An Elementary Account of the Internal Model Principle

Matteo Capucci

University of Strathclyde / Independent ARIA Creator

jww Baltieri, Biehl, Virgo

arxiv:2503.00511

The Classics

Theorem: The simplest optimal regulator R of a reguland S produces events R which are related to the events S by a mapping $h : S \rightarrow R$.

Restated somewhat less rigorously, the theorem says that the best regulator of a system is one which is a model of that system in the sense that the regulator's actions are merely the system's actions as seen through a mapping h. The type of isomorphism here is that expressed (in the form used above) by

$$\exists h : \ \forall i : \ \rho(i) = h[\sigma(i)] \tag{8}$$

where ρ and σ are the mappings that R and S impose on their common input I. This form is essentially that of (5) above.

Conant, Roger C., and Ross W. Ashby. 1970. 'Every Good Regulator of a System Must Be a Model of That System'. *International Journal of Systems Science* 1 (2): 89–97. https://doi.org/10.1080/00207727008920220.

The Classics

Our major conclusion of this section can be summarized as: ***The Internal Model Principle***: A regulator synthesis is structurally stable only if the controller utilizes feedback of the regulated variable, and incorporates in the feedback path a suitably reduplicated model of the dynamic structure of the exogenous signals which the regulator is required to process.

Francis, B. A., and W. M. Wonham. 1976. 'The Internal Model Principle of Control Theory'. *Automatica* 12 (5): 457–65. https://doi.org/10.1016/0005-1098(76)90006-6.

The Classics

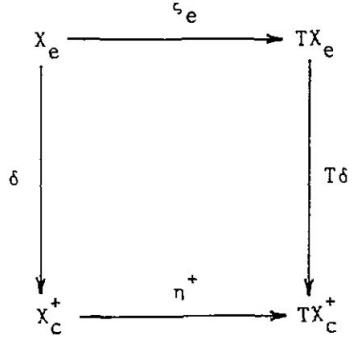
We say that *a map* $A : \mathscr{X} \to \mathscr{X}$ *incorporates an internal model of* A_2 if the minimal polynomial (m.p.) of A_2 divides at least $d(Z)$ invariant factors (i.f.) of A. Thus an internal model is an ℓ-fold reduplication in A of the maximal cyclic component of A_2, where $\ell \geq d(Z) =$ the number of independent outputs to be regulated.

Francis, B. A., and W. M. Wonham. 1976. 'The Internal Model Principle of Control Theory'. *Automatica* 12 (5): 457–65. https://doi.org/10.1016/0005-1098(76)90006-6.

The Classics

V. Existence of Internal Models

On the foregoing basis, it can now be established that the controller incorporates an internal model of some kind, corresponding to different observability conditions. The endomorph will be denoted by X^-, and $\xi|X^-$ by ξ^+, $pr_1|X^+$ by pr_1^+, $pr_1^+(X^-)$ by X_c^+. The controller will be said to *incorporate an internal model of the exosystem*, if there are a diffeomorphism δ: $X_e \to X_c^- \subset X_c$ and a vector field $\eta^+ = \eta|X_c^+$ such that the following diagram commutes.



Hepburn, J., and W. Wonham. 1984. 'Error Feedback and Internal Models on Differentiable Manifolds'. *IEEE Transactions on Automatic Control* 29 (5): 397–403. https://doi.org/10.1109/TAC.1984.1103563.

Ours

+ (light) categorical footing
  > more generality
  > prone to compositionality

+ **clearer assumptions** regarding the structure of the systems
  > we realize some assumptions are not needed!

+ an abstract notion of model
+ a clear link with Bayesian filtering

Still, not the last word & mostly subsumed by older work!

Systems

Definition II.1 (System). A *system* (or more precisely, a *fully observable system*) X is comprised of a set X of *states*, a set I of *inputs* (or *observations*), and an *update* (or *dynamics*) function:

$$\mathsf{upd}_{\mathsf{X}} : X \times I \to X, \tag{1}$$

The pair $\binom{I}{X}$ is collectively referred to as the *interface* of the system, and we write $\mathsf{X} : \mathbf{Sys}\binom{I}{X}$ to mean X has such an interface.

Fully observable, discrete-time,
(deterministic Moore) automata

Systems

Definition II.3 (Map of systems)**.** Let $\mathsf{X} : \mathbf{Sys}\binom{I}{X}$ and $\mathsf{X}' : \mathbf{Sys}\binom{I'}{X'}$ be systems. A *map of systems* $f : \mathsf{X} \to \mathsf{X}'$ is comprised of two parts:

1) a *map on states*, given by a function

$$f_{\mathsf{s}} : X \to X', \tag{2}$$

2) a *map on inputs*, given by a function

$$f_{\mathsf{i}} : X \times I \to I', \tag{3}$$

such that the following diagram commutes:

$$\begin{array}{ccc} X \times I & \xrightarrow{(\pi_X \fatsemi f_{\mathsf{s}}, f_{\mathsf{i}})} & X' \times I' \\ {\scriptstyle \mathsf{upd}_{\mathsf{X}}}\downarrow & & \downarrow{\scriptstyle \mathsf{upd}_{\mathsf{X}'}} \\ X & \xrightarrow{f_{\mathsf{s}}} & X' \end{array} \tag{4}$$

meaning that, for every $x \in X, i \in I$, the following equation is satisfied:

$$f_{\mathsf{s}}(\mathsf{upd}_X(x, i)) = \mathsf{upd}_{X'}(f_{\mathsf{s}}(x), f_{\mathsf{i}}(x, i)), \tag{5}$$

Systems

Construction II.4. In what follows, we will need to compose maps of systems, so we explain here how that happens. Given maps $f : \mathsf{X} \to \mathsf{X}'$ and $g : \mathsf{X}' \to \mathsf{X}''$, where $\mathsf{X} : \mathbf{Sys}\binom{I}{X}$, $\mathsf{X}' : \mathbf{Sys}\binom{I'}{X'}$ and $\mathsf{X}'' : \mathbf{Sys}\binom{I''}{X''}$, their composition $f \fatsemi g : \mathsf{X} \to \mathsf{X}''$ is given on states by the composition of the maps on states:

$$(f \fatsemi g)_{\mathsf{s}} = f_{\mathsf{s}} \fatsemi g_{\mathsf{s}}, \tag{6}$$

while on inputs $(f \fatsemi g)_{\mathsf{i}} : X \times I \to I''$ is defined as follows:

$$(f \fatsemi g)_{\mathsf{i}}(x, i) = g_{\mathsf{i}}(f_{\mathsf{s}}(x), f_{\mathsf{i}}(x, i)). \tag{7}$$

$$X \xrightarrow{f_s} X' \xrightarrow{g_s} X''$$

$$X \times I \xrightarrow{(f_s,\, f_i)} X' \times I' \xrightarrow{g_i} I''$$

Systems

Definition II.5 (Subsystem)**.** A *subsystem* of X is a forward-invariant subset of X together with updates restricted to the subset, thus a map of systems $\gamma : \mathsf{X}' \to \mathsf{X}$ given on states by an inclusion $\gamma_{\mathsf{s}} : X' \to X$ and on inputs by projection $\gamma_{\mathsf{i}} = \pi_I : X' \times I \to I$.

Definition II.6 (Attracting subsystem)**.** An *attracting subsystem* for X is a subsystem $\mathsf{X}^* \to \mathsf{X}$ such that, for each $x \in X$, there exists $n \in \mathbb{N}$ such that for all $t \geq n$ and $\underline{i} \in I^t$, we have $\mathsf{upd}^t_{\mathsf{X}}(x, \underline{i}) \in X^*$.

IMP setup

Assumption 1 (Environment, plant, controller)**.** *The following three components are so defined:*

1) *the environment* $\mathsf{E} : \mathbf{Sys}\binom{1}{E}$ *is an autonomous system*

$$\mathsf{upd}_{\mathsf{E}} : E \to E, \tag{8}$$

2) *the plant* $\mathsf{P} : \mathbf{Sys}\binom{E\times C}{P}$ *is a system*

$$\mathsf{upd}_{\mathsf{P}} : P \times E \times C \to P, \tag{9}$$

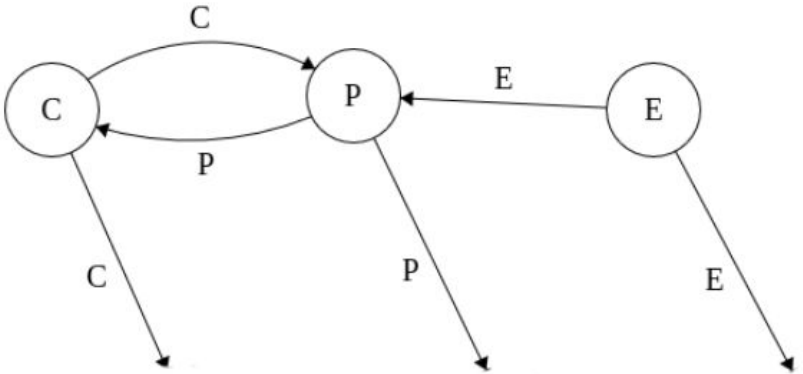
3) *the controller* $\mathsf{C} : \mathbf{Sys}\binom{P}{C}$ *is a system*

$$\mathsf{upd}_{\mathsf{C}} : C \times P \to C. \tag{10}$$

The full system $\mathsf{S} : \mathbf{Sys}\binom{1}{E\times P\times C}$ *is the following composite autonomous system:*

$$\begin{aligned} \mathsf{upd}_{\mathsf{S}} : E \times P \times C &\longrightarrow E \times P \times C \\ (s_{\mathsf{E}}, s_{\mathsf{P}}, s_{\mathsf{C}}) &\longmapsto (\mathsf{upd}_{\mathsf{E}}(s_{\mathsf{E}}), \mathsf{upd}_{\mathsf{P}}(s_{\mathsf{P}}, s_{\mathsf{E}}, s_{\mathsf{C}}), \\ &\qquad \mathsf{upd}_{\mathsf{C}}(s_{\mathsf{C}}, s_{\mathsf{P}})). \end{aligned} \tag{11}$$

Let $S = E \times P \times C$ *denote the state space of* S.



IMP setup

Assumption 1 (Environment, plant, controller). *The following three components are so defined:*

1) *the environment* $\mathsf{E} : \mathbf{Sys}\binom{1}{E}$ *is an autonomous system*

$$\mathsf{upd}_\mathsf{E} : E \to E, \tag{8}$$

2) *the plant* $\mathsf{P} : \mathbf{Sys}\binom{E\times C}{P}$ *is a system*

$$\mathsf{upd}_\mathsf{P} : P \times E \times C \to P, \tag{9}$$

3) *the controller* $\mathsf{C} : \mathbf{Sys}\binom{P}{C}$ *is a system*

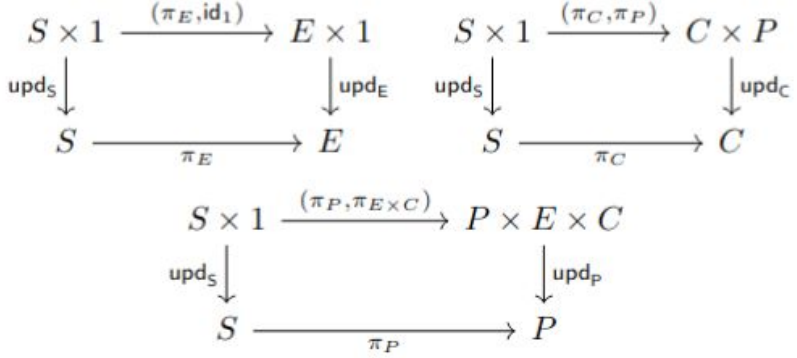
$$\mathsf{upd}_\mathsf{C} : C \times P \to C. \tag{10}$$

The full system $\mathsf{S} : \mathbf{Sys}\binom{1}{E\times P\times C}$ *is the following composite autonomous system:*

$$\begin{aligned} \mathsf{upd}_\mathsf{S} : E \times P \times C &\longrightarrow E \times P \times C \\ (s_\mathsf{E}, s_\mathsf{P}, s_\mathsf{C}) &\longmapsto (\mathsf{upd}_\mathsf{E}(s_\mathsf{E}), \mathsf{upd}_\mathsf{P}(s_\mathsf{P}, s_\mathsf{E}, s_\mathsf{C}), \\ &\qquad \mathsf{upd}_\mathsf{C}(s_\mathsf{C}, s_\mathsf{P})). \end{aligned} \tag{11}$$

Let $S = E \times P \times C$ *denote the state space of* S.

Remark II.7. There are maps of systems $\pi_\mathsf{E} : \mathsf{S} \to \mathsf{E}$, $\pi_\mathsf{P} : \mathsf{S} \to \mathsf{P}$ and $\pi_\mathsf{C} : \mathsf{S} \to \mathsf{C}$ induced by projecting out states of S:



(12)

IMP setup

Definition II.8 (Regulation problem)**.** A *regulation problem* (or *reguland* [1]) is a triple $(\mathsf{E}, \mathsf{P}, \tilde{K} \subseteq E \times P)$.

Assumption 2 (Regulation condition)**.** C *solves its regulation problem, meaning there exists an attracting subsystem (Definition II.6)* $\mathsf{S}^* \to \mathsf{S}$ *such that, on states,* $S^* \subseteq K$.

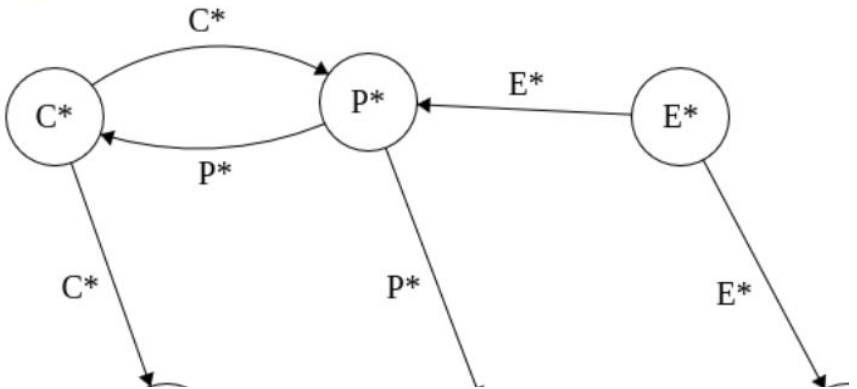
IMP setup

Even though K contains all states of C, this might not be true anymore for S^*. We can define the set of *attracting control states* as

$$C^* := \{s_\mathsf{C} \in C \mid \exists s_\mathsf{E} \in E, s_\mathsf{P} \in P, (s_\mathsf{E}, s_\mathsf{P}, s_\mathsf{C}) \in S^*\}.$$

It is obtained along with a surjection $\pi_{C^*} : S^* \to C^*$ by taking the image of S^* under the projection map $\pi_C : S \to C$:

$$\begin{array}{ccc} S^* & \overset{\pi_{C^*}}{\dashrightarrow} & C^* \\ \downarrow & & \vdots \\ S & \xrightarrow[\pi_C]{} & C \end{array}$$



Model

Definition II.9 (Model). A *model* of a system $\mathsf{X} \in \mathbf{Sys}\binom{I}{X}$ is:

- a system $\mathsf{M} \in \mathbf{Sys}\binom{J}{M}$ (the *archetype*), and
- a map of systems (the *model per se*)

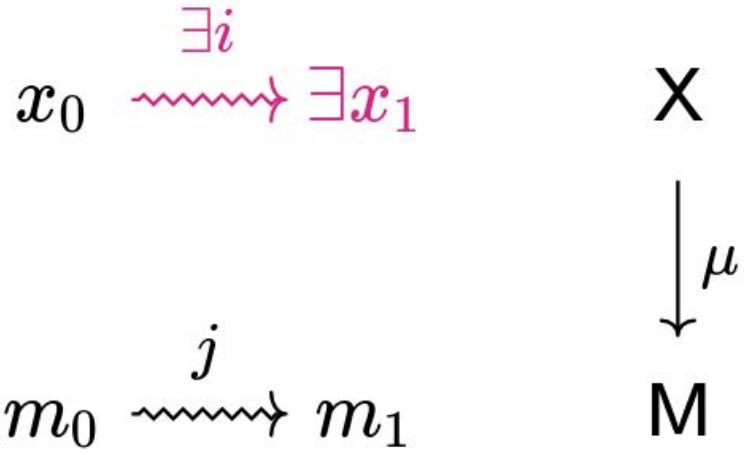
$$\mathsf{X} \xrightarrow{\mu} \mathsf{M} \tag{15}$$

such that

1) its part on states $\mu_{\mathsf{S}} : X \to M$ is surjective, and
2) its part on inputs $\mu_{\mathsf{I}}(x, -) : I \to J$ is surjective for each $x \in X$.

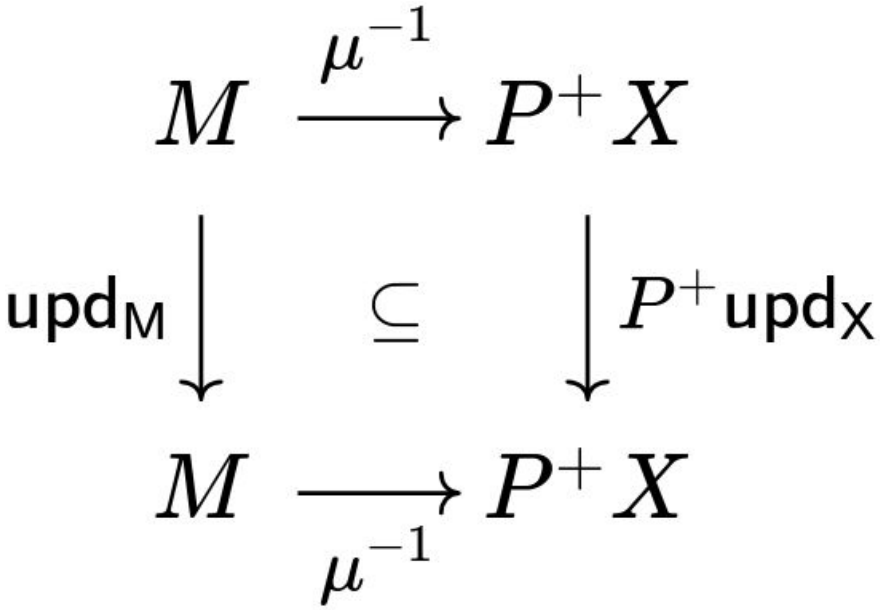
Model

Having a model $\mu : \mathsf{X} \to \mathsf{M}$ means that for each state $m \in M$ we have a set of states $\mu^{-1}(m) \in X$, called the *fibre* of m and which represents a subset of elements of X that are *indistinguishable* from the perspective of the simpler system M as they all map to the same element $m \in M$ via the surjective function μ_{s}. As $m \in M$ varies along the function $\mathsf{upd}_{\mathsf{M}}$, this variation is consistent with the variation described by the function $\mathsf{upd}_{\mathsf{X}}$ for each element x of the fibre $\mu^{-1}(m)$ of m.



everything we do in M can be reproduced in X

Model



The fibers of models are ‘dynamical beliefs’

Model

Remark II.11. Like any good definition, Definition II.9 admits a trivial instance. A *trivial model* is one where $\mu : X \to M$ is a product projection, which means there exists a system $\mathsf{F} \in \mathbf{Sys}\binom{H}{F}$ such that the state space and inputs of the system X factor as $X = M \times F$ and $I = J \times H$ respectively, *and* its dynamics decomposes in that of M and F:

$$\mathsf{upd}_{\mathsf{X}}((m, f), (j, h)) = (\mathsf{upd}_{\mathsf{M}}(m, j), \mathsf{upd}_{\mathsf{F}}(f, h)). \quad (17)$$

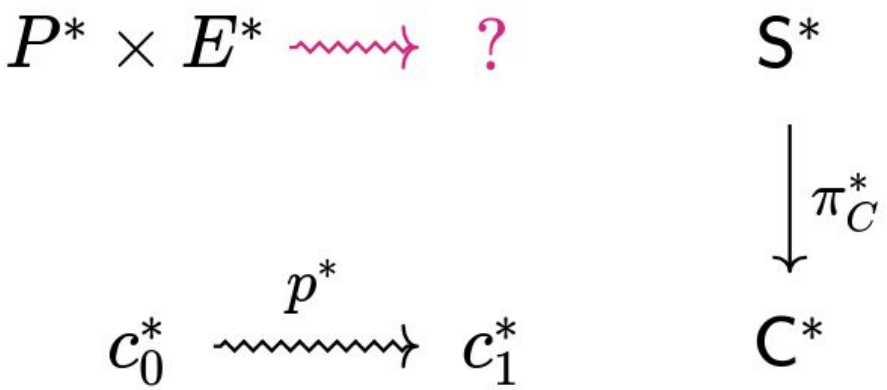
Thus, in a trivial model, knowledge of M does not afford knowledge about the rest of X, because M and F are not coupled.

$$C \times P \rightsquigarrow \mathsf{upd}_{\mathsf{S}}(C, P, e_0) \qquad \mathsf{S}$$

$$\downarrow \pi_E$$

$$e_0 \rightsquigarrow e_1 \qquad \mathsf{E}$$

Model

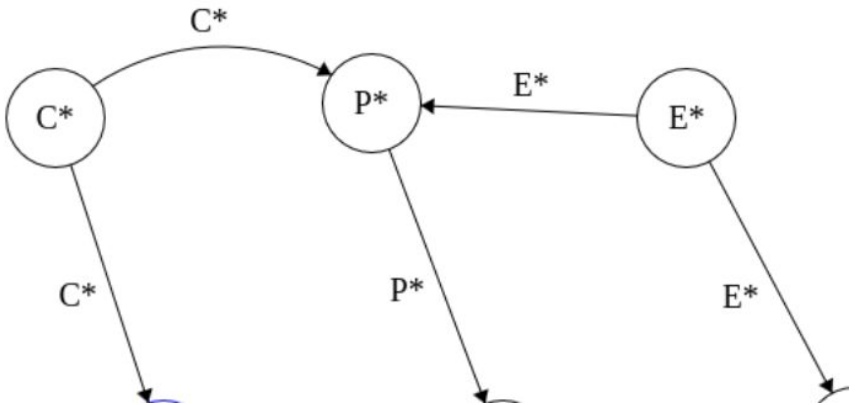


Doesn’t work!

IMP setup

Assumption 3 (Error feedback structure). *C^* supports an autonomous dynamics* $\mathsf{upd}_{C^*} : C^* \to C^*$, *thus defining a system* $\mathsf{C}^* : \mathbf{Sys}\left(\begin{smallmatrix}1\\C^*\end{smallmatrix}\right)$, *that we call attracting controller, and making* π_{C^*} *a full-fledged map of systems:*

$$\mathsf{S}^* \xrightarrow{\pi_{C^*}} \mathsf{C}^*$$



IMP v1

Theorem II.12 (*Internal Model Principle*). *Let* S *be a system subject to Assumptions 1 to 3. Then* C^* *models the attracting full system* S^* *via the projection* π_{C^*}, *as implied by Assumption 3.*

$$P^* \times E^* \rightsquigarrow \mathsf{upd}_{\mathsf{S}}(c_0^*, P^*, E^*) \qquad \mathsf{S}^*$$

$$\Big\downarrow \pi_C^*$$

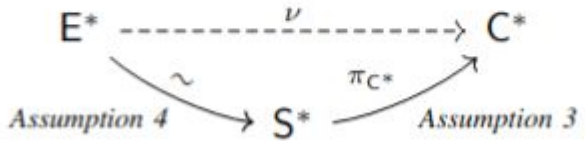
$$c_0^* \rightsquigarrow c_1^* \qquad \mathsf{C}^*$$

IMP v2 – Wonham & Hepburn's version

Assumption 4. *There is an isomorphism of systems* $\mathsf{S}^* \cong \mathsf{E}^*$, *meaning that for each environment state* $s_\mathsf{E} \in E^*$, *there is* exactly one $s \in S^*$ *such that* $\pi_E(s) = s_\mathsf{E}$.

"endomorph"

Theorem II.14 (*Internal Model Principle (Hepburn and Wonham)*). *Let* S *be a system subject to Assumptions 1 to 4. Then* C^* *models the attracting environment* E^* *via the dashed map below:*[2]



(18)

Thanks!

Bayesian interpretation

We now introduce the main ideas involved in the notion of Bayesian interpretations. To aid intuition, we refer to maps of type $p : I \nrightarrow X$ for any object X as *distributions* and to deterministic maps $x : I \to X$ of that type as *elements* of X. We also consider maps $\psi : \Theta \nrightarrow X$ to correspond to *parametrised families of distributions* (also called "channels" in [34], [65]), with each element $\theta : I \to \Theta$ a parameter determining a distribution $\psi(X \mid \theta)$ or $\psi(\theta)$ over X.

Furthermore, given a map $f : X \nrightarrow Y$ and an element $x : I \to Y$ we call $x \fatsemi f : I \nrightarrow Y$ the distribution over Y assigned to the element $x \in X$ by f. We write it as $f(x)$ or $f(Y \mid x)$. In **FinStoch** this terminology and the notation coincide with its common usage in probability theory.

over it. There is a Markov category **FinStoch** *whose objects are finite sets and whose morphisms from X to Y, $X \nrightarrow Y$, called Markov kernels, are defined as functions $X \to D(Y)$. For a Markov kernel $f : X \nrightarrow Y$ we write $f(y \mid x)$ for $f(x)(y)$, the probability assigned to y when the kernel is given x as input.*

Sequential composition of Markov kernels $f : X \nrightarrow Y$, $g : Y \nrightarrow Z$ is given by the Markov kernel so defined (i.e. the Chapman–Kolmogorov equation):

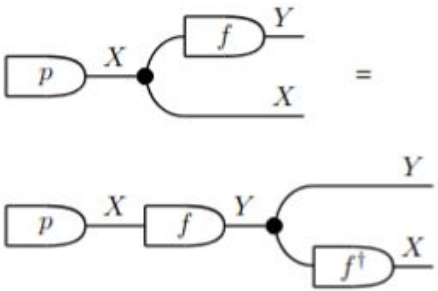
$$f \fatsemi g : X \nrightarrow Z$$
$$(f \fatsemi g)(z \mid x) := \sum_{y \in Y} g(z \mid y) f(y \mid x) \tag{34}$$

There is a Markov category $\mathbf{Rel}^+$ *whose objects are sets and whose morphisms from X to Y are functions $X \to P^+(Y)$, corresponding to left-total relations. Sequential composition of left-total relations $f : X \nrightarrow Y$, $g : Y \nrightarrow Z$ is given by the left-total relation so defined:*

$$f \fatsemi g : X \longrightarrow P^+(Z)$$
$$x \longmapsto \{z \in Z \mid \exists y \in Y, y \in f(x) \text{ and } z \in g(y)\}, \tag{36}$$

Bayesian interpretation

Definition III.6 (Bayesian inversion)**.** In a Markov category, a Bayesian inversion [34] of a map $f : X \to Y$ with respect to a distribution $p : I \to X$ is a map $f^\dagger : Y \to X$ satisfying the following equation:



(38)

To get a more concrete feeling for Bayesian inverses, we can examine the form Eq. (38) takes when our Markov category is **FinStoch**, where it corresponds to the equation

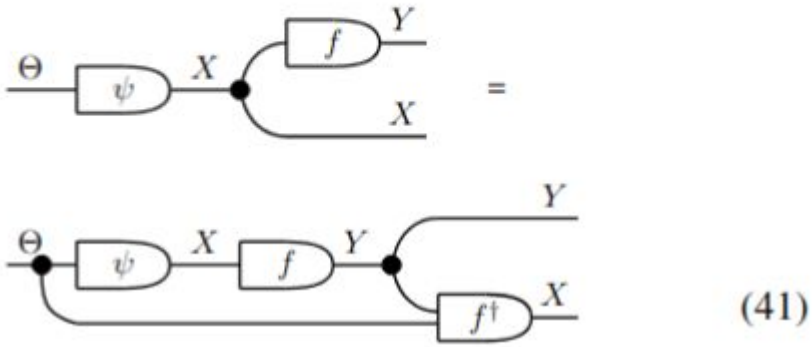
$$\begin{aligned} p(x)f(y \mid x) &= f(y \mid p(x))f^\dagger(x \mid y) \\ &= \sum_{x' \in X} p(x')f(y \mid x')f^\dagger(x \mid y), \end{aligned} \tag{39}$$

from which we can derive the standard Bayes rule by dividing by $f(y \mid p(x))$, assuming it's positive ($f(y \mid p(x)) > 0$):

$$\begin{aligned} f^\dagger(x \mid y) &= \frac{p(x)f(y \mid x)}{f(y \mid p(x))} \\ &= \frac{p(x)f(y \mid x)}{\sum_{x' \in X} p(x')f(y \mid x')}. \end{aligned} \tag{40}$$

Bayesian interpretation

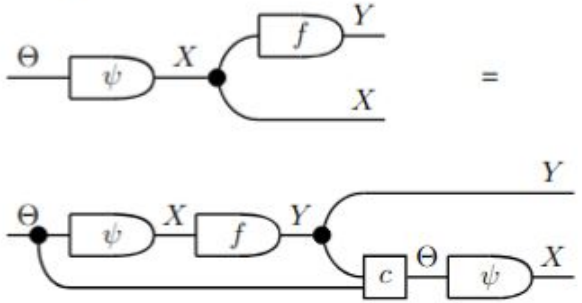
Definition III.7. In a Markov category, we say that a map $f^\dagger : Y \otimes \Theta \to X$ is a Bayesian inversion of a parametrised family $f : \Theta \to X$ if



(41)

Bayesian interpretation

Definition III.8 (Conjugate prior for Bayesian inference)**.** We call ψ a conjugate prior [65] to $f : X \nrightarrow Y$ if there exists a deterministic map $c : Y \otimes \Theta \to \Theta$ that satisfies the following:



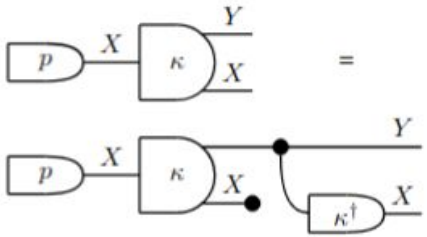
(42)

Note that the map $c \fatsemi \psi$ is a Bayesian inverse of f with respect to ψ, in the sense of Definition III.7. In Bayesian statistics, $\theta \in \Theta$ is sometimes called a *hyperparameter* and $x \in X$ a *parameter*, but we will generally avoid this usage.

If we have a conjugate prior $\psi : \Theta \nrightarrow X$ to $f : X \nrightarrow Y$, then every parameter $\theta : I \to \Theta$ determines a Bayesian inversion $\psi(c(-, \theta)) : Y \nrightarrow X$ of $f : X \nrightarrow Y$ for

Bayesian interpretation

We need then a form of belief updating that corresponds to *Bayesian filtering*, where the hidden variable also changes. For this we can replace $\Delta_X \fatsemi f \otimes \mathrm{id}_X : X \to Y \otimes X$ with a map $\kappa : X \to X \otimes Y$ that produces an observation Y and may change X instead of just copying it. The analogue of the Bayesian inversion (Definition III.6) with respect to a distribution $p : I \to X$ is thus a map $\kappa^\dagger : Y \to X$ satisfying



(43)

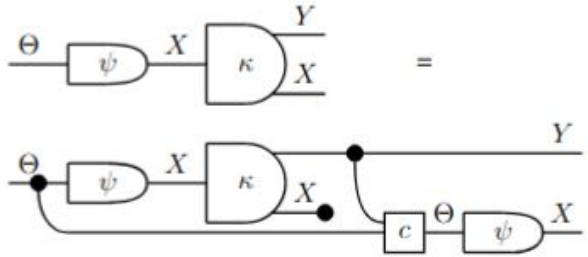
As with ordinary Bayesian inversions (Definition III.6), the map $\kappa^\dagger$ always exists in any Markov category with conditionals, including thus **FinStoch** and $\mathbf{Rel}^+$. Note that setting $\kappa = \Delta_X \fatsemi f \otimes \mathrm{id}_X : X \to Y \otimes X$ recovers Definition III.6.

In Bayesian filtering, the idea is that κ updates X and generates an observation Y. Similarly to the Bayesian inference

$$\kappa^\dagger(x \mid y) = \frac{\sum_{x' \in X} p(x')\kappa(y, x \mid x')}{\sum_{x', x'' \in X} p(x')\kappa(y, x'' \mid x')}.$$

Bayesian interpretation

Definition III.9 (Conjugate prior for Bayesian filtering)**.** We call ψ a conjugate prior for Bayesian filtering to $\kappa : X \to X \otimes Y$ if there exists a deterministic map $c : Y \otimes \Theta \to \Theta$ that satisfies the following:

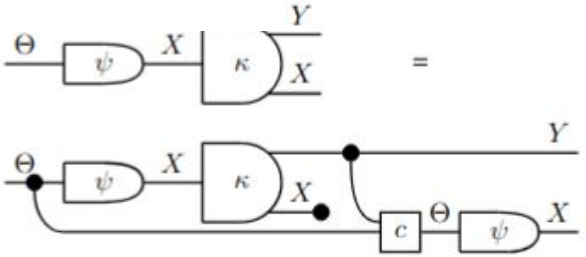


(45)

The map $c : Y \otimes \Theta \to \Theta$ maps parameters of prior distributions over X to parameters of posterior distributions over *updated* X (cf. Definition III.8 where X is instead fixed), while taking into account observations generated by κ. In a probabilistic context (including in **FinStoch**), the map κ where X is updated is often called a *hidden Markov model* (or a *discrete state-space model*), as opposed to the statistical model f where X is fixed.

Bayesian interpretation

Definition III.10 (Bayesian filtering interpretation [33]). Given a deterministic map $c : Y \otimes \Theta \to \Theta$, a *Bayesian filtering interpretation* of c consists of a map $\psi : \Theta \dashrightarrow X$ called the *interpretation map*, together with a map $\kappa : X \dashrightarrow X \otimes Y$ called *hidden Markov model*, such that Eq. (45) holds. In this context, Eq. (45) is called the *consistency equation*. A map $c : Y \otimes \Theta \to \Theta$ together with an interpretation (ψ, κ) is called a *reasoner*.

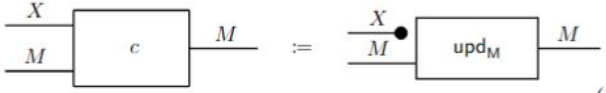


(45)

Virgo, Nathaniel, Martin Biehl, and Simon McGregor. 2021. 'Interpreting Dynamical Systems as Bayesian Reasoners'. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, https://doi.org/10.1007/978-3-030-93736-2_52.

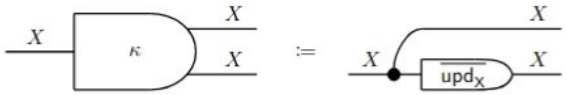
Bayesian interpretation

Theorem IV.5. *Let* M *model* X *with* $\mu : \mathsf{X} \to \mathsf{M}$*, and assume* M *and* X *are autonomous. Define* $c : X \otimes M \to M$ *as*



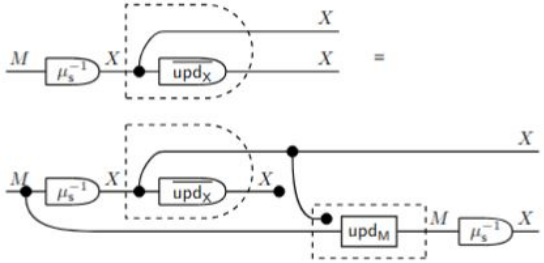
(52)

and $\kappa : X \to X \otimes X$ *as:*



(53)

Then κ *is the hidden Markov model, and* $\mu_s^{-1} : M \to X$ *the interpretation map of a Bayesian filtering interpretation of* c*, i.e. we have:*



(54)

where the dashed lines show, informally, where we replaced the definitions above in Eq. (45).

Thanks again!