilon_\Omega$; and $(k^\dagger)^\dagger = a \circ Ma \circ MMk = a \circ \mu_\Omega \circ MMk = k^\dagger \circ \mu_B$.

For the comparison, let $k : T \to \mathcal{O}$ be a Token map. Then $\mathcal{D}k : \mathcal{D}^{\mathrm{c}} T \to \mathcal{D}^{\mathrm{c}}\mathcal{O}$ is a Token map by Theorem 3.5, and so is $k^\dagger = a \circ \mathcal{D}k$. Hence every Token of $\mathcal{D}^{\mathrm{c}} T$ lies in $(k^\dagger)^* \Upsilon_\Omega$. □

This is the codensity lifting of [21] specialised to the topological functor U. It need not be conservative.

1id1:core:20261122-0005@v1 1id1.net

Example 3.9 (Stochastic dominance)**.** Let $\sqsubseteq$ be a preorder on B, let $P = \langle B, E, \{(\leq, x, y) : x \sqsubseteq y\}\rangle$, and let $I = \langle [0,1], E, \{(\leq, a, b) : a \leq b\}\rangle$ with the expectation map $\mathcal{D}[0,1] \to [0,1]$. Token maps $P \to I$ are monotone functions, so (p, q) is related in $\mathcal{D}^I P$ iff $\mathbb{E}_p k \leq \mathbb{E}_q k$ for every monotone $k : B \to [0,1]$. This is stochastic dominance. By Proposition 3.8 it follows from the existence of a coupling supported in $\sqsubseteq$. The converse is the finite case of Strassen's theorem [30]: a coupling of p and q supported in a relation R exists iff $p(S) \leq q(R[S])$ for every $S \subseteq \operatorname{supp} p$. To see this directly, use a finite flow network with source capacities $p(x)$, target capacities $q(y)$, and capacity 1 on the allowed edges xRy, restricting both sides to the two supports. A cut that crosses an allowed edge has capacity at least 1; every other cut has capacity at least $1 - p(S) + q(R[S])$. The max-flow/min-cut theorem therefore gives a flow of mass 1, which is the desired coupling. For a preorder the required inequality follows by applying dominance to the indicator of the up-set $R[S]$, since $S \subseteq R[S]$. Hence $\mathcal{D}^I P = \mathcal{D}^{\mathrm{c}} P$. In general the two liftings differ, in the same way as the Kantorovich and Wasserstein liftings of [7].

4 Stochastic Token maps

Definition 4.1. Let L be a lifting of $\mathcal{D}$. An *L-stochastic Token map* $T_1 \rightsquigarrow T_2$ is a morphism of $\mathsf{Kl}(\mathcal{D}^L)$, that is, a kernel $f : B_1 \to \mathcal{D}(B_2)$ that is a Token map $T_1 \to \mathcal{D}^L T_2$. For $L = \mathrm{c}$ we say *stochastic Token map*.

Explicitly, f is a stochastic Token map iff for every Token $s[x_1, \ldots, x_n]$ of T_1 there is a distribution on $R_s^{T_2}$ whose marginals are $f(x_1), \ldots, f(x_n)$, and every nullary fact of T_1 holds in T_2. It is $\mathcal{D}^{\times}$-stochastic iff $\prod_i \operatorname{supp} f(x_i) \subseteq R_s^{T_2}$ for every such Token, and $\mathcal{D}^{\delta}$-stochastic iff f is Dirac on every element occurring in a Token and the resulting partial function preserves positive-arity Tokens, with nullary facts also preserved. Composition is composition of kernels. By Proposition 3.7(a) there are faithful functors

$$\mathsf{TS}_E \longrightarrow \mathsf{Kl}(\mathcal{D}^{\delta}) \longrightarrow \mathsf{Kl}(\mathcal{D}^{\times}) \longrightarrow \mathsf{Kl}(\mathcal{D}^{\mathrm{c}}) \longrightarrow \mathsf{Kl}(\mathcal{D}),$$

where the first three are identity on Token classes, the first sends g to $\delta \circ g$, and the last forgets to the carrier via U.

Example 4.2. (i) If T_1 has the empty heap, every kernel is a stochastic Token map, so $\mathsf{Kl}(\mathcal{D})$ embeds fully in $\mathsf{Kl}(\mathcal{D}^{\mathrm{c}})$ through D_E. (ii) Let T_i record a preorder as in Example 3.9. A stochastic Token map is a kernel with $x \sqsubseteq y \Rightarrow f(x) \leq_{\mathrm{st}} f(y)$. (iii) Let T_i record the graph of a binary operation by Tokens $(+, x, y, z)$, as in algebraic tokenization [1]. A kernel is a stochastic Token map iff for all x, y there are random variables $Y_1 \sim f(x)$, $Y_2 \sim f(y)$, $Y_3 \sim f(x+y)$ on a common probability space with $Y_3 = Y_1 + Y_2$ almost surely.

4.1 Causal sequence classes and sampling

Recall the causal sequence class of [1, Section 12.3],

$$\mathsf{Cau}_n(X) = \big\langle X^n, E_n^{\mathrm{pre}}, \{(c_k, s, t) : s_{1:k} = t_{1:k},\ 1 \leq k \leq n\}\big\rangle,$$

whose Token maps are the functions F with $F(s)_i = f_i(s_{1:i})$. Let $\pi_{n,k} : X^n \to X^k$ keep the first k coordinates.

Theorem 4.3 (Stochastic causal maps)**.** *Let X be nonempty. A kernel $F : X^n \to \mathcal{D}(Y^n)$ is a stochastic Token map $\mathsf{Cau}_n(X) \rightsquigarrow \mathsf{Cau}_n(Y)$ iff for every k the law of the first k outputs depends only on the first k inputs, that is, iff there are kernels $F_k : X^k \to \mathcal{D}(Y^k)$ with*

$$\mathcal{D}(\pi_{n,k}) \circ F = F_k \circ \pi_{n,k} \qquad (1 \leq k \leq n).$$

1id1:core:20261122-0005@v1 1id1.net

A kernel is $\mathcal{D}^\times$-stochastic between these classes iff it is $\delta \circ G$ for a Token map $G : \mathsf{Cau}_n(X) \to \mathsf{Cau}_n(Y)$.

Proof. Let $s_{1:k} = t_{1:k}$, and put $p = F(s)$, $p' = F(t)$. A coupling of p and p' supported in $\{(y, y') : y_{1:k} = y'_{1:k}\}$ has equal images under the two maps to $\mathcal{D}(Y^k)$, so $\mathcal{D}(\pi_{n,k})(p) = \mathcal{D}(\pi_{n,k})(p')$. Conversely, if both equal $r \in \mathcal{D}(Y^k)$, then

$$\omega(y, y') = \frac{p(y)\, p'(y')}{r(u)} \quad \text{when } y_{1:k} = y'_{1:k} = u,\ r(u) > 0,$$

and $\omega(y, y') = 0$ otherwise, is such a coupling: summing over y' with $y'_{1:k} = u$ gives $p(y)$. So preservation of (c_k, s, t) is equivalent to equality of the two prefix laws, which defines F_k because every prefix extends. For the last assertion apply the independent condition to (c_n, s, s): all pairs of points in $\operatorname{supp} F(s)$ must be equal, so $F(s)$ is a Dirac distribution, and $\mathcal{D}^\times$ is conservative. □

Corollary 4.4 (Sampling is a stochastic Token map)**.** *For a finite vocabulary V, positionwise sampling*

$$\mathrm{samp}_n : \Delta(V)^n \to \mathcal{D}(V^n), \qquad (p_1, \dots, p_n) \mapsto p_1 \otimes \cdots \otimes p_n,$$

is a stochastic Token map $\mathsf{Cau}_n(\Delta(V)) \rightsquigarrow \mathsf{Cau}_n(V)$. Its composite with a deterministic causal model $\mathsf{Cau}_n(V) \to \mathsf{Cau}_n(\Delta(V))$ is a stochastic Token map $\mathsf{Cau}_n(V) \rightsquigarrow \mathsf{Cau}_n(V)$.

Proof. The law of the first k outputs is $p_1 \otimes \cdots \otimes p_k$. The second statement is closure under composition. □

This gives one precise form of the stochastic extension asked for in [1]. The theorem also shows why the independent lifting is too strict for sequence models: it admits no randomness at all between causal classes. Autoregressive generation, in which a sampled token is fed back, is treated in Section 7.

4.2 Mixtures, and local versus global preservation

Another candidate notion is a random choice of a deterministic Token map.

Definition 4.5. A *mixture* of Token maps $T_1 \to T_2$ is a distribution $\rho \in \mathcal{D}(\mathsf{TS}_E(T_1, T_2))$. Its kernel is $[\rho](x) = \sum_g \rho(g)\, \delta_{g(x)}$. Let $\mathsf{Mix}(T_1, T_2)$ be the set of kernels of mixtures.

Proposition 4.6. (a) *Every kernel of a mixture is a stochastic Token map, and* Mix *is closed under composition, so it is a subcategory of $\mathsf{Kl}(\mathcal{D}^c)$ containing the image of TS_E.*

(b) *If B_1 is finite, every $\mathcal{D}^\times$-stochastic Token map $T_1 \rightsquigarrow T_2$ is the kernel of a mixture.*

(c) *For finite B_1 there are inclusions*

$$\mathsf{Kl}(\mathcal{D}^\times)(T_1, T_2) \subseteq \mathsf{Mix}(T_1, T_2) \subseteq \mathsf{Kl}(\mathcal{D}^c)(T_1, T_2).$$

Each inclusion can be strict. Without the finiteness hypothesis, the first inclusion can fail.

Proof. (a) For a Token $s[x_1, \dots, x_n]$ of T_1, the distribution $\sum_g \rho(g)\, \delta_{(g(x_1), \dots, g(x_n))}$ is supported in $R_s^{T_2}$ and has marginals $[\rho](x_i)$. For composition, $[\sigma] \bullet [\rho]$ is the kernel of the image of $\sigma \otimes \rho$ under composition of Token maps.

(b) Let f be $\mathcal{D}^\times$-stochastic and let $\rho = \bigotimes_{x \in B_1} f(x) \in \mathcal{D}(B_2^{B_1})$. Every g in its support satisfies $g(x) \in \operatorname{supp} f(x)$, so for a Token $s[x_1, \dots, x_n]$ the tuple $(g(x_i))_i$ lies in $\prod_i \operatorname{supp} f(x_i) \subseteq R_s^{T_2}$;

1id1:core:20261122-0005@v1 1id1.net

nullary facts are preserved because f preserves them. Hence ρ is a mixture of Token maps, and $[\rho] = f$.

(c) The inclusions follow from (a) and (b). For strictness of the first take $T_1 = \langle\{a\}, E, \{(a,a)\}\rangle$ and $T_2 = \langle\{0,1\}, E, \{(0,0),(1,1)\}\rangle$: both constant functions are Token maps, their uniform mixture is not $\mathcal{D}^{\times}$-stochastic. For strictness of the second see Examples 4.7 and 4.8.

For failure with infinite source, give $\mathbb{N}$ and $\{0,1\}$ empty heaps and put $f(n)(1) = 1/(n+2)$. Every kernel between these classes is $\mathcal{D}^{\times}$-stochastic. If f were a mixture of m functions with weights $\lambda_1, \ldots, \lambda_m$, each number $f(n)(1)$ would be a subset sum of these weights, giving at most 2^m distinct values. The displayed kernel has infinitely many distinct values, a contradiction. □

Example 4.7 (A stochastic Token map where no Token map exists)**.** Let $r \in E$, let $K_3 = \langle\{a,b,c\}, E, \{(r,x,y) : x \neq y\}\rangle$ and $K_2 = \langle\{0,1\}, E, \{(r,0,1),(r,1,0)\}\rangle$. A Token map $K_3 \to K_2$ is a proper 2-colouring of a triangle, so $\mathsf{TS}_E(K_3, K_2) = \varnothing$ and there are no mixtures. The constant kernel $u(x) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ is a stochastic Token map, since $\frac{1}{2}\delta_{(0,1)} + \frac{1}{2}\delta_{(1,0)}$ couples $u(x)$ and $u(y)$ inside the relation of K_2. Deleting the leading core symbol in both heaps gives the same example for $E = \varnothing$.

Example 4.8 (A stochastic homomorphism that is not a mixture)**.** Let Z be the Token class on $\{0,1\}$ whose Tokens are the words $(+, x, y, x \oplus y)$; it records addition modulo 2. The Token maps $Z \to Z$ are the zero map and the identity, so every mixture sends 0 to δ_0. The constant kernel $\kappa(x) = \frac{1}{3}\delta_0 + \frac{2}{3}\delta_1$ is a stochastic Token map: the uniform distribution on the three triples $(1,0,1)$, $(0,1,1)$ and $(1,1,0)$ is supported in the graph of $\oplus$, and each of its three marginals equals $\kappa(x)$.

Remark 4.9 (Local and global)*.* For finite B_1, a kernel f is the kernel of a mixture iff there is one joint distribution of a family $(Y_x)_{x \in B_1}$ with $Y_x \sim f(x)$ under which $x \mapsto Y_x$ is almost surely a Token map. A stochastic Token map only requires a joint distribution for the carrier *occurrences* of each single Token. If $x_i = x_j$, the corresponding marginals agree, but the chosen coupling need not satisfy $Y_i = Y_j$ almost surely. For example, the witness in Example 4.8 for the source Token $(+, 0, 0, 0)$ is not supported on the diagonal. Thus local preservation does not mean that each distinct source element is sampled once and reused within the Token. The first condition is global and the second local. Supplying a finitely supported random input and requiring every resulting function to be a Token map produces a finite mixture; on a finite source, arbitrary shared randomness with finitely supported coordinate laws also induces a finite mixture. These hypotheses specify the global interpretation of the additional-randomness alternative in [1]; the examples show that the local notion is strictly more permissive. The gap has the same form as the gap between local consistency and global sections in the sheaf-theoretic account of contextuality [4], although a stochastic Token map does not fix the joint distributions on Tokens.

Remark 4.10 (Relation to linear programming relaxations)*.* For finite relational structures, $\mathcal{D}^{\mathrm{c}}$ restricted to rational distributions is the structure $\mathrm{LP}(\mathbf{A})$ of [9, Definition 7.10], and an instance $\mathbf{I}$ has a feasible basic linear programming relaxation over $\mathbf{A}$ iff there is a homomorphism $\mathbf{I} \to \mathrm{LP}(\mathbf{A})$; see also [24]. In our terms these are kernels $\mathbf{I} \rightsquigarrow \mathbf{A}$ together with a coupling for each source tuple. Over the rationals they give exactly the displayed BLP witnesses; allowing real weights gives the same existence criterion because this finite feasibility problem has rational coefficients. We use the occurrence-wise marginal convention of [9, Definition 7.6], which does not add an equality constraint between repeated coordinates beyond their common marginal. Example 4.7 is the standard integrality gap of graph colouring. The same construction for the nonempty finite powerset monad gives the power structure that characterises arc consistency [11]. Existence of a stochastic Token map between finite classes with finitely many Tokens is thus decidable by linear programming, while existence of a Token map is in general NP-hard.

1id1:core:20261122-0005@v1 1id1.net

5 Markov categories of Token classes

Lemma 5.1. *Let M be a commutative monad on* **Set**. *Then $\nabla_{A,B}$ is a Token map $M^{\mathrm{c}}T_A \otimes M^{\mathrm{c}}T_B \to M^{\mathrm{c}}(T_A \otimes T_B)$, and the strength is a Token map $T_A \otimes M^{\mathrm{c}}T_B \to M^{\mathrm{c}}(T_A \otimes T_B)$. If M is affine, $M^{\mathrm{c}}1_E \cong 1_E$. For $M = \mathcal{D}$ the same holds for $\mathcal{D}^\delta$ and $\mathcal{D}^\times$.*

Proof. A Token of $M^{\mathrm{c}}T_A \otimes M^{\mathrm{c}}T_B$ of shape s consists of $(p_i)_i = \mathrm{m}_s(\omega_A)$ and $(q_i)_i = \mathrm{m}_s(\omega_B)$. Let $\omega \in M(R_s^{T_A \otimes T_B})$ be the image of $\nabla(\omega_A, \omega_B)$ under the bijection (1). Its i-th coordinate map is $\pi_i \times \pi_i$, so naturality of ∇ gives $M\pi_i(\omega) = \nabla(p_i, q_i)$. The strength is $\nabla \circ (\eta \times \mathrm{id})$, a composite of Token maps. If $M1 \cong 1$, the relations of $M^{\mathrm{c}}1_E$ are images of the nonempty sets $M(1)$, so its heap is full. For $\mathcal{D}^\delta$ use $\delta_x \otimes \delta_y = \delta_{(x,y)}$, and for $\mathcal{D}^\times$ use $\mathrm{supp}(p \otimes q) = \mathrm{supp}\, p \times \mathrm{supp}\, q$. □

Theorem 5.2. *Let M be a commutative affine monad on* **Set**. *Then $\mathsf{Kl}(M^{\mathrm{c}})$ is a Markov category, with tensor $T_A \otimes T_B$ on objects and with copy and delete maps $\eta \circ \langle \mathrm{id}, \mathrm{id} \rangle$ and $\eta \circ !$. For $M = \mathcal{D}$ and $L \in \{\delta, \times, \mathrm{c}\}$:*

(a) *$\mathsf{Kl}(\mathcal{D}^L)$ is a Markov category, and the functors $\mathsf{Kl}(\mathcal{D}^\delta) \to \mathsf{Kl}(\mathcal{D}^\times) \to \mathsf{Kl}(\mathcal{D}^{\mathrm{c}}) \to \mathsf{Kl}(\mathcal{D})$ preserve the Markov structure strictly;*

(b) *a morphism of $\mathsf{Kl}(\mathcal{D}^L)$ is deterministic iff it is $\delta \circ g$ for a Token map g, so the deterministic subcategory is isomorphic to TS_E;*

(c) *$\mathsf{Kl}(\mathcal{D}^L)$ is representable in the sense of [14, Definition 3.10]: the inclusion of the deterministic subcategory has a right adjoint, which sends T to $\mathcal{D}^L T$.*

Proof. By Lemma 5.1 the lifted monad carries the maps ∇. The equations making it a symmetric monoidal monad hold in **Set**; since U is faithful and preserves finite products strictly, they hold in TS_E. Together with $M^{\mathrm{c}}1_E \cong 1_E$ this makes M^{c} a commutative affine monad on a category with finite products, and [13, Corollary 3.2] applies. The same argument applies to $\mathcal{D}^\delta$ and $\mathcal{D}^\times$, and the comparison functors are the identity on objects and on underlying kernels.

(b) In $\mathsf{Kl}(\mathcal{D})$ the two sides of the determinism equation for f are $x \mapsto \mathcal{D}\langle \mathrm{id}, \mathrm{id} \rangle(f(x))$ and $x \mapsto f(x) \otimes f(x)$. They agree iff $f(x)(y)f(x)(y') = 0$ for $y \neq y'$, that is, iff $f(x)$ is a Dirac distribution. So $f = \delta \circ g$, and $\delta \circ g$ is a Token map $T_1 \to \mathcal{D}^L T_2$ iff g is a Token map $T_1 \to T_2$, because L is conservative.

(c) Under the identification in (b) the inclusion is the left adjoint of the Kleisli adjunction. □

The tensor of $\mathsf{Kl}(\mathcal{D}^L)$ is not a categorical product, as in every Markov category with nondeterministic morphisms. The exponentials of TS_E nevertheless remain available in the form familiar from computational effects [27].

Proposition 5.3 (Stochastic function objects)**.** *For Token classes T_B, T_C put $[T_B, T_C]_L = (\mathcal{D}^L T_C)^{T_B}$. There is a bijection*

$$\mathsf{Kl}(\mathcal{D}^L)(T_A \otimes T_B,\, T_C) \cong \mathsf{TS}_E(T_A,\, [T_B, T_C]_L),$$

natural in $T_A \in \mathsf{TS}_E$. The carrier of $[T_B, T_C]_L$ is the set of all kernels $B \to \mathcal{D}(C)$, and its global elements $1_E \to [T_B, T_C]_L$ are exactly the L-stochastic Token maps $T_B \rightsquigarrow T_C$.

Proof. The bijection is the exponential adjunction of TS_E applied to $\mathcal{D}^L T_C$. The carrier is $\mathcal{D}(C)^B$ because U preserves exponentials. A global element selects a kernel u such that $s[u, \ldots, u]$ is a Token of the exponential for every shape s, which by the description of the exponential heap says that $(u(b_1), \ldots, u(b_n)) \in R_s^{\mathcal{D}^L T_C}$ for every $(b_1, \ldots, b_n) \in R_s^{T_B}$. □

1id1:core:20261122-0005@v1 1id1.net

Hence the reading of an architecture as a composite of products, exponentials and evaluation, with parameters as points of bases [1], extends to stochastic layers: a parameterised stochastic layer is a Token map $\Theta \to [T_B, T_C]_L$. Since $\mathcal{D}^L$ is a strong monad on a cartesian closed category, Kleisli composition is itself a Token map $[T_B, T_C]_L \otimes [T_A, T_B]_L \to [T_A, T_C]_L$ [23, 27]. Explicitly, write $P = [T_B, T_C]_L$ and $Q = [T_A, T_B]_L$. The uncurried map first evaluates $Q \otimes T_A \to \mathcal{D}^L T_B$, then applies strength $P \otimes \mathcal{D}^L T_B \to \mathcal{D}^L(P \otimes T_B)$, then $\mathcal{D}^L$ of evaluation into $\mathcal{D}^L\mathcal{D}^L T_C$, and finally μ. Every step is a Token map; currying gives the stated composition.

Remark 5.4 (Other notions of uncertainty). Theorem 5.2 is stated for an arbitrary commutative affine monad. For the nonempty finite powerset monad it gives a Markov category of *possibilistic* Token maps, in which a multivalued function preserves a Token when its sets of values are the projections of a set of target Tokens. This is the sense in which the Markov-category formulation is more general than, and not equivalent to, the probabilistic one: it axiomatises what the probabilistic, possibilistic and other models share.

6 Convex Token classes

A *convex space* is an Eilenberg–Moore algebra (C, a) of $\mathcal{D}$: a set with an operation assigning a barycentre $a(p) \in C$ to every finitely supported distribution, subject to the algebra laws [29, 31, 12, 19]. Algebra maps are called *affine*, and the category is **Conv**. Convex subsets of real vector spaces are examples, and $\mathcal{D}(X)$ is the free convex space on X. A subset $S \subseteq C$ is *convex* when $a(p) \in S$ for every p with $\operatorname{supp} p \subseteq S$. Powers C^n are convex spaces coordinatewise.

Definition 6.1. A *convex Token class* is a Token class $T = \langle C, E, \Upsilon \rangle$ with a convex space structure a on C such that every relation R_s^T is a convex subset of $C^{|s|}$. Morphisms are the affine Token maps. The category is CTS_E.

Theorem 6.2. $\mathsf{EM}(\mathcal{D}^{\mathrm{c}}) \cong \mathsf{CTS}_E$ *over* TS_E. *Moreover* $\mathsf{EM}(\mathcal{D}^{\delta})$ *consists of all Token classes with a convex space structure on the carrier, and* $\mathsf{EM}(\mathcal{D}^{\times})$ *of those for which* $S_1 \times \cdots \times S_n \subseteq R_s^T$ *implies* $\operatorname{conv} S_1 \times \cdots \times \operatorname{conv} S_n \subseteq R_s^T$. *In all three cases the morphisms are the affine Token maps, so*

$$\mathsf{CTS}_E = \mathsf{EM}(\mathcal{D}^{\mathrm{c}}) \subseteq \mathsf{EM}(\mathcal{D}^{\times}) \subseteq \mathsf{EM}(\mathcal{D}^{\delta})$$

as full subcategories.

Proof. An algebra of a lifted monad $\mathcal{D}^L$ is a Token map $a : \mathcal{D}^L T \to T$ satisfying the algebra laws, and since U is faithful these laws say that (C, a) is a convex space. Algebra morphisms are Token maps that are affine. It remains to express that a is a Token map.

For $\mathcal{D}^{\mathrm{c}}$: the condition is $a^n(\mathrm{m}_s^T(\omega)) \in R_s^T$ for every $\omega \in \mathcal{D}(R_s^T)$. The map $a^n \circ \langle \mathcal{D}\pi_1, \ldots, \mathcal{D}\pi_n \rangle$ is the barycentre map of the convex space C^n, so the condition says that R_s^T contains the barycentre of every distribution supported in it.

For $\mathcal{D}^{\times}$: the condition is $(a(p_i))_i \in R_s^T$ whenever $\prod_i \operatorname{supp} p_i \subseteq R_s^T$, and $\{a(p) : \operatorname{supp} p \subseteq S\} = \operatorname{conv} S$. For $\mathcal{D}^{\delta}$: the condition is $(a(\delta_{x_i}))_i = (x_i)_i \in R_s^T$ for $x \in R_s^T$, which always holds. □

The order of the three categories is the reverse of the order of the liftings: a larger lifted heap imposes a stronger condition on algebras.

Example 6.3. (i) $\mathcal{D}^{\mathrm{c}} T$ is the free convex Token class on T, and $([0,1], \leq)$ of Example 3.9 is isomorphic to $\mathcal{D}^{\mathrm{c}}$ of the two-element chain via $p \mapsto p(1)$. (ii) The tokenization of $\mathbb{R}^d$ as a vector space [1] in which every scalar is a core symbol, with Tokens for the graphs of addition and of each scalar multiplication, is a convex Token class: its relations are linear subspaces. For $d \geq 1$ it is not free, since $\mathbb{R}^d$ has no extreme points. This argument does not apply to the alternative

1id1:core:20261122-0005@v1 1id1.net

tokenization with scalars in the carrier: its disjoint-union carrier first needs a specified convex structure, and the graph of joint scalar multiplication is not convex under the usual product convex structure on scalar and vector variables. (iii) $\mathsf{Cau}_n(\mathbb{R}^d)$ and $\mathsf{Cau}_n(\Delta(V))$ are convex Token classes, because prefix agreement is a linear condition. (iv) Recording the graph of a nonlinear function such as ReLU or multiplication, or an inequality $x \neq y$, gives a heap that is not convex. A heap is convex exactly when the recorded structure is compatible with averaging.

Proposition 6.4 (Expectation)**.** *Let $(\mathcal{C}, a)$ be a convex Token class and $h : T \to \mathcal{C}$ a Token map. Then $h^\dagger = a \circ \mathcal{D}h$ is the unique affine Token map $\mathcal{D}^{\mathrm{c}}T \to \mathcal{C}$ with $h^\dagger \circ \delta = h$. Consequently, for every stochastic Token map $f : T_1 \rightsquigarrow T$ the expected value*

$$x \longmapsto \mathbb{E}_{y \sim f(x)}\, h(y) = h^\dagger(f(x))$$

is a deterministic Token map $T_1 \to \mathcal{C}$.

Proof. This is the universal property of the free algebra $\mathcal{D}^{\mathrm{c}}T$ in $\mathsf{EM}(\mathcal{D}^{\mathrm{c}})$, and $h^\dagger \circ f$ is a composite of Token maps. □

The comparison functor $\mathsf{Kl}(\mathcal{D}^{\mathrm{c}}) \to \mathsf{EM}(\mathcal{D}^{\mathrm{c}})$ is full and faithful, so a stochastic Token map $T_1 \rightsquigarrow T_2$ is the same thing as an affine Token map $\mathcal{D}^{\mathrm{c}}T_1 \to \mathcal{D}^{\mathrm{c}}T_2$. The category CTS_E is strictly larger: it also contains the non-free classes of Example 6.3(ii), which are where the hidden states of a network live. Limits in CTS_E are computed as in TS_E, because the forgetful functor of an Eilenberg–Moore category creates limits.

Remark 6.5 (Attention as a barycentre). Let $v : [n] \to \mathbb{R}^d$ list the value vectors of an attention head and let $\alpha_i \in \mathcal{D}([n])$ be the i-th row of attention weights. The head outputs $\sum_j \alpha_i(j) v_j = v^\dagger(\alpha_i)$. Soft attention is therefore the affine extension of value lookup to distributions over positions, and hard attention is its restriction to Dirac distributions. Proposition 6.4 then says that if value lookup is a Token map into a convex Token class and the weights form a stochastic Token map, the attention output is a Token map. The remark concerns the convex-combination step only; the dependence of the weights on queries and keys is a separate map.

7 Coalgebras: predictors, prefix trees and generation

7.1 Predictors and the final coalgebra

Let V be a vocabulary and O a set of outputs. A *predictor* is a coalgebra $\langle o, t\rangle : X \to O \times X^V$ for the functor $H(X) = O \times X^V$ on **Set**: a set of states, an output for each state, and a transition for each token. The final H-coalgebra is $Z = O^{V^*}$ with $\zeta(\varphi) = (\varphi(\varepsilon),\, v \mapsto \varphi(v\,\cdot))$, and the unique coalgebra map is $\mathrm{beh}(x)(u) = o(t_u(x))$, where t_u is the iterated transition along the word u [28, 20].

For $O = \mathcal{D}(V)$ an element of Z is a function assigning a next-token distribution to every context, that is, a V-branching tree whose node u carries a distribution: a *probabilistic prefix tree*. An idealised language model with unbounded context is one point of Z; a model with context bound n determines the tree up to depth n. A stateful implementation is a predictor, and beh sends each state to the tree it realises; the cached decoder of [1], whose agreement with full-prefix evaluation is expressed there by a morphism of Mealy coalgebras, is a machine of this kind. This is the precise sense in which a language model is a tree with a compressed implicit representation. It is distinct from the unfolding of a single piecewise-linear network into a decision tree [6], which concerns the computation inside one node.

1id1:core:20261122-0005@v1 1id1.net

7.2 Lifting final coalgebras

Theorem 7.1 (Final coalgebras lift)**.** *Let $H : \mathbf{Set} \to \mathbf{Set}$ have a final coalgebra (Z, ζ) and let $\overline{H} : \mathsf{TS}_E \to \mathsf{TS}_E$ be a functor with $U\overline{H} = HU$. For a heap Υ on Z let $\Phi(\Upsilon) = \zeta^*(\text{heap of } \overline{H}\langle Z, E, \Upsilon\rangle)$. Then Φ is monotone, and with ν its greatest fixed point, $(\langle Z, E, \nu\rangle, \zeta)$ is a final $\overline{H}$-coalgebra.*

Proof. If $\Upsilon \subseteq \Upsilon'$, the identity of Z is a Token map, so its image under $\overline{H}$ is a Token map with underlying function id_{HZ}; hence Φ is monotone and ν exists by the Knaster–Tarski theorem. By Lemma 2.1, $\nu \subseteq \Phi(\nu)$ says that $\zeta : \langle Z, E, \nu\rangle \to \overline{H}\langle Z, E, \nu\rangle$ is a Token map.

Let $c : \mathcal{X} \to \overline{H}\mathcal{X}$ be a coalgebra with $\mathcal{X} = \langle X, E, \Upsilon_X\rangle$ and let $h : X \to Z$ be the unique H-coalgebra map. Since U is faithful, a $\overline{H}$-coalgebra morphism into $\langle Z, E, \nu\rangle$ can only be h, so it suffices to show that h is a Token map $\mathcal{X} \to \langle Z, E, \nu\rangle$. Put $\Upsilon' = h_*\Upsilon_X$. Then $h : \mathcal{X} \to \langle Z, E, \Upsilon'\rangle$ is a Token map, and so is $Hh : \overline{H}\mathcal{X} \to \overline{H}\langle Z, E, \Upsilon'\rangle$. For $w \in \Upsilon_X$,

$$F_E(\zeta)F_E(h)(w) = F_E(Hh)F_E(c)(w),$$

which is a Token of $\overline{H}\langle Z, E, \Upsilon'\rangle$. Hence $\Upsilon' \subseteq \Phi(\Upsilon')$, so $\Upsilon' \subseteq \nu$. □

This is an instance of coinduction in a fibration [18, 17]. We apply it to predictors. Let $\mathcal{O} = \langle O, E, \Upsilon_O\rangle$ and $\mathcal{V} = \langle V, E, \Upsilon_V\rangle$ be Token classes and

$$\overline{H}(\mathcal{X}) = \mathcal{O} \otimes \mathcal{X}^{\mathcal{V}},$$

with the exponential of TS_E. Then $U\overline{H} = HU$. A $\overline{H}$-coalgebra is a predictor whose output map $o : \mathcal{X} \to \mathcal{O}$ and transition $\mathcal{X} \otimes \mathcal{V} \to \mathcal{X}$ are Token maps. Call words $u_1, \dots, u_n \in V^m$ *s-related* when $(u_1[j], \dots, u_n[j]) \in R_s^{\mathcal{V}}$ for every position $j \le m$.

Proposition 7.2 (Behavioural heap)**.** *For $\overline{H}(\mathcal{X}) = \mathcal{O} \otimes \mathcal{X}^{\mathcal{V}}$ the final coalgebra is O^{V^*} with the heap ν for which $(\varphi_1, \dots, \varphi_n) \in R_s^{\nu}$ iff*

$$(\varphi_1(u_1), \dots, \varphi_n(u_n)) \in R_s^{\mathcal{O}}$$

for all $m \ge 0$ and all s-related words $u_1, \dots, u_n \in V^m$. Consequently, if $\langle o, t\rangle : \mathcal{X} \to \overline{H}\mathcal{X}$ is a coalgebra and $s[x_1, \dots, x_n]$ is a Token of $\mathcal{X}$, then

$$s\big[o(t_{u_1}x_1), \dots, o(t_{u_n}x_n)\big] \in \Upsilon_O$$

for all s-related words $u_1, \dots, u_n$.

Proof. By the descriptions of products and exponentials, $(\varphi_i)_i \in R_s^{\Phi(\Upsilon)}$ iff $(\varphi_i(\varepsilon))_i \in R_s^{\mathcal{O}}$ and $(\varphi_i(v_i\,\cdot))_i \in R_s^{\Upsilon}$ for all $(v_i)_i \in R_s^{\mathcal{V}}$. The displayed family N satisfies $\Phi(N) = N$, by splitting an s-related tuple of words of length $m+1$ into its first letters and the rest. If $\Upsilon \subseteq \Phi(\Upsilon)$, induction on m shows $\Upsilon \subseteq N$. So $N = \nu$. The consequence is the statement that beh is a Token map, by Theorem 7.1. □

Example 7.3. (i) Let every relation $R_s^{\mathcal{V}}$ of positive arity be the diagonal and let every nullary fact hold in $\mathcal{V}$, so that s-related words are equal. A coalgebra is then a predictor all of whose maps o and t_v are Token maps, and every Token of states is preserved along every common continuation. (ii) Let in addition Υ_O consist of the equality Tokens $(\approx, y, y)$. A heap of Tokens $(\approx, x, x')$ on X that makes $\langle o, t\rangle$ a coalgebra is a bisimulation, and the proposition is the statement that bisimilar states have the same behaviour. (iii) Let $\mathcal{O} = \mathcal{D}^{\mathrm{c}}\mathcal{V}'$ for a Token class $\mathcal{V}'$ on the vocabulary. Then related states have next-token distributions that can be coupled inside the relations of $\mathcal{V}'$. With equality Tokens on $\mathcal{V}'$ and diagonal input relations, related states have equal next-token laws and remain related under every common input token. When the

1id1:core:20261122-0005@v1 1id1.net

state relation is an equivalence, these conditions suffice for probabilistic bisimulation of the associated labelled generative system. This is a sufficient criterion, not a characterization of general Larsen–Skou bisimilarity [26], which compares probability mass on equivalence classes of successors. A tolerance relation on outputs instead gives an invariant coupling condition under the designated continuations. (iv) If $R_s^{\mathcal{V}}$ is a tolerance relation on input tokens, the proposition is a robustness statement: related states remain related in their outputs under letterwise related continuations.

7.3 Generation as a stochastic Token map

Now let $\mathcal{O} = \mathcal{D}^{\mathrm{c}}\mathcal{V}$. A coalgebra $c = \langle o, t\rangle : \mathcal{X} \to \mathcal{D}^{\mathrm{c}}\mathcal{V} \otimes \mathcal{X}^{\mathcal{V}}$ determines the *generation kernel*

$$\mathrm{gen}(x) = \sum_{v \in V} o(x)(v)\, \delta_{(v,\, t_v(x))} \ \in \mathcal{D}(V \times X),$$

which samples a token and moves to the next state. Iteration gives $\mathrm{gen}^{(m)} : X \to \mathcal{D}(V^m \times X)$, the joint law of m generated tokens and the final state.

Proposition 7.4 (Generation)**.** gen *is a stochastic Token map* $\mathcal{X} \rightsquigarrow \mathcal{V} \otimes \mathcal{X}$*, and* $\mathrm{gen}^{(m)}$ *is a stochastic Token map* $\mathcal{X} \rightsquigarrow \mathcal{V}^{\otimes m} \otimes \mathcal{X}$*. Explicitly, for every Token* $s[x_1, \dots, x_n]$ *of* $\mathcal{X}$ *there is a joint distribution of* n *runs of length* m *started in* $x_1, \dots, x_n$ *under which, with probability one, the generated words are* s*-related and the final states form a Token* $s[x'_1, \dots, x'_n]$ *of* $\mathcal{X}$*.*

Proof. Let $\kappa : \mathcal{D}^{\mathrm{c}}\mathcal{V} \otimes \mathcal{X}^{\mathcal{V}} \to \mathcal{D}^{\mathrm{c}}(\mathcal{V} \otimes \mathcal{X})$ be the composite of the strength $\mathcal{D}^{\mathrm{c}}\mathcal{V} \otimes \mathcal{X}^{\mathcal{V}} \to \mathcal{D}^{\mathrm{c}}(\mathcal{V} \otimes \mathcal{X}^{\mathcal{V}})$ of Lemma 5.1 with $\mathcal{D}^{\mathrm{c}}\langle \pi_1, \mathrm{ev}\rangle$, where ev is evaluation. Both are Token maps, and $\mathrm{gen} = \kappa \circ c$. In the Markov category $\mathsf{Kl}(\mathcal{D}^{\mathrm{c}})$ put $\mathrm{gen}^{(m+1)} = (\mathrm{id}_{\mathcal{V}^{\otimes m}} \otimes \mathrm{gen}) \bullet \mathrm{gen}^{(m)}$. The explicit statement is the definition of a stochastic Token map together with (1). □

For $n = 2$ and equality on the vocabulary, the proposition states that two related states generate the same distribution over strings, by the usual coupling argument. In general it transports any recorded relation between states, for example agreement of two implementations or of a teacher and a student on a specified heap, to a relation between sampled continuations.

Remark 7.5 (Trace semantics). For unbounded finite traces one must enlarge the support convention. Let $\mathcal{D}^{\omega}_{\leq 1}(X)$ be the set of countably supported subprobability distributions on X. A model with explicit termination is a coalgebra $X \to \mathcal{D}^{\omega}_{\leq 1}(1 + V \times X)$. In the order-enriched setting of [16], the inverse of the initial $1 + V \times (-)$-algebra, embedded into $\mathsf{Kl}(\mathcal{D}^{\omega}_{\leq 1})$, is final for the lifted functor; the resulting trace map assigns a subdistribution on V^* to each state. Finite branching does not imply finite trace support: stopping with probability $1/2$ and otherwise emitting a and repeating gives mass $2^{-(k+1)}$ to a^k for every $k \geq 0$. The finite-support subdistribution monad therefore does not suffice for this claim. Whether the countably supported trace construction lifts to Token classes with the required order enrichment is not settled here.

8 Discussion and open problems

The four tools answer different questions. The lifting decides what preservation means; the Markov category provides the calculus of stochastic maps and recovers Token Space as its deterministic part; convex Token classes provide the objects in which expectations are taken, including the representation spaces of networks; and coalgebras describe behaviour over time. The coupling lifting has a distinguished position among conservative liftings: it is the least one with convex relations, it admits sampling between causal classes, and its algebras are the convex Token classes.

1id1:core:20261122-0005@v1 1id1.net

Limitations. The monad $\mathcal{D}$ contains only finitely supported distributions. This covers distributions over a finite vocabulary but not continuous noise, for which a measurable version over the Giry monad [15] is needed. Stochastic Token maps are still all-or-nothing: a Token is preserved or it is not. Measuring approximate preservation, as required for distillation in [1], needs quantitative heaps and is not attempted. No statement here concerns learning.

Open problems.

(1) Characterise the conservative liftings of $\mathcal{D}$ to TS_E. Is there a largest one, and where do the codensity liftings sit?

(2) Which Token classes T and shapes admit conditionals in $\mathsf{Kl}(\mathcal{D}^{\mathrm{c}})$, in the sense of [13]? This governs Bayesian inversion at the level of Tokens.

(3) Is there an obstruction theory for the gap between stochastic Token maps and mixtures (Remark 4.9)?

(4) Give explicit tensor and internal-hom constructions for convex Token classes, and determine when they have finite presentations, taking account of the standard monoidal constructions for algebras of commutative monads [22, 23].

(5) Lift the trace semantics of Remark 7.5 to Token classes.

(6) Replace heaps by heaps valued in a quantale, with the Kantorovich and Wasserstein liftings of [7], so that the distance of a kernel from being a stochastic Token map can be measured.

(7) Extend the results to measurable carriers and the Giry monad.

A Finite checks

The script `anc/check_liftings.py` tests finite instances of the statements above with exact rational arithmetic; coupling and mixture existence are decided by an exact simplex method. It checks: closure of coupling-lifted relations under mixtures (Theorem 3.5); the inclusions of Proposition 4.6 on random finite classes together with witnesses of strictness; Examples 4.7 and 4.8; the equivalence of coupling and up-set dominance on random preorders (Example 3.9); closure of stochastic Token maps under composition and the Token-map property of ∇ (Lemma 5.1); the characterisation of stochastic causal maps for $n = 2$ (Theorem 4.3); monotonicity of expectations (Proposition 6.4); the behavioural heap of Proposition 7.2 on random finite predictors; and the joint-sampling property of Proposition 7.4 for two generation steps. The results are recorded in `anc/check_liftings.json`. The supplementary script `anc/check_review_regressions.py` checks the repeated-occurrence witness, core-free strictness witnesses, finite subset-sum obstructions, and the geometric trace-support example introduced in this revision. Such tests cover small instances only and do not replace the proofs.

References

[1] W. Pan, *Token Space: A Category Theory Framework for AI Computations*, arXiv:2404.11624v4, revised October 2, 2026 (first posted 2024).

[2] W. Pan, *Categorical Studies of Token Structures: A Historical Perspective*, 1id1 preprint, October 26, 2026. 1id1:core:20261026-0001@v1.

[3] W. Pan, *Monoidal Token Spaces: Initial Sources and Tensor Adjoints*, 1id1 preprint, October 26, 2026. 1id1:core:20261026-0002@v1.

[4] S. Abramsky and A. Brandenburger, The sheaf-theoretic structure of non-locality and contextuality, *New Journal of Physics* 13 (2011), 113036.

[5] J. Adámek, H. Herrlich, and G. E. Strecker, *Abstract and Concrete Categories: The Joy of Cats*, Wiley, 1990; reprinted in *Reprints in Theory and Applications of Categories* 17, 2006.

[6] C. Aytekin, *Neural networks are decision trees*, arXiv:2210.05189 (2022).

[7] P. Baldan, F. Bonchi, H. Kerstan, and B. König, Coalgebraic behavioral metrics, *Logical Methods in Computer Science* 14(3:20) (2018). arXiv:1712.07511.

[8] M. Barr, Relational algebras, in *Reports of the Midwest Category Seminar IV*, Lecture Notes in Mathematics 137, Springer, 1970, 39–55.

[9] L. Barto, J. Bulín, A. Krokhin, and J. Opršal, Algebraic approach to promise constraint satisfaction, *Journal of the ACM* 68(4) (2021), Article 28. arXiv:1811.00970v3 (the definition numbers cited here refer to this version).

[10] K. Cho and B. Jacobs, Disintegration and Bayesian inversion via string diagrams, *Mathematical Structures in Computer Science* 29(7) (2019), 938–971.

[11] T. Feder and M. Y. Vardi, The computational structure of monotone monadic SNP and constraint satisfaction: a study through Datalog and group theory, *SIAM Journal on Computing* 28(1) (1998), 57–104.

[12] T. Fritz, *Convex spaces I: definition and examples*, arXiv:0903.5522 (2009).

[13] T. Fritz, A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics, *Advances in Mathematics* 370 (2020), 107239. arXiv:1908.07021.

[14] T. Fritz, T. Gonda, P. Perrone, and E. F. Rischel, Representable Markov categories and comparison of statistical experiments in categorical probability, *Theoretical Computer Science* 961 (2023), 113896. arXiv:2010.07416.

[15] M. Giry, A categorical approach to probability theory, in *Categorical Aspects of Topology and Analysis*, Lecture Notes in Mathematics 915, Springer, 1982, 68–85.

[16] I. Hasuo, B. Jacobs, and A. Sokolova, Generic trace semantics via coinduction, *Logical Methods in Computer Science* 3(4:11) (2007). arXiv:0710.2505.

[17] I. Hasuo, T. Kataoka, and K. Cho, Coinductive predicates and final sequences in a fibration, *Mathematical Structures in Computer Science* 28(4) (2018), 562–611.

[18] C. Hermida and B. Jacobs, Structural induction and coinduction in a fibrational setting, *Information and Computation* 145(2) (1998), 107–152.

[19] B. Jacobs, Convexity, duality and effects, in *Theoretical Computer Science* (IFIP TCS 2010), IFIP Advances in Information and Communication Technology 323, Springer, 2010, 1–19.

[20] B. Jacobs, *Introduction to Coalgebra: Towards Mathematics of States and Observation*, Cambridge University Press, 2016.

[21] S. Katsumata and T. Sato, Codensity liftings of monads, in *6th Conference on Algebra and Coalgebra in Computer Science (CALCO 2015)*, LIPIcs 35, 2015, 156–170.

[22] A. Kock, Monads on symmetric monoidal closed categories, *Archiv der Mathematik* 21 (1970), 1–10.

[23] A. Kock, Strong functors and monoidal monads, *Archiv der Mathematik* 23 (1972), 113–120.

[24] G. Kun, R. O'Donnell, S. Tamaki, Y. Yoshida, and Y. Zhou, Linear programming, width-1 CSPs, and robust satisfaction, in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 2012)*, ACM, 2012, 484–495.

[25] A. Kurz and J. Velebil, Relation lifting, a survey, *Journal of Logical and Algebraic Methods in Programming* 85(4) (2016), 475–499.

[26] K. G. Larsen and A. Skou, Bisimulation through probabilistic testing, *Information and Computation* 94(1) (1991), 1–28.

[27] E. Moggi, Notions of computation and monads, *Information and Computation* 93(1) (1991), 55–92.

[28] J. J. M. M. Rutten, Universal coalgebra: a theory of systems, *Theoretical Computer Science* 249(1) (2000), 3–80.

[29] M. H. Stone, Postulates for the barycentric calculus, *Annali di Matematica Pura ed Applicata* 29 (1949), 25–30.

[30] V. Strassen, The existence of probability measures with given marginals, *Annals of Mathematical Statistics* 36(2) (1965), 423–439.

[31] T. Świrszcz, Monadic functors and convexity, *Bulletin de l'Académie Polonaise des Sciences, Série des Sciences Mathématiques, Astronomiques et Physiques* 22 (1974), 39–42.

1id1:core:20261122-0005@v1 1id1.net