Towards Information Theory-Based Discovery of Equivariances

Hippolyte Charvin H,CHARVIN@HERTS.AC.UK
Nicola Catenacci Volpi N.CATENACCI-VOLPI@HERTS.AC.UK
Daniel Polani D.POLANI@HERTS.AC.UK
Adaptive Systems Research Group, University of Hertfordshire

Editors: Sophia Sanborn, Christian Shewmake, Simone Azeglio, Nina Miolane

Abstract

The presence of symmetries imposes a stringent set of constraints on a system. This constrained structure allows intelligent agents interacting with such a system to drastically improve the efficiency of learning and generalization, through the internalisation of the system's symmetries into their information-processing. In parallel, principled models of complexity-constrained learning and behaviour make increasing use of information-theoretic methods. Here, we wish to marry these two perspectives and understand whether and in which form the information-theoretic lens can "see" the effect of symmetries of a system. For this purpose, we propose a novel variant of the Information Bottleneck principle, which has served as a productive basis for many principled studies of learning and information-constrained adaptive behaviour. We show (in the discrete case) that our approach formalises a certain duality between symmetry and information parsimony: namely, channel equivariances can be characterised by the optimal *mutual information-preserving joint compression* of the channel's input and output. This information-theoretic treatment furthermore suggests a principled notion of "soft" equivariance, whose "coarseness" is measured by the amount of input-output mutual information preserved by the corresponding optimal compression. This new notion offers a bridge between the field of bounded rationality and the study of symmetries in neural representations. The framework may also allow (exact and soft) equivariances to be automatically discovered.

Keywords: Channel equivariances, Information Bottleneck, Symmetry Discovery.

1. Introduction

Our work is motivated by a programme of formalising the relationship between the presence of coherent structures in an environment, and the informational efficiency that these structures make possible for an (artificial or biological) agent that learns and interacts with them. Our intuition is that there is a fundamental duality between structure and information: in short, any structure in a system affords a possibility of informational efficiency to an agent interacting with it, and every improvement in an agent's informational efficiency must exploit some kind of structure in the system it interacts with.

As a first step towards the operationalisation of this intuition, we focus on a specific kind of structure: symmetries, and, more precisely, the *equivariances* of probabilistic channels (Bloem-Reddy and Teh, 2020). We seek to first design a formal method to identify the duality between equivariances and information, and will leave the modeling of concrete systems

to future work. Previous results (Achille and Soatto, 2018) exhibited links between invariance extraction and the Information Bottleneck (IB) method (Tishby et al., 2000), which optimally compresses one variable under the constraint of preserving information about a second variable. Here, we adapt this idea to the more general context of equivariances, which increasingly appear crucial to efficient learning and generalisation (Higgins et al., 2022). We propose an extension6 of the IB method whose solutions indeed characterise the equivariances of discrete probabilistic channels. This characterisation provides, as far as we are aware, a novel and intuitively appealing point of view on equivariances, through the notion of *mutual information-preserving optimal joint compression* of the channel's input and output. Namely, our result characterises equivariances as the pairs of transformations made indiscernible from the identity by such a compression.

However, to eventually grasp real-world symmetries, which might be much less stringent than mathematical equivariances in the classic, "exact" sense, we need to consider "soft" notions of equivariance. The problem then arises of *how to measure the "divergence"* from being an exact equivariance. Here, we build on our new characterisation of exact equivariances to define the "coarseness" of soft equivariances through the resolution of the informationally optimal compression that they make possible. Namely, soft equivariances of "granularity" λ are defined as pairs of transformations made indiscernible from the identity by an optimal compression which *partially* preserves the channel's input-output mutual information, to a degree specified by λ.

This information-theoretic point of view on equivariances links the study of symmetries in biological and artificial agents to the field of bounded rationality (Genewein et al., 2015), through the duality between informationally optimal representations and the corresponding extracted equivariances. But crucially, this method might also allow one to *discover* soft equivariances: we will sketch a roadmap towards computing equivariances as defined here.

Assumptions and notations: We fix finite sets $\mathcal{X}$ and $\mathcal{Y}$ and a *fully supported* probability $p(X, Y)$ on $\mathcal{X} \times \mathcal{Y}$.[1] "Bottlenecks" are variables T defined on $\mathcal{T} := \mathbb{N}$. The probability simplex defined by a finite set $\mathcal{A}$ is denoted by $\Delta_{\mathcal{A}}$. Conditional probabilities, also called channels, will often be regarded as functions between probability simplices, or as linear maps between vector spaces (e.g., a channel from $\{1, \dots, n\}$ to itself can be regarded as a function from $\Delta_{\{1,\dots,n\}}$ to itself, or as linear map from $\mathbb{R}^n$ to itself). The set of channels with input space $\mathcal{A}$ and output space $\mathcal{B}$, resp. output space $\mathcal{A}$ itself, are denoted by $C(\mathcal{A}, \mathcal{B})$, resp. $C(\mathcal{A})$. The set of bijections of $\mathcal{A}$ is $\mathrm{Bij}(\mathcal{A})$, and for $\gamma \in \mathrm{Bij}(\mathcal{A})$, $a \in \mathcal{A}$, we write $\gamma \cdot a := \gamma(a)$. The identity map on $\mathcal{A}$ is written $e_{\mathcal{A}}$. The symbol $\circ$ denotes function composition, resp. channel composition, depending on the context (functions are seen as deterministic channels when they are composed with another channel). The symbol δ_P means 1 when the proposition P is true, and 0 otherwise. $D(\cdot||\cdot)$ is the Kullback-Leibler divergence.

1. For now, we work under the hypothesis that in real-world scenarios, there will typically be at least some noise spillover into all possible configurations. We leave to future work a generalisation to non-fully supported $p(X, Y)$ (see Remark 18 in Appendix B.3) and to non-finite $p(X, Y)$ (see Appendix C).

2. The Intertwining Information Bottleneck and exact equivariances

Definition 1 *An (exact)* equivariance *of the channel $p(Y|X)$ is a pair of deterministic permutations $(\sigma, \tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$ such that $p(Y|X) \circ \sigma = \tau \circ p(Y|X)$. An* invariance *of $p(Y|X)$ is some $\sigma \in \mathrm{Bij}(\mathcal{X})$ such that $p(Y|X) \circ \sigma = p(Y|X)$.*

It can be easily verified that the set of equivariances of $p(Y|X)$ is a group for the relation $(\sigma, \tau) \cdot (\sigma', \tau') := (\sigma \circ \sigma', \tau \circ \tau')$. This group will be called the *equivariance group* of $p(Y|X)$, and be denoted $G_{p(Y|X)}$. Now, in the IB method, which, as mentioned above, has been suggested to extract channel invariances, one considers a pair of variables X and Y, but the compressed variable is a function of only one of them, say X, whereas it preserves information about the second variable Y. This is consistent with the idea that the IB might extract invariances, because the latter transform only the space $\mathcal{X}$. However, equivariances clearly transform *both* spaces $\mathcal{X}$ and $\mathcal{Y}$, so that a compression that has any hope of extracting these equivariances should be a function of *both* X and Y. For the same reason, it does not seem natural that, here, the preserved information should be either only that about X, or only that about Y. Rather, we want to formalise the following intuition: the presence of (exact, resp. soft) equivariances of $p(X, Y)$ should correspond to the possibility of compressing the joint variable (X, Y) in a way that (fully, resp. partially) preserves the *mutual* information $I(X;Y) := D(p(X,Y)||p(X)p(Y))$. Thus we propose to consider what we call the *Intertwining Information Bottleneck* (IIB), defined for every $0 \leq \lambda \leq I(X;Y)$:

$$\underset{\substack{\kappa \in C(\mathcal{X}\times\mathcal{Y},\mathcal{T}):\\ D(\kappa(p(X,Y))||\kappa(p(X)p(Y))) = \lambda}}{\arg\min} I_\kappa(X, Y; T), \tag{1}$$

where the mutual information $I_\kappa(X, Y; T)$ is computed from the distribution $p(x, y)\kappa(t|x, y)$. The constraint in (1) means that the channel κ must conserve the divergence between $p(X, Y)$ and its split version $p(X)p(Y)$, to the level specified by λ. On the other hand, the minimisation of $I_\kappa(X, Y; T)$ means that κ implements, under the latter constraint, an optimal compression. In particular, the solutions to (1) for $\lambda = I(X;Y)$ formalise the intuition of largest possible compression of the pair (X, Y) that still preserves the mutual information between these variables. Importantly, both the IB and the Symmetric IB (Slonim et al., 2006) can be recovered from the IIB problem by adding the right constraint on the shape of κ in (1). If we add the requirement that κ can only compress the $\mathcal{X}$ coordinate, we recover the IB problem with source X and relevancy Y; while if we rather impose that κ must compress $\mathcal{X}$ and $\mathcal{Y}$ separately, we recover the Symmetric IB problem (see Appendix A).

Given the structural similarity between (1) and the IB problem, the algorithms for computing the latter might be adaptable to the former. In particular, we leave to future work to prove the convergence of, and implement, an adapted version of the Blahut-Arimoto algorithm used for the IB (Tishby et al., 2000). Another possibility would be to identify, and optimise for, variational bounds (Alemi et al., 2019) on the information quantities from (1). Note that for $\lambda = I(X;Y)$, the set of solutions can be computed explicitly, and, up to trivial transformations, it consists of a unique deterministic clustering (see Corollary 10 in Appendix B.1).

Let us now formalise our intuition of duality between the (exact) equivariance group $G_{p(Y|X)}$ and the information compression that the latter makes possible. To state this result,

let us define, for a pair $(\sigma, \tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$, the *tensor product* $\sigma \otimes \tau : (x, y) \mapsto (\sigma \cdot x, \tau \cdot y)$. We will also need to consider the projection of the equivariance group $G_{p(Y|X)}$ on the space $\mathrm{Bij}(\mathcal{X})$ of input transformations, i.e.,

$$G^{\mathcal{X}}_{p(Y|X)} := \{\sigma \in \mathrm{Bij}(\mathcal{X}) \ : \ \exists \tau \in \mathrm{Bij}(\mathcal{Y}),\ (\sigma, \tau) \in G_{p(Y|X)}\}, \tag{2}$$

and the following notion (Bloem-Reddy and Teh, 2020):

Definition 2 *Let G be a group acting on a finite set $\mathcal{A}$. A probability distribution $p(A)$ on $\mathcal{A}$ is said* G-exchangeable *if for all $a \in \mathcal{A}$, we have $p(g \cdot a) = p(a)$ — i.e., if $p(A)$ is uniform on every orbit of the action of G on $\mathcal{A}$.*

Theorem 3 *Assume that $p(X)$ is $G^{\mathcal{X}}_{p(Y|X)}$-exchangeable, and let $\kappa \in C(\mathcal{X} \times \mathcal{Y}, \mathcal{T})$ be a solution to the IIB problem for $\lambda = I(X;Y)$. Then a pair $(\sigma, \tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$ is an equivariance of $p(Y|X)$ if and only if*

$$\kappa \circ (\sigma \otimes \tau) = \kappa. \tag{3}$$

Proof See Appendix B. ■

Intuitively, the essentially unique solution κ to the IIB for $\lambda = I(X;Y)$ is the deterministic coarse-graining of the product space $\mathcal{X} \times \mathcal{Y}$ satisfying the following property: a pair of permutations $(\sigma, \tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$ is an equivariance of $p(X, Y)$ if and only if this coarse-graining "filters out" the effect of simultaneously transforming $\mathcal{X}$ with σ and $\mathcal{Y}$ with τ — thus making the pair (σ, τ) indiscernible from the identity on $\mathcal{X} \times \mathcal{Y}$. In particular, the equivariance group of $p(X, Y)$ is characterised by the optimal compression of the joint variable (X, Y) that still preserves the mutual information $I(X;Y)$. Note that the assumption of $G^{\mathcal{X}}_{p(Y|X)}$-exchangeability on $p(X)$ means intuitively that the input distribution "respects the symmetries" of the channel $p(Y|X)$. In particular, this assumption is satisfied if $p(X)$ is uniform, but also if $p(X)$ achieves the capacity of the channel $p(Y|X)$ — at least in the case where $p(Y|X)$ defines an injective transition matrix.[2]

3. Towards soft equivariances discovery

To soften the notion of channel equivariance, we first allow the transformations on resp. $\mathcal{X}$ and $\mathcal{Y}$ to be non-invertible and stochastic. But more importantly, we have to choose *the right notion of "divergence"* from Definition 1's exact equivariance requirement being achieved. Following the dual point of view developed in Section 2, we assume, intuitively, that soft equivariances should be characterised by an optimal compression of (X, Y) under the constraint of, here, *partially* preserving $I(X;Y)$. To make the statement precise, let us define, for $\mu \in C(\mathcal{X})$ and $\eta \in C(\mathcal{Y})$, the tensor product $\mu \otimes \eta(x', y'|x, y) := \mu(x'|x)\eta(y'|y)$.

Definition 4 *Let $p(Y|X)$ be given, let $0 \leq \lambda \leq I(X;Y)$, and let κ be a solution to the IIB problem (1) with uniform[3] $p(X)$ and parameter λ. A* (λ, κ)-equivariance *of $p(X, Y)$ is*

2. See Theorem III.1 in (Pernice, 2022).
3. In view of Theorem 3, it might seem more natural to require $p(X)$ to be only $G^{\mathcal{X}}_{p(Y|X)}$-exchangeable. But at this stage, it is not clear yet that this would result in a well-posed definition (see Appendix D).

a pair $(\mu, \eta) \in C(\mathcal{X}) \times C(\mathcal{Y})$ *such that*

$$\kappa \circ (\mu \otimes \eta) = \kappa. \tag{4}$$

We will also call a pair (μ, η) *a* λ-equivariance *if there exists some solution* κ *to the IIB problem* (1)*, with parameter* λ*, such that* (μ, η) *is a* (λ, κ)*-equivariance.*

Intuitively, a pair (μ, η) is a (λ, κ)-equivariance if the channel κ, which implements a joint optimal compression of X and Y under the constraint of partially preserving their mutual information, "filters out" the simultaneous stochastic transformations of $\mathcal{X}$ through μ and $\mathcal{Y}$ through η — thus making (μ, η) indiscernible from the identity on $\mathcal{X} \times \mathcal{Y}$. Moreover, it is clear from Theorem 3 that exact equivariances are λ-equivariances with $\lambda = I(X;Y)$.

For fixed λ and corresponding κ, the set of (λ, κ)-equivariances is clearly a *semigroup*. Intuitively, we expect this semigroup to get larger when λ decreases: indeed, the IIB channel κ then enforces a larger compression of X and Y, thus allowing more transformations $\mu \otimes \eta$ of $\mathcal{X} \times \mathcal{Y}$ to be "filtered out" by this compression. More precisely, equation (4) is equivalent to $\mathrm{Im}(\mu \otimes \eta - e_{\mathcal{X}\times\mathcal{Y}}) \subseteq \ker(\kappa)$,[4] and we conjecture that the dimension of $\ker(\kappa)$ increases for decreasing λ, thus allowing it to contain the image of more tranformations of the form $\mu \otimes \eta - e_{\mathcal{X}\times\mathcal{Y}}$. Note for instance that for $\lambda = 0$, the IIB solutions are the channels κ such that $\kappa(T|x, y)$ does not depend on (x, y). Their kernel is the direction of the whole simplex $\Delta_{\mathcal{X}\times\mathcal{Y}}$, so that the corresponding set of $(0, \kappa)$-equivariances is the whole $C(\mathcal{X}) \otimes C(\mathcal{Y})$.

Now, assuming that a solution κ to the IIB is known, how can we explicitly compute the corresponding (λ, κ)-equivariances? The equation (4) which defines soft equivariances is a polynomial equation, made of quadratic homogeneous polynomials — more precisely, linear combinations of elements of the form $\mu_{x',x}\eta_{y',y}$. To this homogeneous polynomial equation, we must add the requirement that μ and η are conditional probabilities: i.e., they must satisfy the linear equations $\sum_{x'} \mu_{x',x} = 1$ and $\sum_{y'} \eta_{y',y} = 1$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, along with the linear inequalities defining the non-negativity constraints. Overall, the pair of real matrices (μ, η) that satisfy the conditions of Definition 4 thus correspond to the intersection of the positive orthant $\{\forall x, x' \in \mathcal{X},\ \forall y, y' \in \mathcal{Y},\ \mu_{x',x} \geq 0,\ \eta_{y',y} \geq 0\}$ with the solutions of a degree 2 polynomial system of equations. We leave to future work a more involved study of this problem, and of algorithms that might solve it.

As a first step for assessing the relevance of our method to equivariance discovery, one could also study scenarios where specific exact equivariances are known, and verify that IIB solutions do "filter them out" — in the sense of equation (4). If this is the case, one could then perturb the channel $p(Y|X)$, and investigate whether the exact equivariances of the unperturbed channel are still soft equivariances of the perturbed channel — still in the sense of equation (4).

In short, in this work we have formalised the duality between channel equivariances and the informational efficiency that they make possible for capturing the relationship between the channel's input and output. We achieved this with a novel extension of the IB principle, which leads to a principled generalisation of exact equivariances into "soft" ones. The proposed approach might help understand the emergence of symmetries in neural systems through the lens of information parsimony, and potentially opens a new path towards the automatic discovery of exact and soft equivariances.

4. Here, the discrete conditional probabilities are seen as transition matrices acting on real vectors.

Funding H.C. and D.P. were funded by the Pazy Foundation under grant ID 195.

Acknowledgements D.P. thanks Naftali Tishby for early discussions leading to the present studies.

References

Alessandro Achille and Stefano Soatto. Emergence of Invariance and Disentanglement in Deep Representations. pages 1–9, February 2018. doi: 10.1109/ITA.2018.8503149.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep Variational Information Bottleneck, October 2019. Comment: 19 pages, 8 figures, Accepted to ICLR17.

Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information Geometry*, volume 64 of *Ergebnisse Der Mathematik Und Ihrer Grenzgebiete 34*. Springer International Publishing, Cham, 2017. ISBN 978-3-319-56477-7 978-3-319-56478-4. doi: 10.1007/978-3-319-56478-4.

Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, NY, 3. ed edition, 1995. ISBN 978-0-471-00710-4.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *J. Mach. Learn. Res*, 21:61, January 2020. ISSN 1532-4435.

Imre Csiszár and János Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, Cambridge, 2 edition, 2011. ISBN 978-0-521-19681-9. doi: 10.1017/CBO9780511921889.

Tim Genewein, Felix Leibfried, Jordi Grau-Moya, and Daniel Braun. Bounded Rationality, Abstraction, and Hierarchical Decision-Making: An Information-Theoretic Optimality Principle. *Frontiers in Robotics and AI*, 2, November 2015. doi: 10.3389/frobt.2015.00027.

Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. An Information Theoretic Tradeoff between Complexity and Accuracy. In Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, Bernhard Schölkopf, and Manfred K. Warmuth, editors, *Learning Theory and Kernel Machines*, volume 2777, pages 595–609. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. ISBN 978-3-540-40720-1 978-3-540-45167-9. doi: 10.1007/978-3-540-45167-9_43.

Robert M. Gray. *Entropy and Information Theory*. Springer New York, NY, 2 edition, September 2014. ISBN 978-1-4899-8132-5.

Irina Higgins, Sébastien Racanière, and Danilo Rezende. Symmetry-Based Representations for Artificial and Biological General Intelligence. *Frontiers in Computational Neuroscience*, 16, 2022. ISSN 1662-5188.

Olav Kallenberg. *Random Measures, Theory and Applications*, volume 77 of *Probability Theory and Stochastic Modelling*. Springer International Publishing, Cham, 2017. ISBN 978-3-319-41596-3 978-3-319-41598-7. doi: 10.1007/978-3-319-41598-7.

Claude Lemaréchal. Lagrangian Relaxation. In Michael Jünger and Denis Naddef, editors, *Computational Combinatorial Optimization: Optimal or Provably Near-Optimal Solutions*, pages 112–156. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001. ISBN 978-3-540-45586-8. doi: 10.1007/3-540-45586-8_4.

Francisco Pernice. On the Symmetries of the Deletion Channel. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–6, Monticello, IL, USA, September 2022. IEEE. ISBN 9798350399981. doi: 10.1109/Allerton49937.2022.9929324.

Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, Inc., January 1987.

Noam Slonim, Nir Friedman, and Naftali Tishby. Multivariate Information Bottleneck. *Neural Computation*, 18(8):1739–1789, August 2006. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.2006.18.8.1739.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, April 2000.

Appendix A. Relation between IB, Symmetric IB and Intertwining IB

Let us start with the following lemma, which will prove useful below:

Lemma 5 *Let f and g be continuous real functions defined on a convex subspace C of a topological vector space, such that g is convex and non-negative, the image of g contains 0, and $g^{-1}(0) \subseteq f^{-1}(0)$. Let $\lambda \geq 0$, and consider the constrained optimisation problem*

$$\arg\min_{\substack{v \in C: \\ f(v) \geq \lambda}} g(v). \tag{5}$$

Then every solution v to (5) (i.e., every minimiser of (5)) must satisfy $f(v) = \lambda$. In other words, the set of solutions to (5) coincides with the set of solutions to

$$\arg\min_{\substack{v \in C: \\ f(v) = \lambda}} g(v).$$

Proof If f is bounded from above by λ, then a solution v to (5) must satisfy both $f(v) \geq \lambda$ and $f(v) \leq \lambda$, so that $f(v) = \lambda$ and the proof is done. Let us thus consider a vector $v \in C$ such that $f(v) > \lambda$, and fix also some $v_0 \in g^{-1}(0)$. By convexity of g, for all $0 < \epsilon < 1$, we have, with $v^\epsilon := \epsilon v_0 + (1-\epsilon)v \in C$,

$$\begin{aligned} g(v^\epsilon) &\leq \epsilon g(v_0) + (1-\epsilon)g(v) = (1-\epsilon)g(v) \\ &< g(v), \end{aligned} \tag{6}$$

where the equality comes from $g(v_0) = 0$, and the last inequality uses the fact that, because of the assumption $g^{-1}(0) \subseteq f^{-1}(0)$ and $f(v) > \lambda \geq 0$, we must have $g(v) \neq 0$ — i.e., taking into account the non-negativity assumption, $g(v) > 0$. Moreover for small enough ϵ, by continuity of f, the inequality $f(v) > \lambda$ implies that $f(v^\epsilon) \geq \lambda$.

Therefore, we proved that whenever $f(v) > \lambda$, there exists some $v^\epsilon \in C$ satisfying both $f(v^\epsilon) \geq \lambda$ and $g(v^\epsilon) < g(v)$: i.e., $g(v)$ cannot be a minimum of (5). In other words, for v to achieve the minimum in (5), the condition $f(v) = \lambda$ is necessary — which means that the inequality in (5) can be replaced by an equality. ■

A.1. IIB and classic IB

We want to impose, in the IIB problem (1), an additional restriction on κ that reduces the latter problem to the Information Bottleneck (IB) problem with source X and relevancy Y, i.e., (Gilad-Bachrach et al., 2003)

$$\underset{\substack{q(T_{\mathrm{IB}}|X) \in C(\mathcal{X}, \mathcal{T}_{\mathrm{IB}}) : \\ I_q(Y;T_{\mathrm{IB}}) \geq \lambda}}{\arg\min} I_q(X; T_{\mathrm{IB}}), \tag{7}$$

where T_{IB} is defined on $\mathcal{T}_{\mathrm{IB}} := \mathbb{N}$, and $I_q(Y; T_{\mathrm{IB}})$ is computed from the marginal $q(Y, T_{\mathrm{IB}})$ of the extension $q(X, Y, T_{\mathrm{IB}})$ of $p(X, Y)$ defined through the Markov chain condition $T_{\mathrm{IB}} - X - Y$, i.e., $q(x, y, t_{\mathrm{IB}}) := p(x, y)q(t_{\mathrm{IB}}|x)$. Let us define the set

$$C_{\mathrm{IB}(X,Y)} := \{\kappa_{\mathcal{X}} \otimes e_{\mathcal{Y}} \ : \ \ \kappa_{\mathcal{X}} \in C(\mathcal{X}, \mathcal{T}_{\mathrm{IB}})\} \ \subset \ C(\mathcal{X} \times \mathcal{Y}, \mathcal{T}_{\mathrm{IB}} \times \mathcal{Y}).$$

of channels that can compress the $\mathcal{X}$ coordinate but leave the $\mathcal{Y}$ coordinate unchanged. Note that for such channels, the output T, defined on $\mathcal{T}_{\mathrm{IB}} \times \mathcal{Y}$, can be written $T = (T_{\mathrm{IB}}, Y')$, where T_{IB} is defined on $\mathcal{T}_{\mathrm{IB}}$ and Y' is a copy of Y.[5] We now consider the problem

$$\underset{\substack{\kappa \in C_{\mathrm{IB}(X,Y)} : \\ D(\kappa(p(X,Y))||\kappa(p(X)p(Y)))=\lambda}}{\arg\min} I_\kappa(X, Y; T), \tag{8}$$

which is the IIB problem (1) where we added the constraint that κ must be of the form $\kappa_{\mathcal{X}} \otimes e_{\mathcal{Y}}$. It turns out that (8) does coincide with the IB problem, in the following sense:

Proposition 6 *For every $0 \leq \lambda \leq I(X;Y)$, a channel $\kappa_{\mathcal{X}} \otimes e_{\mathcal{Y}} \in C_{IB}(X, Y)$ solves the problem (8) if and only if $\kappa_{\mathcal{X}} = \kappa_{\mathcal{X}}(T_{IB}|X)$ solves the IB problem (7).*

Crucially, note that here $\kappa_{\mathcal{X}} \otimes e_{\mathcal{Y}} \in C_{\mathrm{IB}}(X, Y)$ is entirely determined by $\kappa_{\mathcal{X}}$ through its tensor product with the fixed identity channel $e_{\mathcal{Y}}$, while conversely, $\kappa_{\mathcal{X}}$ is entirely determined by $\kappa_{\mathcal{X}} \otimes e_{\mathcal{Y}}$ through the marginalisation relation

$$\kappa_{\mathcal{X}}(t_{\mathrm{IB}}|x) = \sum_{y'} \kappa_{\mathcal{X}}(t_{\mathrm{IB}}|x)p(y') = \sum_{y,y'} \kappa_{\mathcal{X}}(t_{\mathrm{IB}}|x)\delta_{y=y'}p(y) = \sum_{y,y'} \kappa_{\mathcal{X}} \otimes e_{\mathcal{Y}}(t_{\mathrm{IB}}, y'|x, y)p(y).$$

Informally, the only difference between $\kappa_{\mathcal{X}}$ and $\kappa_{\mathcal{X}} \otimes e_{\mathcal{Y}}$ is that $\kappa_{\mathcal{X}} \otimes e_{\mathcal{Y}}$ concatenates the output of $\kappa_{\mathcal{X}}$ with a copy of Y. Let us now prove Proposition 6.

Proof For $\kappa = \kappa_{\mathcal{X}} \otimes e_{\mathcal{Y}} \in C_{\mathrm{IB}}(X, Y)$, let us write $q(X, Y, T_{\mathrm{IB}}, Y')$ the distribution defined by

$$q(x, y, t_{\mathrm{IB}}, y') := p(x, y)\kappa(t_{\mathrm{IB}}, y'|x, y) = p(x, y)\kappa_{\mathcal{X}}(t_{\mathrm{IB}}|x)\delta_{y'=y}. \tag{9}$$

5. As we defined $\mathcal{T} := \mathbb{N}$, $\mathcal{T}_{\mathrm{IB}} := \mathbb{N}$ and as there is a bijection between $\mathbb{N} \times \mathcal{Y}$ and $\mathbb{N}$, writing here the bottleneck space as $\mathcal{T}_{\mathrm{IB}} \times \mathcal{Y}$ rather than $\mathcal{T}$ is just a difference of presentation.

It can be easily verified that then $\kappa(p(X,Y)) = q(T_{\mathrm{IB}}, Y)$ and $\kappa(p(X)p(Y)) = q(T_{\mathrm{IB}})p(Y)$, so that

$$D(\kappa(p(X,Y))||\kappa(p(X)p(Y))) = D(q(T_{\mathrm{IB}}, Y)||q(T_{\mathrm{IB}})p(Y))) = I_q(T_{\mathrm{IB}}; Y). \tag{10}$$

On the other hand,

$$\begin{aligned} I_\kappa(X, Y; T) &= I_q(X, Y; T_{\mathrm{IB}}, Y') \\ &= I_q(X, Y; T_{\mathrm{IB}}) + I_q(X, Y; Y'|T_{\mathrm{IB}}) && (11) \\ &= I_q(X; T_{\mathrm{IB}}) + I_q(Y; Y'|T_{\mathrm{IB}}) + I(X; Y'|T_{\mathrm{IB}}, Y) && (12) \\ &= I_q(X; T_{\mathrm{IB}}) + H(Y|T_{\mathrm{IB}}) && (13) \\ &= I_q(X; T_{\mathrm{IB}}) - I(Y; T_{\mathrm{IB}}) + H(Y), && (14) \end{aligned}$$

where line (11) uses the chain rule for mutual information, line (12) uses the chain rule again and the fact that from the definition (9), under q, the Markov chain $T_{\mathrm{IB}} - X - Y$ holds, while line (13) uses $I(X; Y|T_{\mathrm{IB}}, Y') = 0$ and $I_q(Y; Y'|T_{\mathrm{IB}}) = H(Y|T_{\mathrm{IB}})$, which are both consequences of Y' being a copy of Y. Therefore, combining (10), (14) and the fact that $H(Y)$ does not depend on κ, the problem (8) has the same solutions as

$$\underset{\substack{\kappa \in C_{\mathrm{IB}(X,Y)}\,: \\ I_q(Y;T_{\mathrm{IB}})=\lambda}}{\arg\min} \; I_q(X; T_{\mathrm{IB}}) - I_q(Y; T_{\mathrm{IB}}), \tag{15}$$

where q is defined from κ through (9). But in (15), as the value of $I_q(Y; T_{\mathrm{IB}})$ is fixed by the constraint, it can be removed from the target function. Moreover, the definition (9) shows that κ is entirely determined by $q(T_{\mathrm{IB}}|X) = \kappa_{\mathcal{X}}$. These two latter facts show that κ solves (15) (i.e., solves (8)) if and only if $q(T_{\mathrm{IB}}|X)$ solves

$$\underset{\substack{q(T_{\mathrm{IB}}|X) \in C(\mathcal{X}, \mathcal{T}_{\mathrm{IB}})\,: \\ I_q(Y;T_{\mathrm{IB}})=\lambda}}{\arg\min} \; I_q(X; T_{\mathrm{IB}}). \tag{16}$$

Eventually, it can be easily verified that the convex set $C := C(\mathcal{X}, \mathcal{T}_{\mathrm{IB}})$, together with the functions $f(q(T_{\mathrm{IB}}|X)) := I_q(Y; T_{\mathrm{IB}})$ and $g(q(T_{\mathrm{IB}}|X)) := I_q(X; T_{\mathrm{IB}})$, satisfy the assumptions of Lemma 5. Thus the equality $I_q(Y; T_{\mathrm{IB}}) = \lambda$ in (16) can be replaced by the inequality $I_q(Y; T_{\mathrm{IB}}) \geq \lambda$: in other words, the problem (16) can be replaced by the IB problem (7). This ends the proof of the proposition. ■

Let us point out that while the IIB problem is symmetric in X and Y, this is not the case for the IB problem, where the source variable and the relevancy variable play different roles. Here, we proved that the IB with source X and relevancy Y can be recovered by adding to (1) the constraint defined by $C_{\mathrm{IB}}(X, Y)$, but similarly, the IB with source Y and relevancy X can be recovered by replacing, in (8), the set $C_{\mathrm{IB}}(X, Y)$ with the set

$$C_{\mathrm{IB}(Y,X)} := \{e_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}} \; : \;\; \kappa_{\mathcal{Y}} \in C(\mathcal{Y}, \mathcal{T}_{\mathrm{IB}})\} \;\subset\; C(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{T}_{\mathrm{IB}}).$$

of channels that compress the $\mathcal{Y}$ coordinate but leave the $\mathcal{X}$ coordinate unchanged.

A.2. IIB and Symmetric IB

Let us consider a different restriction on κ which will lead to the Symmetric IB (Slonim et al., 2006). With $\mathcal{T}_\mathcal{X} := \mathbb{N}$ and $\mathcal{T}_\mathcal{Y} := \mathbb{N}$, we define the set

$$C_{sIB(X,Y)} := \{\kappa_\mathcal{X} \otimes \kappa_\mathcal{Y} : \quad \kappa_\mathcal{X} \in C(\mathcal{X}, \mathcal{T}_\mathcal{X}),\, \kappa_\mathcal{Y} \in C(\mathcal{Y}, \mathcal{T}_\mathcal{Y})\} \;\subset\; C(\mathcal{X}\times\mathcal{Y}, \mathcal{T}_\mathcal{X}\times\mathcal{T}_\mathcal{Y})$$

of split channels, i.e., of channels that transform $\mathcal{X}$ and $\mathcal{Y}$ separately.[6] Note that for such channels, the output T can be written $T = (T_X, T_Y)$, where T_X is defined on $\mathcal{T}_\mathcal{X}$ and T_Y on $\mathcal{T}_\mathcal{Y}$. We consider the problem

$$\underset{\substack{\kappa\in C_{\mathrm{sIB}(X,Y)}\,:\\ D(\kappa(p(X,Y))||\kappa(p(X)p(Y)))=\lambda}}{\arg\min} \qquad I_\kappa(X,Y;T). \tag{17}$$

We want to show that this problem has the same set of solutions as

$$\underset{\substack{q(T_X|X),\,q(T_Y|Y)\,:\\ I_q(T_X;T_Y)\geq\lambda}}{\arg\min} \quad I_q(X;T_X) + I_q(Y;T_Y). \tag{18}$$

Proposition 7 *Let $0 \leq \lambda \leq I(X;Y)$. Then:*

(i) In (18), the inequality in the constraint can be replaced by the equality constraint $I_q(T_X;T_Y) = \lambda$.

(ii) The set of solutions of the problems (17) and (18) are identical.

Proof It can be easily verified that the convex set $C := C(\mathcal{X}, \mathcal{T}_\mathcal{X}) \times C(\mathcal{Y}, \mathcal{T}_\mathcal{Y})$, together with the functions

$$f \,: (q(T_X|X), q(T_Y|Y)) \mapsto I_q(T_X;T_Y)$$

and

$$g \,: (q(T_X|X), q(T_Y|Y)) \mapsto I_q(X;T_X) + I(Y;T_Y),$$

satisfy the assumptions of Lemma 5. Thus, the latter proves point (i).

Let us now prove (ii). For $\kappa = \kappa_\mathcal{X} \otimes \kappa_\mathcal{Y}$, we define the joint distribution $q(X,Y,T_X,T_Y)$ on $\mathcal{X}\times\mathcal{Y}\times\mathcal{T}_\mathcal{X}\times\mathcal{T}_\mathcal{Y}$ through

$$q(x,y,t_X,t_Y) := q(x,y)\kappa_\mathcal{X}(t_X|x)\kappa_\mathcal{Y}(t_Y|y). \tag{19}$$

In particular, $q(X,Y,T_X,T_Y)$ is such that the Markov chain $T_X - X - Y - T_Y$ holds. From the latter Markov chain, the chain rule for mutual information and the equality

6. As we defined $\mathcal{T} := \mathbb{N}$ and as there is a bijection between $\mathbb{N}\times\mathbb{N}$ and $\mathbb{N}$, writing here the bottleneck space as $\mathcal{T}_\mathcal{X}\times\mathcal{T}_\mathcal{Y}$ rather than $\mathcal{T}$ is just a difference of presentation.

$I(A;B) = H(A) - H(A|B)$, we get

$$\begin{aligned}
I_q(X,Y;T_X,T_Y) &= I(X;T_X,T_Y) + I(Y;T_X,T_Y|X) \\
&= I(X;T_X) + I(X;T_Y|T_X) + I(Y;T_X|X) + I(Y;T_Y|X,T_X) \\
&= I(X;T_X) + I(X;T_Y|T_X) + 0 + I(Y;T_Y|X) \qquad (20)\\
&= I(X;T_X) + H(T_Y|T_X) - H(T_Y|T_X,X) + H(T_Y|X) - H(T_Y|X,Y) \\
&= I(X;T_X) + H(T_Y|T_X) - H(T_Y|X) + H(T_Y|X) - H(T_Y|Y) \qquad (21)\\
&= I(X;T_X) + H(T_Y|T_X) - H(T_Y|Y) \\
&= I(X;T_X) + I(Y;T_Y) - I(T_Y;T_X), \qquad (22)
\end{aligned}$$

where line (20) uses $T_X - X - Y$ and $T_X - X - (T_Y, Y)$, while line (21) uses $T_X - X - T_Y$ and $X - Y - T_Y$. Moreover, it can be verified that, for $\kappa = \kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}} = q(T_X|T) \otimes q(T_Y|Y)$, we have $\kappa(p(X,Y)) = q(T_x, T_Y)$ and $\kappa(p(X)p(Y)) = q(T_X)q(T_Y)$, so that

$$D(\kappa(p(X,Y))||\kappa(p(X)p(Y))) = D(q(T_X,T_Y)||q(T_X)q(T_Y)) = I_q(T_X;T_Y). \qquad (23)$$

Combining (22) and (23) above, we get that the solutions of (17) are also those of

$$\underset{\substack{q(T_X|X),\, q(T_Y|Y):\\ I_q(T_X;T_Y)=\lambda}}{\arg\min} \; I_q(X;T_X) + I_q(Y;T_Y) - I_q(T_X;T_Y), \qquad (24)$$

But in the latter problem, as the value of $I_q(T_X;T_Y)$ is fixed by the constraint, it can be removed from the target function. Eventually, we can use point (i) to conclude that the solutions of (24) coincide with those of the problem (18), which ends the proof. ■

Crucially, the problem (18) is the Symmetric IB — more precisely, Ref. (Slonim et al., 2006) defines the Lagrangian relaxation (Lemaréchal, 2001) of (18), i.e.,

$$\underset{q(T_X|X),\, q(T_Y|Y)}{\arg\min} \; I_q(X;T_X) + I_q(Y;T_Y) - \beta I_q(T_X;T_Y), \qquad (25)$$

for varying parameter $\beta \geq 0$. In this sense, the IIB problem (1) with additional constraint of split channel $\kappa = \kappa_{\mathcal{X}} \otimes \kappa_{\mathcal{Y}}$, i.e., the problem (17), is the Symmetric IB problem.

Appendix B. Proof of Theorem 3

In most of the proof (Sections B.1 and B.2), we will set ourselves in the more general framework of fully supported marginals $p(X)$ and $p(Y)$, but not necessarily fully supported joint distribution $p(X,Y)$. This more general formulation might help for future work to generalise this paper's results. However, at the end the proof (Section B.3) we will use the assumption of fully supported $p(X,Y)$.

Notations In this proof, we denote channels in $C(\mathcal{X} \times \mathcal{Y}, \mathcal{T})$ by $q(T|X,Y)$ rather than κ. For $(x,y,t) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{T}$, we write

$$q(x,y,t) := p(x,y)q(t|x,y), \quad \tilde{q}(x,y,t) := p(x)p(y)q(t|x,y),$$

and $q(T)$, resp. $\tilde{q}(T)$, the corresponding marginals on the bottleneck space $\mathcal{T}$. The symbols $\mathcal{S}$ and $\text{supp}(p(X,Y))$ both denote the support of the distribution $p(X,Y)$. For a subset $\mathcal{A}$, we denote by $\mathcal{A}^c$ the complement of $\mathcal{A}$. We consider the equivalence relation

$$(x,y) \sim (x',y') \quad \Leftrightarrow \quad \frac{p(x,y)}{p(x)p(y)} = \frac{p(x',y')}{p(x')p(y')}, \tag{26}$$

which is always well-defined, because we assumed that $p(X)$ and $p(Y)$ are fully supported. The relation $\sim$ defines a partition of $\mathcal{X}\times\mathcal{Y}$. If $\mathcal{S}^c \neq \emptyset$, then $\mathcal{S}^c$ is an element of this partition, and we write $\{\mathcal{S}_j\}_{j=1,\dots,n}$ for the other elements of the partition, which together thus define a partition of the support $\mathcal{S}$. The latter partition can be seen as the deterministic clustering

$$\begin{aligned} \pi_{\mathcal{S}} : \quad & \mathcal{S} \longrightarrow \{1,\dots,n\} \\ & (x,y) \mapsto \sum_{j=1}^{n} j\, \delta_{(x,y)\in\mathcal{S}_j}. \end{aligned} \tag{27}$$

We also denote by π the deterministic clustering defined by the relation $\sim$ on the whole space $\mathcal{X}\times\mathcal{Y}$: explicitly, we set $\pi_{|\mathcal{S}} := \pi_{\mathcal{S}}$, and if $\mathcal{S}^c \neq \emptyset$, we set $\pi(x,y) = 0$ for $(x,y) \in \mathcal{S}^c$.

As we will see, the clustering $\pi_{\mathcal{S}}$ happens to be the essentially unique solution to (1) for $\lambda = I(X;Y)$. To make this statement precise, we need the following notion (Ay et al., 2017):

Definition 8 *Fo finite sets $\mathcal{A}$ and $\mathcal{B}$, a channel γ from $\mathcal{A}$ to $\mathcal{B}$ is called* congruent *if for $a \neq a'$, the supports of $\gamma(B|a)$ and $\gamma(B|a')$ are disjoint. We will denote by $C_{cong}(\mathcal{A},\mathcal{B})$ the set of congruent channels from $\mathcal{A}$ to $\mathcal{B}$.*

The definition says that, observing an outcome $b \in \mathcal{B}$ with nonzero probability, one can reconstruct unambiguously the $a \in \mathcal{A}$ which was originally transmitted through the channel. Thus, intuitively, a congruent channel $p(B|A)$ defines a splitting of each symbol $a \in \mathcal{A}$ into the symbol(s) of $\text{supp}(p(B|a))$. Note that permutations of $\mathcal{A}$ are congruent channels with $\mathcal{A} = \mathcal{B}$ and $|\text{supp}(p(B|a))| = 1$ for all $a \in \mathcal{A}$.

B.1. Explicit form of IIB solutions for $\lambda = I(X;Y)$

Theorem 9 *Let $\lambda = I(X;Y)$. The solutions to the IIB problem (1) are the channels of the form*

$$q(t|x,y) = \begin{cases} (\gamma \circ \pi_{\mathcal{S}})(t|x,y) & \text{if } (x,y) \in \mathcal{S} \\ q_0(t|x,y) & \text{if } (x,y) \in \mathcal{S}^c \end{cases}$$

for any congruent channel $\gamma \in C_{cong}(\{1,\dots,n\},\mathcal{T})$, and any arbitrary channel $q_0 \in C(\mathcal{S}^c,\mathcal{T})$ on the support's complement.

In short, a solution $q(T|X,Y)$ to the IIB for $\lambda = I(X;Y)$ can have an arbitrary effect on the zero probability symbols, but its restriction to the support $\mathcal{S}$ must be, up to permuting or splitting the symbols in $\mathcal{T}$, the clustering $\pi_{\mathcal{S}}$ from (27). The following corollary is then straightforward:

Corollary 10 *Assume that $p(X,Y)$ is fully supported, and let $\lambda = I(X;Y)$. Then the solutions to the IIB problem (1) are the channels of the form*

$$q(t|x,y) = (\gamma \circ \pi)(t|x,y)$$

for any congruent channel $\gamma \in C_{cong}(\{1,\dots,n\}, \mathcal{T})$, where π is the deterministic clustering defined by the relation $\sim$ (see equation (26)).

Let us come back to the proof of Theorem 9.

Proof

The following sets, defined for $j = 1,\dots,n$, will be central to the proof:

$$\mathcal{T}_j^q = \{t \in \mathcal{T} : \ \exists (x,y) \in \mathcal{S}_j, \ q(t|x,y) > 0\}. \tag{28}$$

Intuitively, $\mathcal{T}_j^q$ is the "probabilistic image set" of $\mathcal{S}_j$ through $q(T|X,Y)$: i.e, it is the subset of $\mathcal{T}$ that can be achieved with nonzero probability starting from inputs (x,y) in $\mathcal{S}_j$ and using the channel $q(T|X,Y)$. It will also be useful to consider, for $t \in \mathcal{T}$,

$$\mathcal{S}_t^q := \{(x,y) \in \mathcal{S} : \ q(t|x,y) > 0\}, \tag{29}$$

which can be seen as the "probabilistic pre-image set" of t through $q(T|X,Y)$. The following lemma shows that $D(q(T)||\tilde{q}(T)) = I(X;Y)$ is characterised by the fact that $\frac{p(x,y)}{p(x)p(y)}$ is constant on the "pre-image" $\mathcal{S}_t^q$ of every symbol t:

Lemma 11 *Let $q(T|X,Y) \in C(\mathcal{X}\times\mathcal{Y}, \mathcal{T})$. Then we always have $D(q(T)||\tilde{q}(T)) \leq I(X;Y)$, and $D(q(T)||\tilde{q}(T)) = I(X;Y)$ if and only if, for all $t \in \mathcal{T}$, there exists some $\mathcal{S}_j$ such that*

$$\mathcal{S}_t^q \subseteq \mathcal{S}_j. \tag{30}$$

Proof We have

$$\begin{aligned} D(q(T)||\tilde{q}(T)) &= \sum_t \left(\sum_{x,y} q(t|x,y)p(x,y) \right) \log \left(\frac{\sum_{x,y} q(t|x,y)p(x,y)}{\sum_{x,y} q(t|x,y)p(x)p(y)} \right) \\ &= \sum_t \left(\sum_{(x,y)\in\mathcal{S}} q(t|x,y)p(x,y) \right) \log \left(\frac{\sum_{(x,y)\in\mathcal{S}} q(t|x,y)p(x,y)}{\sum_{(x,y)\in\mathcal{S}} q(t|x,y)p(x)p(y)} \right), \end{aligned}$$

while

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \\ &= \sum_{(x,y)\in\mathcal{S}} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \\ &= \sum_{(x,y)\in\mathcal{S}} \left(\sum_t q(t|x,y)p(x,y) \right) \log \left(\frac{q(t|x,y)p(x,y)}{q(t|x,y)p(x)p(y)} \right) \\ &= \sum_t \sum_{(x,y)\in\mathcal{S}} q(t|x,y)p(x,y) \log \left(\frac{q(t|x,y)p(x,y)}{q(t|x,y)p(x)p(y)} \right), \end{aligned}$$

where we use the convention $0 \log(\frac{0}{0}) = 0$. But from the log-sum inequality, for all $t \in \mathcal{T}$,

$$\left(\sum_{(x,y)\in\mathcal{S}} q(t|x,y)p(x,y) \right) \log \left(\frac{\sum_{(x,y)\in\mathcal{S}} q(t|x,y)p(x,y)}{\sum_{(x,y)\in\mathcal{S}} q(t|x,y)p(x)p(y)} \right) \\ \leq \sum_{(x,y)\in\mathcal{S}} q(t|x,y)p(x,y) \log \left(\frac{q(t|x,y)p(x,y)}{q(t|x,y)p(x)p(y)} \right). \tag{31}$$

So that $D(q(T)||\tilde{q}(T)) \leq I(X;Y)$, with equality if and only if, for all $t \in \mathcal{T}$, it holds in (31). From the equality case of the log-sum inequality (Csiszár and Körner, 2011), the latter is equivalent to the existence of nonzero constants $(\alpha_t)_{t\in\mathcal{T}}$ such that

$$\forall (x,y) \in \mathcal{S}, \qquad q(t|x,y)p(x,y) \;=\; \alpha_t \, q(t|x,y)p(x)p(y),$$

i.e., such that, for every t, the quantity $\frac{p(x,y)}{p(x)p(y)}$ is constant on the subset of elements (x,y) such that $q(t|x,y) > 0$. Recalling the definitions (29) of $\mathcal{S}_t^q$ and (26) of the relation $\sim$ defining the sets $\mathcal{S}_j$, we thus proved the following: we have $D(q(T)||\tilde{q}(T)) = I(X;Y)$ if and only if, for all $t \in \mathcal{T}$, there exists some $\mathcal{S}_j$ such that

$$\mathcal{S}_t^q \;\subseteq\; \mathcal{S}_j, \tag{32}$$

■

Lemma 12 *Let $q(T|X,Y) \in C(\mathcal{X}\times\mathcal{Y},\mathcal{T})$. Then the following are equivalent:*

(i) $D(q(T)||\tilde{q}(T)) = I(X;Y)$,

(ii) $\{\mathcal{T}_j^q\}_{j=1,\dots,n}$ *is a partition of* $\mathrm{supp}(q(T)) \subseteq \mathcal{T}$,

(iii) For all $j, j' \in \{1,\dots,n\}$, *we have* $q(\mathcal{T}_{j'}|\mathcal{S}_j) = \delta_{j'=j}$.

Proof Note that it clearly follows from the definition (28) that the union of the sets $\mathcal{T}_j^q$ is $\mathrm{supp}(q(T))$, so that these sets define a partition of $\mathrm{supp}(q(T))$ if and only if they are disjoint. Moreover, the definition (28) of $\mathcal{T}_j^q$ can be reformulated as

$$\mathcal{T}_j^q = \{t \in \mathrm{supp}(q(T)) : \; \mathcal{S}_t^q \cap \mathcal{S}_j \neq \emptyset\}. \tag{33}$$

which means, intuitively, that a symbol t is in the (probabilistic) image $\mathcal{T}_j^q$ of $\mathcal{S}_j$ through $q(T|X,Y)$ if and only if the (probabilistic) pre-image S_t^q of t intersects the set $\mathcal{S}_j$.

Assume that $D(q(T)||\tilde{q}(T)) = I(X;Y)$ holds. Then Lemma 11 and the fact that the $\mathcal{S}_j$ are disjoint imply that $\mathcal{S}_t^q \cap \mathcal{S}_j \neq \emptyset \Leftrightarrow \mathcal{S}_t^q \subseteq \mathcal{S}_j$, so that

$$\mathcal{T}_j^q = \{t \in \mathrm{supp}(q(T)) : \; \mathcal{S}_t^q \subseteq \mathcal{S}_j\}. \tag{34}$$

Therefore, once again because the $\mathcal{S}_j$ are disjoint, the sets $\mathcal{T}_j^q$ must also be disjoint, and they define a partition of $\mathrm{supp}(q(T))$.

Conversely, assume that $\{\mathcal{T}_j^q\}_{j=1,\dots,n}$ is a partition of $\mathrm{supp}(q(T))$. If there is some $t \in \mathrm{supp}(q(T))$ such that we have both $\mathcal{S}_t^q \cap \mathcal{S}_j \neq \emptyset$ and $\mathcal{S}_t^q \cap \mathcal{S}_{j'} \neq \emptyset$, for $j \neq j'$, then from (33), we have $t \in \mathcal{T}_j^q \cap \mathcal{T}_{j'}^q$, which is a contradiction. Thus for all $t \in \mathrm{supp}(q(T))$, there is a unique $j \in \{1, \cdots, n\}$ such that $\mathcal{S}_t^q \cap \mathcal{S}_j \neq \emptyset$. As the union of the $\mathcal{S}_j$ over $j \in \{1, \dots, n\}$ is $\mathcal{S}$, and as by definition, $\mathcal{S}_t^q$ is included in $\mathcal{S}$, this means that $\mathcal{S}_t^q \subseteq \mathcal{S}_j$. Therefore, from Lemma 11, we must have $D(q(T)||\tilde{q}(T)) = I(X;Y)$, and the equivalence of points (i) and (ii) is proven.

Let us now prove the equivalence of points (ii) and (iii). For $1 \leq j, j' \leq n$,

$$q(\mathcal{T}_{j'}^q|\mathcal{S}_j) = \frac{q(\mathcal{T}_{j'}^q, \mathcal{S}_j)}{p(\mathcal{S}_j)} = \frac{1}{p(\mathcal{S}_j)} \sum_{t \in \mathcal{T}_{j'}^q, (x,y) \in \mathcal{S}_j} q(t|x,y)p(x,y),$$

where we recall that for all $(x,y) \in \mathcal{S}_j$, we have $p(x,y) > 0$, because $\mathcal{S}_j \subseteq \mathcal{S}$. Thus

$$q(\mathcal{T}_{j'}^q|\mathcal{S}_j) = 0 \quad \Leftrightarrow \quad \forall t \in \mathcal{T}_{j'}^q, \forall (x,y) \in \mathcal{S}_j,\ q(t|x,y) = 0.$$

But $q(t|x,y) = 0$ for all $(x,y) \in \mathcal{S}_j$ if and only if $t \notin \mathcal{T}_j^q$ (by definition of $\mathcal{T}_j^q$). Therefore

$$\forall 1 \leq j, j' \leq n : j' \neq j,\ \ q(\mathcal{T}_{j'}^q|\mathcal{S}_j) = 0 \quad \Leftrightarrow \quad \forall 1 \leq j, j' \leq n : j' \neq j,\ \forall t \in \mathcal{T}_{j'}^q,\ \ t \notin \mathcal{T}_j^q.$$

The right-hand-side above clearly means that the sets $\mathcal{T}_q^j$, for varying $1 \leq j \leq n$, are pairwise disjoint. As we saw in the begining of the proof, this is equivalent to these sets forming a partition of $\mathrm{supp}(\mathcal{T})$. Thus point (ii) is equivalent to

$$\forall 1 \leq j, j' \leq n : j \neq j',\ \ q(\mathcal{T}_{j'}^q|\mathcal{S}_j) = 0.$$

To end the proof, let us show that $q(\mathcal{T}_j^q|\mathcal{S}_j) = 1$ always holds. We have

$$\begin{aligned} q(\mathcal{T}_j^q, \mathcal{S}_j) &= \sum_{(x,y) \in \mathcal{S}_j} \sum_{t \in \mathcal{T}_j^q} p(x,y)q(t|x,y) \\ &= \sum_{(x,y) \in \mathcal{S}_j} \sum_{t \in \mathcal{T}} p(x,y)q(t|x,y) \\ &= \sum_{(x,y) \in \mathcal{S}_j} p(x,y) = p(\mathcal{S}_j), \end{aligned}$$

where the second equality uses the definition (28) of $\mathcal{T}_j^q$ as "probabilistic image" of $\mathcal{S}_j$ through $q(T|X,Y)$. Therefore, $q(\mathcal{T}_j^q|\mathcal{S}_j) = 1$. ■

Lemma 13 *Let $q(T|X,Y) \in C(\mathcal{X} \times \mathcal{Y}, \mathcal{T})$. Then $q(T|X,Y)$ solves the IIB probelm (1) with $\lambda = I(X;Y)$ if and only if*

1. *$\{\mathcal{T}_j^q\}_{j=1,\dots,n}$ is a partition of $\mathrm{supp}(q(T)) \subseteq \mathcal{T}$*
2. *For all $j \in \{1, \dots, n\}$, all $(x,y) \in \mathcal{S}_j$ and all $t \in \mathcal{T}_j^q$,*

$$q(t|x,y) = \sum_{j'=1}^{n} q(t|\mathcal{T}_{j'}^q)q(\mathcal{T}_{j'}^q|x,y) \tag{35}$$

Proof We have

$$\begin{aligned} I_q(X,Y;T) &= \sum_{x,y,t} p(x,y)q(t|x,y)\log\left(\frac{q(t|x,y)}{q(t)}\right) \\ &= \sum_{(x,y)\in\mathcal{S}} p(x,y) \sum_{t\in\text{supp}(q(T))} q(t|x,y)\log\left(\frac{q(t|x,y)}{q(t)}\right) \\ &= \sum_{j=1}^{n} \sum_{(x,y)\in\mathcal{S}_j} p(x,y) \sum_{t\in\text{supp}(q(T))} q(t|x,y)\log\left(\frac{q(t|x,y)}{q(t)}\right) \\ &= \sum_{j=1}^{n} \sum_{(x,y)\in\mathcal{S}_j} p(x,y) \sum_{t\in\mathcal{T}_j^q} q(t|x,y)\log\left(\frac{q(t|x,y)}{q(t)}\right), \end{aligned}$$

where the second equality uses the fact that if $(x,y)\in\mathcal{S}$, then $q(t)=0$ implies that $q(t|x,y)=0$; and the last equality follows from the definition of T_j^q as the "probabilistic image set" of $\mathcal{S}_j$ (see equation (28)). Yet, using once again the log-sum inequality, we have, for all $j=1,\dots,n$ and all $(x,y)\in\mathcal{S}_j$,

$$\sum_{t\in\mathcal{T}_j^q} q(t|x,y)\log\left(\frac{q(t|x,y)}{q(t)}\right) \geq \Big(\sum_{t\in\mathcal{T}_j^q} q(t|x,y)\Big)\log\left(\frac{\sum_{t\in\mathcal{T}_j^q} q(t|x,y)}{\sum_{t\in\mathcal{T}_j^q} q(t)}\right),$$

i.e,

$$\sum_{t\in\mathcal{T}_j^q} q(t|x,y)\log\left(\frac{q(t|x,y)}{q(t)}\right) \geq q(\mathcal{T}_j^q|x,y)\log\left(\frac{q(\mathcal{T}_j^q|x,y)}{q(\mathcal{T}_j^q)}\right), \tag{36}$$

with equality if and only if (Csiszár and Körner, 2011) for all $t\in\mathcal{T}_j^q$,

$$\frac{q(t|x,y)}{q(t)} = \frac{q(\mathcal{T}_j^q|x,y)}{q(\mathcal{T}_j^q)},$$

i.e.,

$$q(t|x,y) = q(t|\mathcal{T}_j^q)q(\mathcal{T}_j^q|x,y).$$

Thus $q(T|X,Y)$ minimises $I_q(X,Y;T)$ among all the channels $q'(T|X,Y)$ such that $D(q'(T)||\tilde{q}'(T)) = I(X;Y)$ if and only if

1. It satisfies $D(q(T)||\tilde{q}(T)) = I(X;Y)$.
2. For all $j\in\{1,\dots,n\}$, all $(x,y)\in\mathcal{S}_j$ and all $t\in\mathcal{T}_j^q$,

$$q(t|x,y) = q(t|\mathcal{T}_j^q)q(\mathcal{T}_j^q|x,y). \tag{37}$$

From point (ii) in Lemma 12, point 1 above is equivalent to point 1 in Lemma 13. Again from point (ii) in Lemma 12, we have $q(t|\mathcal{T}_{j'}) = 0$ for $t \in \mathcal{T}_j^q$ and $j \neq j'$, so that the equality (37) can be reformulated as

$$q(t|x,y) = \sum_{j'=1}^{n} q(t|\mathcal{T}_{j'}^q) q(\mathcal{T}_{j'}^q|x,y), \tag{38}$$

where in the sum above, only one term is actually nonzero. In other words, point 2 above is equivalent to point 2 in Lemma 13. ■

Let us now keep reformulating points 1 and 2 in Lemma 13. For a channel $q(T|X,Y)$ that solves the IIB problem (1) with $\lambda = I(X;Y)$, we define $c \in C(\mathcal{S}, \{1,\dots,n\})$ and $\gamma \in C(\{1,\dots,n\})$ through

$$c(j|x,y) := q(\mathcal{T}_j^q|x,y), \quad \gamma(t|j) := q(t|\mathcal{T}_j^q). \tag{39}$$

Then (35) can be reformulated as

$$\forall (x,y) \in \mathcal{S},\, t \in \mathcal{T}, \quad q(t|x,y) = (\gamma \circ c)(t|x,y).$$

Yet, using point (iii) in Lemma 12,

$$\begin{aligned} q(\mathcal{T}_j^q|x,y) &= \sum_{j'} q(\mathcal{T}_j^q|\mathcal{S}_{j'}) q(\mathcal{S}_{j'}|x,y) = \sum_{j'} \delta_{j=j'} q(\mathcal{S}_{j'}|x,y) = q(\mathcal{S}_j|x,y) \\ &= p(\mathcal{S}_j|x,y) = \delta_{(x,y)\in\mathcal{S}_j}, \end{aligned}$$

so that c must be the deterministic clustering $\pi_{\mathcal{S}}$ of the partition defined by $\sim$ (see equation (27)).

On the other hand, requiring that $\{\mathcal{T}_j^q\}_{j=1,\dots,n}$ defines a partition of $\mathrm{supp}(q(T))$ is equivalent to requiring that the supports of the distributions $q(T|\mathcal{T}_j^q)$ are disjoint. Recalling our definition (39) of γ, this means requiring that γ is a congruent channel (see Definition 8).

Eventually, note that Lemma 13 imposes no constraint on $q(t|x,y)$ if $(x,y) \in \mathcal{S}^c = \mathrm{supp}(p(X,Y))^c$. We have thus proved that the solution set of the IIB problem (1) for $\lambda = I(X;Y)$ coincides with the channels of the form

$$q(t|x,y) = \begin{cases} (\gamma \circ \pi_{\mathcal{S}})(t|x,y) & \text{if } (x,y) \in \mathcal{S} \\ q_0(t|x,y) & \text{if } (x,y) \in \mathcal{S}^c \end{cases}$$

for any arbitrary channel $q_0 \in C(\mathcal{S}^c, \mathcal{T})$ on the support's complement, and any congruent channel $\gamma \in C_{\mathrm{cong}}(\{1,\dots,n\}, \mathcal{T})$.

■

B.2. Characterisation of equivariances with the equivalence relation

In this part, we characterise the equivariance group of $p(Y|X)$ with the equivalence relation $\sim$ (see equation (26)). We recall that the equivariance group of the channel $p(Y|X)$ is denoted by $G_{p(Y|X)}$. Similarly we denote by $G_{p(X|Y)}$ the equivariance group of the channel $p(X|Y)$. Besides the projection $G^{\mathcal{X}}_{p(Y|X)}$ of $G_{p(Y|X)}$ on $\mathrm{Bij}(\mathcal{X})$ (see equation (2)), we also consider the projection of $G_{p(Y|X)}$ on $\mathrm{Bij}(\mathcal{Y})$, i.e.,

$$G^{\mathcal{Y}}_{p(Y|X)} := \{\tau \in \mathrm{Bij}(\mathcal{Y}) \,:\, \exists\, \sigma \in \mathrm{Bij}(\mathcal{X}),\, (\sigma, \tau) \in G_{p(Y|X)}\},$$

and similarly, the projections of $G_{p(X|Y)}$ on resp. $\mathrm{Bij}(\mathcal{X})$ and $\mathrm{Bij}(\mathcal{Y})$, i.e.,

$$G^{\mathcal{X}}_{p(X|Y)} := \{\sigma \in \mathrm{Bij}(\mathcal{X}) \,:\, \exists\, \tau \in \mathrm{Bij}(\mathcal{Y}),\, (\tau, \sigma) \in G_{p(X|Y)}\}$$

and

$$G^{\mathcal{Y}}_{p(X|Y)} := \{\tau \in \mathrm{Bij}(\mathcal{Y}) \,:\, \exists\, \sigma \in \mathrm{Bij}(\mathcal{X}),\, (\tau, \sigma) \in G_{p(X|Y)}\},$$

Moreover, the (column) permutation matrix defined by a bijection γ on some ordered finite set $\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\}$ will be denoted P_γ. For a probability $p(A)$ on $\mathcal{A}$, we write D_A the diagonal matrix with $\big(p(a_1), \dots, p(a_{|\mathcal{A}|})\big)$ on its diagonal. The (column) transition matrices defined by $p(Y|X)$ and $p(X|Y)$ will be denoted by resp. $P_{Y|X}$ and $P_{X|Y}$. The transpose of a matrix M will be denoted by $M^\dagger$. Eventually, the following notation will be central to the proof: for any $\sigma \in \mathrm{Bij}(\mathcal{X})$ and $\tau \in \mathrm{Bij}(\mathcal{Y})$,

$$\bar{P}_\sigma := D_X P_\sigma D_X^{-1}, \qquad \bar{P}_\tau := D_Y P_\tau D_Y^{-1}. \tag{40}$$

Note that $(\sigma, \tau) \in G_{p(Y|X)}$ means, in matrix terms, that $P_{Y|X} P_\sigma = P_\tau P_{Y|X}$ (see Definition 1). Let us start with the following lemmas:

Lemma 14 *For a pair of permutations $(\sigma, \tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$, we have*

$$P_{Y|X} P_\sigma = P_\tau P_{Y|X} \quad \Leftrightarrow \quad \bar{P}_{\sigma^{-1}} P_{X|Y} = P_{X|Y} \bar{P}_{\tau^{-1}}. \tag{41}$$

Proof The Bayes rule $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$, for all $x \in \mathcal{X}, y \in \mathcal{Y}$ can be written as

$$P_{Y|X} = D_Y P_{X|Y}^\dagger D_X^{-1}.$$

Thus

$$\begin{aligned} P_{Y|X} P_\sigma = P_\tau P_{Y|X} \quad &\Leftrightarrow \quad D_Y P_{X|Y}^\dagger D_X^{-1} P_\sigma = P_\tau D_Y P_{X|Y}^\dagger D_X^{-1} \\ &\Leftrightarrow \quad D_X P_\sigma^\dagger D_X^{-1} P_{X|Y} = P_{X|Y} D_Y P_\tau^\dagger D_Y^{-1}. \end{aligned} \tag{42}$$

But because P_σ and P_τ are permutation matrices, we have that $P_\sigma^\dagger = P_\sigma^{-1} = P_{\sigma^{-1}}$ and $P_\tau^\dagger = P_\tau^{-1} = P_{\tau^{-1}}$. Therefore $D_X P_\sigma^\dagger D_X^{-1} = \bar{P}_{\sigma^{-1}}$ and $D_Y P_\tau^\dagger D_Y^{-1} = \bar{P}_{\tau^{-1}}$, so that equation (42) yields the result. ∎

Lemma 15 *For a channel $\gamma \in C(\mathcal{A}, \mathcal{B})$, a pair $(\sigma, \tau) \in \mathrm{Bij}(\mathcal{A}) \times \mathrm{Bij}(\mathcal{B})$ is an equivariance of the channel γ if and only if for all $(a, b) \in \mathcal{A} \times \mathcal{B}$,*

$$\gamma(b|a) = \gamma(\tau \cdot b|\sigma \cdot a).$$

Proof We have, writing $P_{B|A}$ the column transition matrix corresponding to the channel $\gamma = \gamma(B|A)$ and G_γ the equivariance group of γ ,

$$\begin{aligned}(\sigma, \tau) \in G_\gamma \quad &\Leftrightarrow \quad P_{B|A} P_\sigma = P_\tau P_{B|A} \\ &\Leftrightarrow \quad P_{B|A} = P_\tau P_{B|A} P_{\sigma^{-1}} \\ &\Leftrightarrow \quad P_{B|A} = P_{\tau \cdot B|\sigma \cdot A},\end{aligned}$$

where the last equivalence comes from the fact that the *left* multiplication of $P_{B|A}$ by the column permutation matrix P_τ induces the permutation τ of the rows of $P_{B|A}$; whereas the *right* multiplication of $P_{B|A}$ by the column permutation matrix $P_{\sigma^{-1}}$ induces the permutation $(\sigma^{-1})^{-1} = \sigma$ of the columns of $P_{B|A}$. ■

Let us now state the main result of this section.

Theorem 16 *Let $p(Y|X)$ be fixed. If $p(X)$ is $G^{\mathcal{X}}_{p(Y|X)}$-exchangeable, then:*

(i) *$p(Y)$ is $G^{\mathcal{Y}}_{p(Y|X)}$-exchangeable.*

(ii) *For any pair $(\sigma, \tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$,*

$$(\sigma, \tau) \in G_{p(Y|X)} \;\Leftrightarrow\; (\tau, \sigma) \in G_{p(X|Y)}.$$

(iii) *$p(X)$ is $G^{\mathcal{X}}_{p(X|Y)}$-exchangeable.*

(iv) *For any pair $(\sigma, \tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$,*

$$(\sigma, \tau) \in G_{p(Y|X)} \;\Leftrightarrow\; \forall (x, y) \in \mathcal{X} \times \mathcal{Y},\ (x, y) \sim (\sigma \cdot x, \tau \cdot y).$$

(v) *$p(Y)$ is $G^{\mathcal{Y}}_{p(X|Y)}$-exchangeable.*

Intuitively, point *(ii)* means that the equivariance groups of $p(Y|X)$ and $p(X|Y)$ are the same (up to, of course, exchanging the order of the transformations on $\mathcal{X}$ and $\mathcal{Y}$, because we exchanged the role of X and Y). Note that the only result which we will use for the proof of Theorem 3 is point *(iv)*.

Proof *(i)*. For an equivariance $(\sigma, \tau) \in G_{p(Y|X)}$, we know that $p(\sigma \cdot X) = p(X)$, because $p(X)$ is $G^{\mathcal{X}}_{p(Y|X)}$-exchangeable. So that for all $y \in \mathcal{Y}$,

$$\begin{aligned}p(\tau \cdot y) &= \sum_x p(x) p(\tau \cdot y|x) = \sum_x p(x) p(y|\sigma^{-1} \cdot x) = \sum_x p(\sigma \cdot x) p(y|x) \\ &= \sum_x p(x) p(y|x) = p(y),\end{aligned}$$

where the second equality uses Lemma 15. Thus $p(Y)$ is $G^{\mathcal{Y}}_{p(Y|X)}$-exchangeable, which is point (i).

(ii). Let us consider an ordering on $\mathcal{X} = \{x_1, \dots, x_{|\mathcal{X}|}\}$, such that the r orbits of $G^{\mathcal{X}}_{p(Y|X)}$ are $\mathcal{O}_{\mathcal{X},1} := \{x_1, \dots, x_{m_1}\}$, $\mathcal{O}_{\mathcal{X},2} := \{x_{m_1+1}, \dots, x_{m_2}\}$, $\dots$, up to $\mathcal{O}_{\mathcal{X},r} := \{x_{m_{r-1}+1}, \dots, x_{m_r}\}$, where $m_r := |\mathcal{X}|$. Similarly, we consider an ordering of $\mathcal{Y} = \{y_1, \dots, y_{|\mathcal{Y}|}\}$ such that the s orbits of $G^{\mathcal{Y}}_{p(Y|X)}$ are $\mathcal{O}_{\mathcal{Y},1} := \{y_1, \dots, y_{n_1}\}$, $\mathcal{O}_{\mathcal{Y},2} := \{y_{n_1+1}, \dots, y_{n_2}\}$, $\dots$, up to $\mathcal{O}_{\mathcal{Y},s} := \{y_{n_{s-1}+1}, \dots, y_{n_s}\}$, where $n_s := |\mathcal{Y}|$. With theses orderings of $\mathcal{X}$ and $\mathcal{Y}$, for each equivariance $(\sigma, \tau) \in G_{p(Y|X)}$, the permutation matrices P_σ and P_τ have a diagonal block structure corresponding to the orbits of $G^{\mathcal{X}}_{p(Y|X)}$ on $\mathcal{X}$, resp. of $G^{\mathcal{Y}}_{p(Y|X)}$ on $\mathcal{Y}$. Let us write $P_{\sigma,i}$ for the bloc of P_σ corresponding to the orbit $\mathcal{O}_{\mathcal{X},i}$, and resp. write $P_{\tau,j}$ for the bloc of P_τ corresponding to the orbit $\mathcal{O}_{\mathcal{Y},j}$. Then the corresponding matrices $\bar{P}_\sigma$ and $\bar{P}_\tau$ (see definition (40)) must also have a diagonal block structure, with blocs $\bar{P}_{\sigma,i}$, resp. $\bar{P}_{\tau,j}$, given by

$$\bar{P}_{\sigma,i} = D_{X,i} P_{\sigma,i} D_{X,i}^{-1} \tag{43}$$

and

$$\bar{P}_{\tau,j} = D_{Y,j} P_{\tau,j} D_{Y,j}^{-1}, \tag{44}$$

where the diagonal matrices $D_{X,i}$ and $D_{Y,j}$ have, along their diagonal, the values of $p(X)$ corresponding to the orbit $\mathcal{O}_{\mathcal{X},i}$, resp. the values of $p(Y)$ corresponding to the orbit $\mathcal{O}_{\mathcal{Y},j}$. Explicitly, the diagonal values of $D_{X,i}$ are $\big(p(x_{m_{i-1}+1}), \dots, p(x_{m_i})\big)$, and those of $D_{Y,j}$ are $\big(p(y_{n_{j-1}+1}), \dots, p(y_{n_j})\big)$.

Now, the assumption that $p(X)$ is $G^{\mathcal{X}}_{p(Y|X)}$-exchangeable means that for every $1 \leq i \leq r$, the probability $p(X)$ is constant on the orbit $\mathcal{O}_{\mathcal{X},i}$: in other words, the diagonal vector defining $D_{X,i}$ from equation (43) is constant. Thus, from equation (43), for all $\sigma \in G^{\mathcal{X}}_{p(Y|X)}$, we have $\bar{P}_{\sigma,i} = P_{\sigma,i}$ for all i, which implies that $\bar{P}_\sigma = P_\sigma$. Moreover, as we already proved point (i), we know that $p(Y)$ is $G^{\mathcal{Y}}_{p(Y|X)}$-exchangeable. Thus similarly, the diagonal vector defining $D_{Y,j}$ from equation (44) is constant, and for all $\tau \in G^{\mathcal{Y}}_{p(Y|X)}$, we have $\bar{P}_\tau = P_\tau$. Therefore, using Lemma 14, we get, for all pair (σ, τ),

$$P_{Y|X} P_\sigma = P_\tau P_{Y|X} \quad \Leftrightarrow \quad P_{\sigma^{-1}} P_{X|Y} = P_{X|Y} P_{\tau^{-1}},$$

i.e.,

$$(\sigma, \tau) \in G_{p(Y|X)} \quad \Leftrightarrow \quad (\tau^{-1}, \sigma^{-1}) \in G_{p(X|Y)}.$$

So that, recalling that an element belongs to a group if and only if its inverse does, we proved point (ii).

(iii). Let $\sigma \in G^{\mathcal{X}}_{p(X|Y)}$: i.e., there exists some $\tau \in \mathrm{Bij}(\mathcal{Y})$ such that $(\tau, \sigma) \in G_{p(X|Y)}$. From point (ii), we have $(\sigma, \tau) \in G_{p(Y|X)}$, which means in particular that $\sigma \in G^{\mathcal{X}}_{p(Y|X)}$. Thus from our assumption that $p(X)$ is $G^{\mathcal{X}}_{p(Y|X)}$-exchangeable, we get that $p(X) = p(\sigma \cdot X)$. So that $p(X)$ is $G^{\mathcal{X}}_{p(X|Y)}$-exchangeable.

(iv). We have

$$
\begin{aligned}
(\sigma,\tau)\in G_{p(Y|X)} \quad &\Leftrightarrow \quad (\tau,\sigma)\in G_{p(X|Y)} && (45)\\
&\Leftrightarrow \quad p(X|Y)\circ\tau=\sigma\circ p(X|Y)\\
&\Leftrightarrow \quad \forall(x,y)\in\mathcal{X}\times\mathcal{Y},\ p(x|y)=p(\sigma\cdot x|\tau\cdot y) && (46)\\
&\Leftrightarrow \quad \forall(x,y)\in\mathcal{X}\times\mathcal{Y},\ \frac{p(x,y)}{p(y)}=\frac{p(\sigma\cdot x,\tau\cdot y)}{p(\tau\cdot y)}\\
&\Leftrightarrow \quad \forall(x,y)\in\mathcal{X}\times\mathcal{Y},\ \frac{p(x,y)}{p(x)p(y)}=\frac{p(\sigma\cdot x,\tau\cdot y)}{p(\sigma\cdot x)p(\tau\cdot y)}, && (47)
\end{aligned}
$$

where line (45) is the already proven point (ii); line (46) uses Lemma 15; and line (47) uses the already proven point (iii), i.e., the fact that $p(X)$ is $G^{\mathcal{X}}_{p(X|Y)}$-exchangeable. Thus, recalling the definition of the relation $\sim$ (see equation (26)), point (iv) holds.

(v). Let $\tau\in G^{\mathcal{Y}}_{p(X|Y)}$: i.e., there exists some $\sigma\in\mathrm{Bij}(\mathcal{Y})$ such that $(\tau,\sigma)\in G_{p(X|Y)}$. From point (ii), we have $(\sigma,\tau)\in G_{p(Y|X)}$, which means in particular that $\tau\in G^{\mathcal{Y}}_{p(Y|X)}$. But from point (i), we know that $p(Y)$ is $G^{\mathcal{Y}}_{p(Y|X)}$-exchangeable. Thus $p(Y)=p(\tau\cdot Y)$. I.e., we proved that $p(Y)$ is $G^{\mathcal{Y}}_{p(X|Y)}$-exchangeable. ■

B.3. Conclusion of the proof

Here, we assume again that $p(X,Y)$ is fully supported, i.e., that $\mathcal{S}=\mathcal{X}\times\mathcal{Y}$.

Moreover, we recall that in Theorem 3, we assume that $p(X)$ is $G^{\mathcal{X}}_{p(Y|X)}$-exchangeable. Thus from point (iv) in Theorem 16, we have

$$
(\sigma,\tau)\in G_{p(Y|X)} \quad \Leftrightarrow \quad \forall(x,y)\in\mathcal{X}\times\mathcal{Y},\ (x,y)\sim(\sigma\cdot x,\tau\cdot y). \tag{48}
$$

But recalling that $(\sigma\otimes\tau)(x,y):=(\sigma\cdot x,\tau\cdot y)$ and that by definition of the deterministic clustering π (see the beginning of Appendix B), we have $(x,y)\sim(x',y')$ if and only if $\pi(x,y)=\pi(x',y')$, we get that the right-hand-side in (48) is equivalent to

$$
\pi\circ(\sigma\otimes\tau)=\pi. \tag{49}
$$

Now, from Corollary 10 and the fact that $p(X,Y)$ is fully supported, the solutions to the IIB for $\lambda=I(X;Y)$ are the channels of the form $\gamma\circ\pi$, for any congruent channel $\gamma\in C_{\mathrm{cong}}(\{1,\dots,n\},\mathcal{T})$. Thus, if we prove that, for any congruent channel γ, equation (49) is equivalent to

$$
\gamma\circ\pi\circ(\sigma\otimes\tau)=\gamma\circ\pi, \tag{50}
$$

this would prove that for any solution κ to the IIB for $\lambda=I(X;Y)$, we have $(\sigma,\tau)\in G_{p(Y|X)}$ if and only if $\kappa\circ(\sigma\otimes\tau)=\kappa$: this is exactly the statement of Theorem 3. Therefore, we only need to prove the following lemma:

Lemma 17 *Let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be finite sets. Consider two functions $f,g:\mathcal{A}\to\mathcal{B}$, and a congruent channel $\gamma\in C_{cong}(\mathcal{B},\mathcal{C})$. Then $f=g$ if and only if $\gamma\circ f=\gamma\circ g$.*

Proof Clearly, $f = g$ implies $\gamma \circ f = \gamma \circ g$. Conversely, assume that $\gamma \circ f = \gamma \circ g$. As γ is congruent, the supports of the $\gamma(C|b)$, where $b \in \mathcal{B}$, are disjoint sets $\mathcal{C}_b \subseteq \mathcal{C}$. Let us consider the deterministic clustering $h \in C(\bigsqcup_{b\in\mathcal{B}} \mathcal{C}_b, \mathcal{B})$ defined by $h(b|c) := \delta_{c\in\mathcal{C}_b}$. Then $h \circ \gamma$ is the identity of $\mathcal{B}$. But $\gamma \circ f = \gamma \circ g$ implies that

$$h \circ \gamma \circ f = h \circ \gamma \circ g, \tag{51}$$

which thus means exactly $f = h$. ■

Remark 18 *The only part of the proof where we used the full support assumption on $p(X, Y)$ was Appendix B.3, which is thus the only part which would need, in future work, to be adapted to non-necessarily full-support distributions $p(X, Y)$.*

Appendix C. Towards generalisations to non-finite variables

This work is set in the finite case, but it provides a basis for generalisations to more general settings. Indeed, the notions and tools used in this paper have straightforward generalisations to, for instance, the measure-theoretic setting — which include finite, countable and continuous spaces. In particular, one can directly generalise to Borel spaces (Rudin, 1987) probabilities and conditional probabilities (Billingsley, 1995), as well as the Kullback-Leibler divergence and mutual information (Gray, 2014). Thus it seems that the IIB problem (1) can be defined for Borel spaces. Moreover, the tools used in Appendix B.1 to describe explicitly the case $\lambda = I(X;Y)$ seem to adapt well to Borel spaces: namely, the log-sum inequality and its equality case; partitions induced by an equivalence relation; and the switching of the integration order for probability measures (Billingsley, 1995).

Eventually, one can consider the action of measurable groups on Borel spaces (Kallenberg, 2017), along with the corresponding partition defined by the group action's orbits (which is crucial to Appendix B.2). One could then consider measurable equivariances of conditional probabilities between Borel spaces, and similarly a notion of exchangeability analogous to Definition 2. These concepts would allow the statement of Theorem 3 to be given a meaning in this general setting. We leave to future work to fully adapt the proof of Theorem 3 to such a generalised statement.

Appendix D. Towards a more general assumption for soft equivariances

In Definition 4, we require $p(X)$ to be uniform, whereas Theorem 3 suggests to require $p(X)$ to be only $G^{\mathcal{X}}_{p(Y|X)}$-exchangeable. However, it is not clear that such a definition would be well-posed. Indeed, distinct $G^{\mathcal{X}}_{p(Y|X)}$-exchangeable distributions $p(X)$ might *a priori* result in distinct sets of solutions κ to the IIB problem (1): in this case, equation (4) would depend on the specific choice of $p(X)$, so that requiring $p(X)$ to only be $G^{\mathcal{X}}_{p(Y|X)}$-exchangeable would make Definition 4 ill-posed.

We conjecture that this is actually not the case: i.e., that for fixed $p(Y|X)$, any $G^{\mathcal{X}}_{p(Y|X)}$-exchangeable distribution $p(X)$ results in the same solutions to the IIB problem (1). Yet, at this stage, we restrict ourselves to uniform $p(X)$ to ensure Definition 4 to be well-posed,

and will consider the generalization to $G^{\mathcal{X}}_{p(Y|X)}$-exchangeable $p(X)$ once the conjecture can be proven.