Proceedings Track

An Informational Parsimony Perspective on Symmetry-Based Structure Extraction

Editors: List of editors' names

Abstract

Extraction of structure, in particular of group symmetries, is increasingly crucial to understanding and building intelligent models. In particular, some information-theoretic models of complexity-constrained learning have been argued to induce invariance extraction. Here, we formalise these arguments from a group-theoretic perspective, and extend them to the study of more general probabilistic symmetries through dedicated structure-preserving compressions. More precisely, we consider compressions that are optimal under the constraint of preserving the divergence from a given exponential family, yielding a novel generalisation of the Information Bottleneck framework. Through appropriate choices of exponential families, we fully characterise (in the discrete and full support case) channel invariance, channel equivariance and distribution invariance under permutation. Allowing imperfect divergence preservation then leads to principled definitions of "soft symmetries", where the "coarseness" corresponds to the degree of compression of the system. In simple synthetic experiments, we demonstrate that our method successively recovers, at increasingly compressed "resolutions", nested but increasingly perturbed equivariances, where new equivariances emerge at bifurcation points of the distortion parameter. Our framework opens a new path towards the extraction of generalised probabilistic symmetries.

Keywords: Probabilistic symmetries, Equivariance, Information Bottleneck, Geometric Complexity

1. Introduction

Group symmetries have become highly relevant to the study of intelligence, from neuronal circuit dynamics (Stewart, 2022) to perception (Pizlo and de Barros, 2021) and the structure of representations (Higgins et al., 2022), to equivariant neural networks (Gerken et al., 2023) and structure-discovering AI models (Liu and Tegmark, 2022). This relevance is often understood as a consequence of the pervasiveness of symmetries in the natural world: biological and artificial systems interacting with such a highly structured environement should *leverage* its symmetries, e.g., by internalising them into their own information-processing. But what leverage do symmetries provide exactly? Intuitively, the presence of symmetries in a system allows for a *simpler* description of it — thus potentially improving the efficiency of learning and generalisation about this system. In other words, a system's symmetries afford the possibility of *informationally parsimonious* descriptions of it.

More precisely, the *projection on orbits* of a symmetry group's action can be seen as an information-preserving compression, in that it is a clustering which preserves the information about anything invariant under the group action. This motivates the search of dedicated rate-distortion-inspired frameworks whose optimal compressions mimick the projections on orbits of specific symmetry groups, and thus hopefully characterise the symmetries themselves. We implement this program with the introduction of a new generalisation of the Information Bottleneck (IB) framework (Tishby et al., 2000), which we call the Divergence Information

Bottleneck (dIB). Here optimal encoder channels (partially or fully) preserve the divergence of the data distribution from a given exponential family, and one can potentially impose constraints on the shape of encoders. With adequate choices of exponential families and channel shape constraints, we obtain encoders which, in the full divergence preservation case, characterise various group-theoretic symmetries. Relaxing the full divergence preservation requirement then leads to a principled definition of *soft symmetries*, where the "coarseness" is measured by the degree of compression of the system. Moreover, this framework has the potential to generalise existing notions of group symmetries not only by softening them, but also by capturing qualitatively different structures — through dIB problems with well-chosen exponential family and constraints on encoders.

The classic IB method has previously been argued to extract invariances (Achille and Soatto, 2018). In Section 2, we formally prove that the IB method, for full information preservation, characterises group-theoretic channel invariances. This motivates the introduction (Section 3) of the more general dIB framework, which we specialise to characterise the equivariances of a channel $p(Y|X)$ and the permutation invariances of a distribution $p(A)$. We then present simple synthetic numerical experiments on exact and soft equivariances (Section 4). These experiments show how the corresponding dIB framework can extract approximate equivariances by allowing imperfect preservation of the divergence. We study channels satisfying a series of nested equivariances that have been perturbed to various degrees. Our framework recovers the perturbed equivariances, at successive bifurcation points of the trade-off parameter corresponding to increasingly compressed resolutions. Finally, we present the limitations of our approach (Section 5), and conclude in Section 6.

Assumptions and notations All alphabets are finite, except bottleneck alphabets $\mathcal{T} := \mathbb{N}$. The probability simplex defined by an alphabet $\mathcal{A}$ is denoted by $\Delta_{\mathcal{A}}$. The set of conditional probabilities, also called channels, from $\mathcal{A}$ to $\mathcal{B}$, resp. to $\mathcal{A}$ itself, is denoted by $C(\mathcal{A}, \mathcal{B})$, resp. $C(\mathcal{A})$. Functions are seen as deterministic channels. Channels are often seen as linear maps between vector spaces of measures, where the image of the distribution p through the channel γ is written $\gamma \cdot p$. By extension, $f \cdot a := f(a)$ for an element a and a function f. The symbol $\circ$ denotes channel composition. The set of bijections of $\mathcal{A}$ is $\mathrm{Bij}(\mathcal{A})$, the uniform distribution $\mathcal{U}(\mathcal{A})$, the identity map $e_{\mathcal{A}}$, and $\mathcal{S}^c := \mathcal{A} \setminus \mathcal{S}$ for $\mathcal{S} \subseteq \mathcal{A}$. For $p_1 \in \Delta_{\mathcal{A}}, p_2 \in \Delta_{\mathcal{B}}$, their *tensor product* is defined through $(p_1 \otimes p_2)(a, b) := p_1(a)p_2(b)$. Similarly $(\mu \otimes \eta)(b, b'|a, a') := \mu(b|a)\eta(b'|a')$ for $\mu \in C(\mathcal{A}, \mathcal{B})$, $\eta \in C(\mathcal{A}', \mathcal{B}')$.

2. Information Bottleneck and Group Invariances

Let $p(X, Y) \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ such that $p(X)$ is full-support. The IB problem with source X and relevancy Y is defined, for $0 \leq \lambda \leq \Lambda := I(X; Y)$, as (Gilad-Bachrach et al., 2003)

$$\mathrm{IB}(\lambda) := \underset{\substack{q(T|X) \in C(\mathcal{X}, \mathcal{T}) : \\ I_q(T;Y) \geq \lambda}}{\arg\min} \quad I_q(X; T), \tag{1}$$

where $q(X, Y, T) := p(X, Y)q(T|X)$. This problem implements a trade-off between compressing X and preserving the information that X carries about the Y.

On the other hand, the *channel invariance group* G_{ci} of $p(Y|X)$ is the group of bijections $\sigma \in \mathrm{Bij}(\mathcal{X})$ such that $p(Y|X) \circ \sigma = p(Y|X)$, with projection on orbits written $\pi_{\mathrm{ci}} : \mathcal{X} \to \mathcal{X}/G$. Crucially, here π_{ci} characterises G_{ci}: it can be easily verified that $\sigma \in G_{\mathrm{ci}} \Leftrightarrow \pi_{\mathrm{ci}} \circ \sigma = \pi_{\mathrm{ci}}$.

We now want to show that the solutions $\kappa \in \mathrm{IB}(\Lambda)$ essentially coincide with π_{ci}, thus yielding a characterisation of G_{ci} through such κ. We will need the equivalence relation

$$x \sim_{\mathcal{X}} x' \quad \Leftrightarrow \quad p(Y|x) = p(Y|x'), \tag{2}$$

with corresponding partition $\bar{\mathcal{X}}$ and projection $\pi_{\mathcal{X}} : \mathcal{X} \to \bar{\mathcal{X}}$; along with the following notion.

Definition 1 *The set of* congruent channels *(Ay et al., 2017) from $\mathcal{A}$ to $\mathcal{B}$, denoted by $C_{cong}(\mathcal{A}, \mathcal{B})$, is that of channels γ such that there exists a function $f : \mathcal{B} \to \mathcal{A}$ with $f \circ \gamma = e_{\mathcal{A}}$.*

In particular, composing an encoder κ with a congruent channel γ can be seen as a trivial operation, in that the output of κ can be unambigously recovered from that of $\gamma \circ \kappa$.

Theorem 2 *For $\Lambda := I(X; Y)$ and all $\sigma \in \mathrm{Bij}(\mathcal{X})$, the following holds:*

(i) $IB(\Lambda) = \{\gamma \circ \pi_{\mathcal{X}} \ : \ \ \gamma \in C_{cong}(\bar{X}, \mathcal{T})\}$.

(ii) Let $\kappa \in IB(\Lambda)$. Then $\sigma \in G_{ci}$ if and only if $\kappa \circ \sigma = \kappa$.

(iii) If $\sigma \in G_{ci}$, then $\kappa \circ \sigma = \kappa$ also holds for all $0 \leq \lambda \leq \Lambda$ and $\kappa \in IB(\lambda)$.

(iv) The projection $\pi_{\mathcal{X}}$ defined by $\sim_{\mathcal{X}}$ coincides with the projection on orbits π_{ci}.

Proof See Appendices A and B. ■

Crucially, point (ii) means that invariances are thoses bijections σ such that the effect of transforming $\mathcal{X}$ with σ is "quotiented out" by $\kappa \in \mathrm{IB}(\Lambda)$. Point (i) explains why: these κ implement precisely (up to trivial transformations) the quotient of $\mathcal{X}$ by the equivalence relation $\sim_{\mathcal{X}}$ that equates elements of $\mathcal{X}$ providing the same information about Y (see (2)). Point (iii) shows that the "quotienting out" of invariances by bottlenecks also occurs for all values of the trade-off parameter λ, even though it is only a full characterisation for $\lambda = I(X; Y)$. Point (iv), combined with point (i), means that the projection π_{ci}, defined purely in group-theoretic terms, is characterised as the solution to the zero-distortion case of a generalised rate-distortion problem, here the IB (Zaidi et al., 2020). Note that point (i) is redundant with existing results (Shamir et al., 2010);[1] and point (ii) is not surprising as previous work already linked the IB method to invariance extraction (Achille and Soatto, 2018). However, our group-theoretic formalisation provides guidance for generalisations of this phenomenon: i.e., for reformulating and softening probabilistic symmetries with the language of information theory. The following sections provides first steps in this direction.

3. Divergence Information Bottleneck and Group Symmetries

3.1. General framework

Fix a distribution $p = p(A) \in \Delta_{\mathcal{A}}$, an exponential family $\mathcal{E} \subseteq \Delta_{\mathcal{A}}$, and a convex subset of encoders $C \subseteq C(\mathcal{A}, \mathcal{T})$. We then define the *Divergence Information Bottleneck* (dIB) as

$$\mathrm{dIB}(\lambda) := \underset{\substack{\kappa \in C \\ D(\kappa \cdot p || \kappa \cdot \mathcal{E}) \geq \lambda}}{\arg\min} \quad I_\kappa(A; T), \tag{3}$$

1. Point (i) can be seen as the fact that $\mathrm{IB}(\Lambda)$ consists of minimal sufficient statistics of X w.r.t. Y, proven in (Shamir et al., 2010). But our new proof also yields point (iii) and mirrors that of Theorem 3 below.

where $0 \leq \lambda \leq \Lambda := D(p||\mathcal{E})$, and, denoting by $\operatorname{cl}\mathcal{E}$ the topological closure of $\mathcal{E}$ in $\Delta_{\mathcal{A}}$,

$$D(p||\mathcal{E}) := \inf_{r \in \operatorname{cl}\mathcal{E}} D\big(p(A)||r(A)\big) = D\big(p(A)||\tilde{p}(A)\big), \tag{4}$$

$$D(\kappa \cdot p||\kappa \cdot \mathcal{E}) := \inf_{r \in \operatorname{cl}\mathcal{E}} D\big((\kappa \cdot p)(T)||(\kappa \cdot r)(T)\big) = D\big((\kappa \cdot p)(T)||(\kappa \cdot \tilde{p})(T)\big). \tag{5}$$

Here $\tilde{p} \in \operatorname{cl}\mathcal{E}$ is the unique distribution which achieves the minimum in both (4) and (5) (see Appendix C.1 for details). While $D(p||\mathcal{E})$ is the divergence of p from $\mathcal{E}$, here $D(\kappa \cdot p||\kappa \cdot \mathcal{E})$ measures the "divergence of p from $\mathcal{E}$ *in the latent space* $\mathcal{T}$, through the lens of the channel κ". Solutions to (3) can thus be seen as optimal compressions of A under the constraint of (partially or wholly) preserving the divergence of $p(A)$ from the exponential family $\mathcal{E}$. The choice of C allows to potentially enforce constraints on the shape of encoder channels κ.

Intuitively, $D(p||\mathcal{E})$ measures the presence of a specific structure in $p(A)$, formalised as the divergence from the family $\mathcal{E}$ of distributions which do not have such structure. E.g., for $\mathcal{A} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{E} := \Delta_{\mathcal{X}} \otimes \Delta_{\mathcal{Y}}$, we have $D(p(X,Y)||\mathcal{E}) = I(X;Y)$: the corresponding dIB (with e.g. $C = C(\mathcal{X} \times \mathcal{Y}, \mathcal{T})$) is a mutual information-preserving joint compression of X and Y (Charvin et al., 2023). More generally, the divergence from a *hierarchical model* $\mathcal{E}$ measures the complexity of a system's given set of interdependencies (Ay et al., 2011). This structure of dependencies should be made *salient* by an optimal compression which preserves only the corresponding complexity measure. Our dIB framework is primarily tailored for $\mathcal{E}$ being a hierarchical model, even though this assumption is not relevant to the next theorem.

Let us define, on $\mathcal{A}$, the equivalence relation

$$a \sim a' \quad \Leftrightarrow \quad p(a)\tilde{p}(a') = p(a')\tilde{p}(a), \tag{6}$$

with $\bar{\mathcal{A}} := \{\mathcal{A}_j\}_{j=1,\dots,n}$, and $\pi : \mathcal{A} \to \bar{\mathcal{A}}$, resp., the corresponding partition and projection.

Theorem 3 *If* $C = C(\mathcal{A}, \mathcal{T})$ *and* $\operatorname{supp}(p(A)) = \mathcal{A}$, *then* $dIB(\Lambda) = \big\{\gamma \circ \pi \;:\; \gamma \in C_{cong}(\bar{\mathcal{A}}, \mathcal{T})\big\}$.

Proof See Appendices A and C.2. ■

I.e., for full support $p(A)$ and no constraints on the shape of encoders, the fully divergence-preserving solutions $\kappa \in \mathrm{dIB}(\Lambda)$ coincide, up to trivial transformations, with the clustering of $\mathcal{A}$ defined by the relation (6). This will yield the results from Sections 3.2 and 3.3.

3.2. Application to equivariances

Consider now $\mathcal{A} = \mathcal{X} \times \mathcal{Y}$ equipped with a full support distribution $p(X,Y)$; in particular, $p(Y|X)$ is well-defined. The group G_{ce} of *channel equivariances* is the group of pairs $(\sigma, \tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$ such that $p(Y|X) \circ \sigma = \tau \circ p(Y|X)$.

We want to design a dIB problem that mimics the projection π_{ce} on orbits of the equivariance group. Crucially, π_{ce} does not compress $\mathcal{X}$ and $\mathcal{Y}$ separately but rather "intertwines" them (Charvin et al., 2023), which motivates the choice of *joint* compression channels, i.e., we impose no constraint on their shape: $C = C(\mathcal{A}, \mathcal{T})$. Moreover, it can be verified[2] $\pi_{\mathrm{ce}}(x,y) = \pi_{\mathrm{ce}}(x',y')$ implies $p(y|x) = p(y'|x')$. Based on this observation, we search for an exponential family $\mathcal{E}$ such that the relation $\sim$ from equation (6) becomes

$$(x,y) \sim (x',y') \quad \Leftrightarrow \quad p(y|x) = p(y'|x'). \tag{7}$$

2. See Lemma 15 in (Charvin et al., 2023).

This is achieved by choosing

$$\mathcal{E} = \mathcal{E}_{\rm ce} := \{r(X)\mathcal{U}(\mathcal{Y}),\ r(X) \in \Delta_{\mathcal{X}}\},$$

which yields $\tilde{p}(X,Y) = p(X)\mathcal{U}(\mathcal{Y})$, so that $p(x,y)\tilde{p}(x',y') = p(x',y')\tilde{p}(x,y)$ if and only if $p(y|x) = p(y'|x')$. Borrowing from the geometric approach to complexity (Ay et al., 2011), we note that $\mathcal{E}_{\rm ce}$ coincides with the hierarchical model of probability distributions on $\mathcal{X} \times \mathcal{Y}$ that depend only on $\mathcal{X}$, which allows us to interpret $D(p(X,Y)||\mathcal{E})$ as measuring the "degree to which the system (X,Y) is more than just X" (see Appendix C.3 for details). As desired, this dIB problem, which we denote by $\mathrm{dIB}_{\rm ce}$, does characterise equivariances:[3]

Theorem 4 *The following holds, for $\Lambda := D(p||\mathcal{E})$ and all $(\sigma,\tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$:*

(i) Let $\kappa \in \mathrm{dIB}_{ce}(\Lambda)$. Then $(\sigma,\tau) \in G_{ce}$ if and only if $\kappa \circ (\sigma \otimes \tau) = \kappa$.

(ii) If $(\sigma,\tau) \in G_{ce}$, then $\kappa \circ (\sigma \otimes \tau) = \kappa$ also holds for all $0 \le \lambda \le \Lambda$ and $\kappa \in \mathrm{dIB}_{ce}(\lambda)$.

(iii) The projection π defined by $\sim$ in equation (7) does not, in general, coincide with π_{ce}.

Proof See Appendix C.4. The proof crucially relies on Theorem 3. ■

Point (i) means that equivariances of $p(Y|X)$ are those pairs of transformations (σ,τ) such that the effect of simultaneously transforming $\mathcal{X}$ with σ and $\mathcal{Y}$ with τ is "quotiented out" by the coarse-grainings $\kappa \in \mathrm{dIB}(\Lambda)$, making these transformations indiscernible from the identity. This is because from Theorem 3 and relation (7), these κ only distinguish pairs (x,y) with distinct $p(y|x)$, while equivariances (σ,τ) satisfy $p(\tau \cdot Y|\sigma \cdot X) = p(Y|X)$. Point (ii) means that the "quotienting out" of equivariances happens actually for all granularity λ, even though it is only a full characterisation for $\lambda = \Lambda$. But even though the problem $\mathrm{dIB}(\Lambda)$ characterises equivariances, point (iii) says that the corresponding clustering π does not always coincide with the projection on orbits $\pi_{\rm ce}$. This comes from the fact that here,the action of $G_{\rm ce}$ on a given orbit depends on its action on other orbits.

We can now draw upon our new information-theoretic characterisation of equivariances to soften this group-theoretic notion, where each granularity λ defines a corresponding set of soft equivariances. Let $0 \le \lambda \le \Lambda$. We define a λ-*equivariance* as a pair $(\mu,\eta) \in C(\mathcal{X}) \otimes C(\mathcal{Y})$ such that there exists $\kappa \in \mathrm{dIB}_{\rm ce}(\lambda)$ with $\kappa \circ (\mu \otimes \eta) = \kappa$. In other words, a soft equivariance is defined through the very same equation $\kappa \circ (\sigma \otimes \tau) = \kappa$ that characterises exact equivariances, but where the fully information-preserving compression κ is now only a *partially* information-preserving compression. Moreover, we allow μ and ν to be non-invertible and stochastic.

To conclude this section, let us point out that the classic IB can be recovered as a dIB with the same exponential family $\mathcal{E}_{\rm ce}$ as for equivariances, but with shape constraints $C \subsetneq C(\mathcal{A},\mathcal{T})$: i.e., by imposing that κ can only compress $\mathcal{X}$ and not $\mathcal{Y}$. See Appendix C.6.

3.3. Application to distribution invariances

We now apply a similar process to transformations that leave a given distribution invariant. I.e., let $p \in \Delta_{\mathcal{A}}$ be full support, and define the group $G_{\rm di}$ of *distribution invariances* as

3. Appendix C.5 clarifies how our results relate to (Charvin et al., 2023), from which this work is inspired.

the group of $\Phi \in \mathrm{Bij}(\mathcal{A})$ such that $p(\Phi \cdot A) = p(A)$. Because we do not consider any structure on $\mathcal{A}$, it is natural to choose $C = C(\mathcal{A}, \mathcal{T})$. Moreover, as $\Phi \in G_{\mathrm{di}}$ if and only if $p(a) = p(\Phi \cdot a)$ for all a, it is natural to search for an exponential family yielding the equivalence relation $a \sim a' \Leftrightarrow p(a) = p(a')$. It can be easily verified that this is achieved by choosing merely the uniform distribution: $\mathcal{E} = \mathcal{E}_{\mathrm{di}} := \{\mathcal{U}(\mathcal{A})\}$. Intuitively, here the dIB problem, which we denote by $\mathrm{dIB}_{\mathrm{di}}$, preserves (partially or wholly) the divergence $D(p(A)||\mathcal{U}(\mathcal{A})) = H(\mathcal{U}(\mathcal{A})) - H(A)$ of $p(A)$ from the uniform distribution: i.e., it preserves the "degree to which $p(A)$ is deterministic".

Theorem 5 *The following holds, for $\Lambda := D(p||\mathcal{E})$ and all $\Phi \in \mathrm{Bij}(\mathcal{A})$:*

(i) Let $\kappa \in \mathrm{dIB}_{di}(\Lambda)$. Then $\Phi \in G_{di}$ if and only if $\kappa \circ \Phi = \kappa$.

(ii) If $\Phi \in G_{ci}$, then $\kappa \circ \Phi = \kappa$ also holds for all $0 \leq \lambda \leq \Lambda$ and $\kappa \in \mathrm{dIB}_{di}(\lambda)$.

(iii) The projection π defined by $\sim$ coincides with the projection on orbits π_{di}.

Proof See Appendix C.7. The proof crucially relies on Theorem 3. ■

Interpretations of points (i) and (ii) are analogous to those for equivariances. Point (iii) highlights that here, π_{di} and π do coincide: contrarily to equivariances, the elements of G_{di} are not required to respect any factorisation of $\mathcal{A}$. Eventually, one can directly adapt the definition of soft equivariances to one for soft distribution invariances.

3.4. Relevant computational and conceptual tools

Here, we present an iterative algorithm (for unconstrained encoder shape) and two important concepts for the dIB problem. Consider the Lagrangian relaxation of the dIB problem,

$$\arg\min_{\kappa \in C} \Big[I_\kappa(A;T) - \beta D(\kappa \cdot p || \kappa \cdot \tilde{p}) \Big], \tag{8}$$

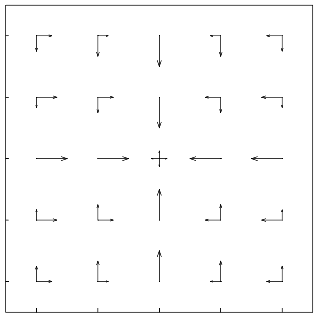
where $\beta \geq 0$. For $C = C(\mathcal{A}, \mathcal{T})$, we obtain, deriving w.r.t. κ, the following necessary condition for local minimisers $\kappa = q(T|A)$ of (8): for all $a \in \mathrm{supp}(p(A))$ and $t \in \mathrm{supp}(q(T))$,

$$q(t|a) = \frac{1}{Z(a,\beta)} q(t) \exp\left[-\beta \left(\frac{q(t)\tilde{p}(a)}{\tilde{q}(t)p(a)} - \log\left(\frac{q(t)\tilde{p}(a)}{\tilde{q}(t)p(a)} \right) - 1 \right) \right], \tag{9}$$

where $q(t) := \sum_a p(a)q(t|a)$ and $\tilde{q}(t) := \sum_a \tilde{p}(a)q(t|a)$, with $Z(a,\beta)$ a normaliser. From this fixed-point equation, we obtain a Blahut-Arimoto (BA) algorithm which is not provably convergent to a global minimum but has the same guarantees as BA for the classic IB (Tishby et al., 2000) (see Appendix D.2). This algorithm is used in the experiments of Section 4.1[4]. In the following, we will write κ_β the output of the BA algorithm with parameter β, and also $I_\beta := I_{\kappa_\beta}(A;T)$ and $D_\beta := D(\kappa_\beta \cdot p || \kappa_\beta \cdot \tilde{p})$. Note that both I_β and D_β increase with β.

Now the *effective cardinality* (Zaslavsky and Tishby, 2019) of some $\kappa \in \mathrm{dIB}(\lambda)$ is defined as the minimum number of symbols t necessary to describe the output of κ (see Appendix D.3 for a formal definition). In all our numerical experiments, we observed that: (i) similarly

4. In our experiments, we choose $|\mathcal{T}| = \mathrm{supp}(p(A)) + 1$.



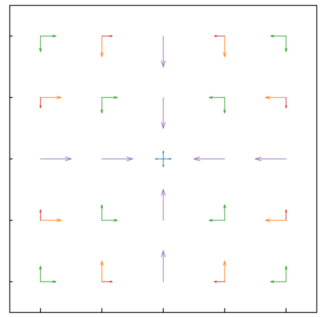


Figure 1: Left: representation of $p(Y|X)$, where X is the position on the grid, Y the gradient direction, and probabilities are proportional to arrow lengths. Right: same figure with colors representing a clustering of $\mathcal{X} \times \mathcal{Y}$; the cluster $\text{supp}(p(X,Y))^c$ corresponds to all arrows with length 0, and thus has no color. This clustering is obtained in 3 distinct ways: (i) as the projection on orbits of the equivariance group of $p(Y|X)$; (ii) as the clustering defined by relation (7); and (iii) by perturbing $p(X,Y)$ as described in the main text, and then computing $q_{\beta_{\text{ce}}}(T|X,Y)$.

as for the classic IB, effective cardinality monotically increases with β, and (ii) changes of effective cardinality coincide with discontinuities in the slope of the curve $\beta \mapsto (I_\beta, D_\beta)$, which is reminiscent of the second-order bifurcations observed for the IB (Zaslavsky and Tishby, 2019). We will thus here refer to changes of effective cardinalities as bifurcations.

Eventually, we want to investigate whether the equations $\kappa_\beta \circ \Phi = \kappa_\beta$ is satisfied for varying β and varying $\Phi \in G$, with G some fixed subgroup of $\text{Bij}(\mathcal{A})$. But numerically, it is also important to quantify, when this equation is not exactly satisfied, the extent of the deviation. We propose to use the divergence defined for all channel $\kappa \in C(\mathcal{A}, \mathcal{T})$ as

$$D^p(\kappa||C_G) := \min_{\nu \in C_G} D^p(\kappa||\nu) := \min_{\nu \in C_G} \sum_{a \in \text{supp}(p(A))} p(a) D(\kappa(T|a)||\nu(T|a)),$$

where $C_G := \{\nu : \forall \Phi \in G, \nu \circ \Phi = \nu\}$ is the family of channels that are exactly input-symmetric w.r.t G. Intuitively, $D^p(\kappa||C_G)$ measures the divergence of the channel κ from being input-symmetric for the action of G on the distribution $p(A)$. In particular, $D^p(\kappa||C_G) = 0$ if and only if $\kappa \circ \Phi = \kappa$ for all $\Phi \in G$. See Appendix D.4 for more details.

4. Numerical experiments

4.1. Synthetic numerical experiments on equivariances

To illustrate the relevance of the dIB_{ce} characterising equivariances, we propose to study soft equivariances in a simple synthetic grid-world scenario. Here $\mathcal{X}$ stands for positions on a 5×5 grid, and $\mathcal{Y}$ for a gradient with 4 possible directions. Thus $p(Y|X)$ describes the probability of a direction at a given position, which can be thought of, e.g., as a nutrient gradient sensed by a bacteria. We choose uniform $p(X)$ (choosing non-uniform $p(X)$ resulted

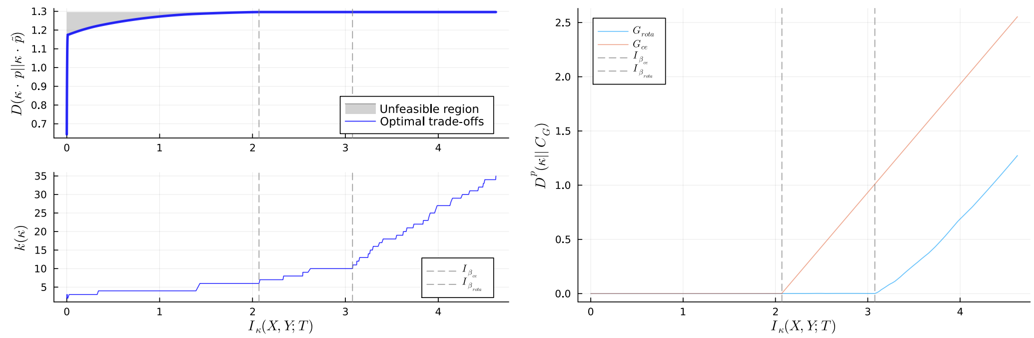


Figure 2: $D_\beta := D(\kappa_\beta \cdot p||\kappa_\beta \cdot \tilde{p})$ as a function of $I_\beta := I_{\kappa_\beta}(X,Y;T)$. Bottom left: Effective cardinality $k(\kappa)$ as a function of I_β. Right: Divergence of compression channels κ_β as a function of I_β, for the groups G_{ce} and G_{rota}. The vertical dashed lines represent specific bifurcations of the parameter β at which $D^p(\kappa_\beta||C_{G_{\text{rota}}})$, resp. $D^p(\kappa_\beta||C_{G_{\text{ce}}})$, approximately vanishes (in decreasing order of I_β).

in similar results). As seen in Figure 1, left, $p(Y|X)$ has many symmetries: it can be verified that the equivariance group of $p(Y|X)$ has 6 distinct orbits (one is $\text{supp}(p)^c$), represented in Figure 1, right. Moreover, even though we saw in Theorem 4, point (iii), that the projection on orbits π_{ce} does not generally coincide with the projection π defined by relation (7), here the two projections actually coincide. Thus Figure 1, right, also represents π.

From Section 3.4, we have $D^p(\kappa_\beta||C_{G_{\text{ce}}}) = 0$ if and only if $\kappa_\beta \circ (\sigma, \tau) = \kappa_\beta$ for all $(\sigma, \tau) \in G_{\text{ce}}$. Theorem 4, point (ii), suggests that this equation should indeed hold for all β.[5] As a sanity check, we thus computed the dIB$_{\text{ce}}$ bottlenecks κ_β for $0 \leq D_\beta \leq \Lambda$, and indeed obtained $D^p(\kappa_\beta||C_{G_{\text{ce}}}) \leq 3e^{-16}$ for all β. We also noted that the bottlenecks' effective cardinality monotonically increases from 1 for $D_\beta = 0$ to 6 for $D_\beta = \Lambda$.

We then perturb $p(Y|X)$ with two distinct random perturbations. The first one, of larger amplitude, satisfies the constraint that after the perturbation, $p(Y|X)$ still exactly satisfies the equivariances defined by a subgroup $G_{\text{rota}} \subsetneq G_{\text{ce}}$ of it. This subgroup is generated by rotating both the positions and the gradient directions by 90 degrees. The second perturbation applied to $p(Y|X)$, of smaller amplitude, breaks all the remaining equivariances of $p(Y|X)$. We obtain a new $p(Y|X)$ which, intuitively, is still "approximately" equivariant, but where the approximate equivariances in $G_{\text{ce}} \setminus G_{\text{rota}}$ are *coarser* than those in G_{rota}, because the perturbation was larger for the former than for the latter.

We compute 1000 dIB$_{\text{ce}}$-bottlenecks for varying β. The resulting information curve (I_β, D_β), along with the corresponding effective cardinalities, are shown in Figure 2, left. As for the classic IB, we obtain a non-decreasing and concave information curve, and an increasing effective cardinality (except for small I_β, which could be due to numerical errors).

Crucially, we then observe (Figure 2, right) that for decreasing β, the divergences $D^p(\kappa_\beta||C_{G_{\text{rota}}})$ and $D^p(\kappa_\beta||C_{G_{\text{ce}}})$ successively vanish, at bifurcation values β_{rota}, resp. $\beta_{\text{ce}} < \beta_{\text{rota}}$. Thus the perturbed equivariances are here recovered by the dIB$_{\text{ce}}$ method as *soft*

5. Here the full support assumption, which is required in Theorem 4, does not hold for $p(X,Y)$. We leave to future work the theoretical study of the non full support case.

equivariances, for low enough I_β. Moreover, as the equivariances from G_{rota} have been less perturbed that those in the remaining of G_{ce}, here the degree of compression required to recover an approximate equivariance scales with the "coarseness" of that equivariance.

The fact that all the original equivariances are recovered for $\beta \leq \beta_{\text{ce}}$ is further supported by inspecting the the bottleneck for $\beta = \beta_{\text{ce}}$. Indeed, $q_{\beta_{\text{ce}}}(T|X,Y)$ coincides exactly with π_{ce} (and can thus be represented by Figure 1, right). This is not the case for $q_\beta(T|X,Y)$ when $\beta > \beta_{\text{ce}}$. Let us stress, though, that this situation partially stems from the fact that here, for the unperturbed equivariant $p(Y|X)$, the projections π_{ce} and π coincide (see above).

Eventually, note that, in Figure 2, left, the gain in divergence D_β from $I_\beta = I_{\beta_{\text{ce}}}$ to the maximum value $I_{\max}$ of I_β is negligible, whereas $I_{\max} - I_{\beta_{\text{ce}}}$ is large. This resonates with the intuition that coarse symmetries in raw data allow for a potentially drastic informational compression (I_β here), under a negligible loss in informational accuracy (D_β here).

5. Limitations

Our core results are of theoretical nature, and hold in the discrete and full support case. At this stage, it is still unclear whether and how they extend to continuous and non fully supported distributions. Numerically, the BA class of algorithms addresses only the discrete case and generally scales unfavourably in larger scenarios. Future work could make the dIB problem amenable to deep network optimisation by adapting the classic IB's variational bounds (Alemi et al., 2019). Moreover, to use the algorithm to extract concrete equivariances, one would need to solve the symmetry equations from, e.g., point (i) in resp. Theorem 4 and 5. The computational feasibility of solving these equations partly or in whole is not clear yet. Finally, recall that from point (ii) in Theorem 4, the dIB corresponding to equivariances does not yield the projection on orbits under the group's action. It is an open question whether a suitable instance or variation of the divergence IB would achieve this.

6. Conclusion

Motivated by the ability of the classic IB to implicitly extract channel invariances, we investigated generalizations of this phenomenon. For this, we introduced the Divergence IB, a novel and substantial generalization of the classic IB. We show how this method can generalize the informational characterisation of invariances to that of channel equivariances and distribution invariances. Crucially, expressing these symmetries through IB-like trade-offs yields a natural softening of these very stringent group-theoretic notions. This suggests a principled route to extract coarse-grained symmetries through *structure-preserving coarse-grainings* of the given data, thus exposing its "platonic core", so to say.

However, while we only investigated some canonical examples of symmetries, the dIB framework is highly versatile. With other exponetial families $\mathcal{E}$ and channel shape constraints C, this method could help discover novel kinds of scientifically relevant structures.

Eventually, our work opens a new path for formalisations of the intuition that informationally parsimonious systems, e.g., biological agents, should internalise the coarse symmetries in their environment — in other words, for understanding the emergence of symmetries in neural systems (Bertoni et al., 2021) from the point of view of bounded rationality (Genewein et al., 2015).

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.

Nihat Ay. Information Geometry on Complexity and Stochastic Interaction. *Entropy*, 17(4): 2432–2458, April 2015. ISSN 1099-4300. doi: 10.3390/e17042432.

Nihat Ay, Eckehard Olbrich, Nils Bertschinger, and Jürgen Jost. A geometric approach to complexity. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(3):037103, September 2011. ISSN 1054-1500. doi: 10.1063/1.3638446.

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.

Federico Bertoni, Noemi Montobbio, Alessandro Sarti, and Giovanna Citti. Emergence of Lie Symmetries in Functional Architectures Learned by CNNs. *Frontiers in Computational Neuroscience*, 15, November 2021. ISSN 1662-5188. doi: 10.3389/fncom.2021.694505.

Hippolyte Charvin, Nicola Catenacci Volpi, and Daniel Polani. Towards Information Theory-Based Discovery of Equivariances, November 2023.

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.

Jan E. Gerken, Jimmy Aronsson, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Geometric deep learning and equivariant neural networks. *Artificial Intelligence Review*, 56(12):14605–14662, December 2023. ISSN 1573-7462. doi: 10.1007/s10462-023-10502-7.

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.

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.

Ziming Liu and Max Tegmark. Machine-learning hidden symmetries. *Physical Review Letters*, 128(18):180201, May 2022. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.128.180201.

Zygmunt Pizlo and J. Acacio de Barros. The Concept of Symmetry and the Theory of Perception. *Frontiers in Computational Neuroscience*, 15, August 2021. ISSN 1662-5188. doi: 10.3389/fncom.2021.681162.

Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29):2696–2711, 2010. ISSN 0304-3975. doi: 10.1016/j.tcs.2010.04.006.

Ian Stewart. Symmetry and Network Topology in Neuronal Circuits: Complicity of Form and Function. *International Journal of Bifurcation and Chaos*, November 2022. doi: 10.1142/S0218127422300336.

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

Raymond W. Yeung. *Information Theory and Network Coding*. Springer, 2008.

Abdellatif Zaidi, Iñaki Estella-Aguerri, and Shlomo Shamai (Shitz). On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views. *Entropy*, 22(2), 2020. ISSN 1099-4300. doi: 10.3390/e22020151.

Noga Zaslavsky and Naftali Tishby. Deterministic annealing and the evolution of Information Bottleneck representations. August 2019.

Appendix A. General notations, definitions and results

The proofs of Theorem 2 (see Appendix B) and Theorem 3 (see Appendix C.2.) are very similar. Here, we collect definitions and pieces of reasoning that are common to both.

We fix a finite set $\mathcal{A}$, a probability $p \in \Delta_{\mathcal{A}}$ which we assume full support, and a channel $\kappa = q(T|A) \in C(\mathcal{A}, \mathcal{T})$. We also consider a partition $\bar{\mathcal{A}} = \{\mathcal{A}_j\}_{j=1,\dots,n}$ of $\mathcal{A}$, and denote by $\pi : \mathcal{A} \to \bar{\mathcal{A}}$ the corresponding projection. Whenever it can simplify notations, we will identify $\bar{\mathcal{A}}$ with $\{1, \dots, n\}$. We associate to $\kappa = q(T|A)$ a corresponding channel $\bar{\kappa} = \bar{q}(T|\mathcal{A}_J) \in C(\bar{\mathcal{A}}, \mathcal{T})$ defined through

$$\bar{q}(t|j) := \bar{q}(t|\mathcal{A}_j) := q(t|\mathcal{A}_j) := \frac{\sum_{a\in\mathcal{A}_j} q(t|a)p(a)}{p(\mathcal{A}_j)}. \tag{10}$$

Intuitively, $\bar{q}(T|\mathcal{A}_J)$ is the "quotient of $q(T|A)$ defined by the partition $\bar{\mathcal{A}}$ and $p(A)$". We also define a channel $\kappa_\pi = q_\pi(T|A) \in C(\mathcal{A}, \mathcal{T})$ through, for all $a \in \mathcal{A}$, $t \in \mathcal{T}$,

$$q_\pi(t|a) := (\bar{\kappa} \circ \pi)(t|a) = \sum_j q(t|\mathcal{A}_j)\delta_{a\in\mathcal{A}_j}. \tag{11}$$

Intuitively, $q_\pi(T|A)$ is the "enforced factorisation of $q(T|A)$ through π". Indeed: (i) it is a channel defined from $q(T|A)$ that factorises through π; and (ii), whenever $q(T|A)$ itself factorises through π, then we must have $q(T|A) = q_\pi(T|A)$ (see point (i) in Lemma 8).

Lemma 6 *We have*

(i) $q_\pi(T) = q(T)$.

(ii) $I_{q_\pi}(A;T) \le I_q(A;T)$, *where equality holds if and only if* $q_\pi(T|A) = q(T|A)$.

Proof (i). For all $t \in \mathcal{T}$,

$$\begin{aligned} q_\pi(t) &= \sum_{a\in\mathcal{A}} p(a)q_\pi(t|a) = \sum_{a\in\mathcal{A}} p(a)q_\pi(t|a) \\ &= \sum_j \sum_{a\in\mathcal{A}_j} p(a)q_\pi(t|a) = \sum_j \sum_{a\in\mathcal{A}_j} p(a)q(t|\mathcal{A}_j) \\ &= \sum_j q(t|\mathcal{A}_j)p(\mathcal{A}_j) = q(t), \end{aligned}$$

where the third and last equality use the fact that $\{\mathcal{A}_j\}_j$ is a partition of $\mathcal{A}$.

(ii). We have

$$\begin{aligned} I_q(A;T) &= \sum_{a\in\mathcal{A}, t\in\mathrm{supp}(q(T))} p(a)q(t|a)\log\left(\frac{q(t|a)}{q(t)}\right) \\ &= \sum_{t\in\mathrm{supp}(q(T))} \sum_j \sum_{a\in\mathcal{A}_j} p(a)q(t|a)\log\left(\frac{q(t|a)}{q(t)}\right). \end{aligned}$$

But from the log-sum inequality (Csiszár and Körner, 2011), for fixed $t \in \mathrm{supp}(q(T))$ and fixed j,

$$\begin{aligned} \sum_{a\in\mathcal{A}_j} p(a)q(t|a)\log\left(\frac{q(t|a)}{q(t)}\right) &= \sum_{a\in\mathcal{A}_j} p(a)q(t|a)\log\left(\frac{p(a)q(t|a)}{p(a)q(t)}\right) \\ &\ge \left(\sum_{a\in\mathcal{A}_j} p(a)q(t|a)\right)\log\left(\frac{\sum_{a\in\mathcal{A}_j} p(a)q(t|a)}{\sum_{a\in\mathcal{A}_j} p(a)q(t)}\right) \qquad (12) \\ &= p(\mathcal{A}_j)q(t|\mathcal{A}_j)\log\left(\frac{q(t|\mathcal{A}_j)}{q(t)}\right), \end{aligned}$$

with equality in (12) if and only if $\frac{q(t|a)p(a)}{q(t)p(a)}$ is constant for $a \in \mathcal{A}_j$, i.e., if and only if $q(t|a)$ is constant for $a \in \mathcal{A}_j$. Note that the last line of (12) can be rewritten

$$\begin{aligned} p(\mathcal{A}_j)q(t|\mathcal{A}_j)\log\left(\frac{q(t|\mathcal{A}_j)}{q(t)}\right) &= \sum_{a\in\mathcal{A}_j} p(a)q_\pi(t|a)\log\left(\frac{q_\pi(t|a)}{q(t)}\right) \\ &= \sum_{a\in\mathcal{A}_j} p(a)q_\pi(t|a)\log\left(\frac{q_\pi(t|a)}{q_\pi(t)}\right), \end{aligned}$$

where the second equality uses point (i) proven above. Thus, summing (12) over j and $t \in \mathrm{supp}(q(T))$, we get

$$\begin{aligned} I_q(A;T) &\geq \sum_{t \in \mathrm{supp}(q(T))} \sum_j \sum_{a \in \mathcal{A}_j} p(a) q_\pi(t|a) \log\left(\frac{q_\pi(t|a)}{q_\pi(t)}\right) \\ &= I_{q_\pi}(A;T), \end{aligned}$$

with equality if and only if for all $t \in \mathrm{supp}(q(T))$ and all j, the quantity $q(t|a)$ is constant for $a \in \mathcal{A}_j$ — which means more precisely that for all $a \in \mathcal{A}_j$, we have $q(t|a) = q(t|\mathcal{A}_j)$. In other words, there is equality if and only if $q(t|a) = q_\pi(t|a)$ for all $a \in \mathcal{A}$ and $t \in \mathrm{supp}(q(T))$. As the definition of $q_\pi(T|A)$ clearly implies $\mathrm{supp}(q(T)) = \mathrm{supp}(q_\pi(T))$, the latter is equivalent to $q(T|A) = q_\pi(T|A)$. ■

We now introduce, for all $t \in \mathcal{T}$, the set

$$\mathcal{A}_t^q := \{a \in \mathcal{A} : \quad q(t|a) > 0\}\,, \tag{13}$$

which can be seen as the "probabilistic pre-image of t through the channel $q(T|A)$".

Lemma 7 *The following are equivalent:*

(i) The channel $\bar{q}(T|\mathcal{A}_J) \in C(\bar{\mathcal{A}}, \mathcal{T})$ defined in (10) is congruent.

(ii) For all $t \in \mathrm{supp}(q(T))$, there exists a partition element $\mathcal{A}_j \in \bar{\mathcal{A}}$ such that $\mathcal{A}_t^q \subseteq \mathcal{A}_j$.

Note that the $\mathcal{A}_j$ satisfying $\mathcal{A}_t^q \subseteq \mathcal{A}_j$ is actually unique, because $\{\mathcal{A}_j\}_j$ is a partition of $\mathcal{A}$, and $\mathcal{A}_t^q \neq \emptyset$ for $t \in \mathrm{supp}(q(T))$.
Proof For any function $f : \mathcal{T} \to \{1, \ldots, n\}$ and all j, j',

$$(f \circ \bar{q})(j'|j) = \sum_{t \in \mathcal{T}} \delta_{f(t)=j'} \bar{q}(t|j) = \sum_{t \in \mathcal{T}} \delta_{f(t)=j'} q(t|\mathcal{A}_j) = q(f^{-1}(j')|\mathcal{A}_j)$$

Thus

$$\begin{aligned} (f \circ \bar{q})(j'|j) = \delta_{j'=j} \quad &\Leftrightarrow \quad q(f^{-1}(j')|\mathcal{A}_j) = \delta_{j'=j} && (14)\\ &\Leftrightarrow \quad q(f^{-1}(j')|\mathcal{A}_j) > 0 \text{ only if } j' = j && (15)\\ &\Leftrightarrow \quad \forall t \in \mathrm{supp}(q(T)),\quad q(t|\mathcal{A}_j) > 0 \text{ only if } f(t) = j && (16)\\ &\Leftrightarrow \quad \forall t \in \mathrm{supp}(q(T)),\quad (\exists a \in \mathcal{A}_j : q(t|a) > 0) \text{ only if } f(t) = j && (17)\\ &\Leftrightarrow \quad \forall t \in \mathrm{supp}(q(T)),\quad (\mathcal{A}_t^q \cap \mathcal{A}_j \neq \emptyset) \text{ only if } f(t) = j && (18) \end{aligned}$$

where line (15) uses the fact that $q(\cdot|\mathcal{A}_j)$ is a probability measure and $\{f^{-1}(j)\}_j$ a partition of $\mathcal{T}$; line (16) that $q(f^{-1}(j')|\mathcal{A}_j) = \sum_{t:f(t)=j'} q(t|\mathcal{A}_j)$; line (17) that $q(t|\mathcal{A}_j) = \frac{1}{p(\mathcal{A}_j)} \sum_{a \in \mathcal{A}_j} q(t|a)p(a)$ with $\mathcal{A}_j \subseteq \mathcal{A}$; and line (18) the definition of $\mathcal{A}_t^q$. Therefore,

$$\begin{aligned} &\exists f : \mathcal{T} \to \{1, \ldots, n\},\ \forall j, j',\ (f \circ \bar{q})(j'|j) = \delta_{j'=j} && (19)\\ &\quad\Leftrightarrow \quad \exists f : \mathcal{T} \to \{1, \ldots, n\},\ \forall j, \forall t \in \mathrm{supp}(q(T)),\ (\mathcal{A}_t^q \cap \mathcal{A}_j \neq \emptyset) \text{ only if } f(t) = j \\ &\quad\Leftrightarrow \quad \exists f : \mathcal{T} \to \{1, \ldots, n\},\ \forall t \in \mathrm{supp}(q(T)),\ \mathcal{A}_t^q \subseteq \mathcal{A}_{f(t)} && (20) \end{aligned}$$

But on the one hand, the statement (19) is the definition of $\bar{q}(T|\mathcal{A}_J)$ being a congruent channel. On the other hand, because any function on $\mathrm{supp}(q(T))$ can be arbitrarily extended to the whole $\mathcal{T}$, the statement (20) is equivalent to

$$\exists f : \mathrm{supp}(q(T)) \to \{1, \ldots, n\},\ \forall t \in \mathrm{supp}(q(T)),\ \mathcal{A}_t^q \subseteq \mathcal{A}_{f(t)},$$

which is clearly a reformulation of point (ii). ■

Lemma 8 *Fix a channel $\gamma \in C(\bar{\mathcal{A}}, \mathcal{T})$, and define $\kappa := q(T|A) := \gamma \circ \pi$. Then*

(i) γ coincides with the channel $\bar{\kappa} := \bar{q}(T|\mathcal{A}_J)$ defined in equation (10).

(ii) If moreover γ is congruent, then $I_q(A;T) = H(\pi(A))$.

Proof (i). For all $\mathcal{A}_j \in \bar{\mathcal{A}}$,

$$\bar{q}(t|\mathcal{A}_j) := \sum_{a \in \mathcal{A}_j} \frac{p(a)q(t|a)}{p(\mathcal{A}_j)} = \sum_{a \in \mathcal{A}_j} \frac{p(a)\gamma(t|\pi(a))}{p(\mathcal{A}_j)} = \sum_{a \in \mathcal{A}_j} \frac{p(a)\gamma(t|\mathcal{A}_j)}{p(\mathcal{A}_j)} = \gamma(t|\mathcal{A}_j).$$

(ii). Fix a deterministic function $f : \mathcal{T} \to \bar{\mathcal{A}}$ such that $f \circ \gamma = e_{\bar{\mathcal{A}}}$. Then

$$I_q(A;T) \geq I_q(A; f(T)) = I_q(A; f \circ \gamma(\pi(A))) = I(A; \pi(A)) = H(\pi(A)),$$

where the last equality holds because π is deterministic. On the other hand, as a direct consequence of the factorisation $q(T|A) = \gamma \circ \pi$, we have the Markov chain $A - \pi(A) - T$. Thus $I_q(A;T) \leq I(A; \pi(A)) = H(\pi(A))$. ■

Appendix B. Proof of Theorem 2

The following sections present the successive steps of the proof. Let us first note that all the definitions and statements from Appendix A can be applied here, with $\mathcal{A} = \mathcal{X}$ and $p(A) = p(X)$ (we assumed that $p(A)$ is full support in Appendix A, but also that $p(X)$ is full support in Section 2). We write here $\bar{\mathcal{A}} = \bar{\mathcal{X}} = \{\mathcal{X}_j\}_{j=1,\ldots,n}$ and $\pi = \pi_{\mathcal{X}} : \mathcal{X} \to \bar{\mathcal{X}}$ (these notations coincides with those defined in Section 2); also $q(T|A) = q(T|X)$, $\bar{q}(T|\mathcal{A}_J) = \bar{q}(T|\mathcal{X}_J)$, $q_\pi(T|A) = q_{\pi_{\mathcal{X}}}(T|X)$. Moreover, recall that here $q(T|X)$ defines not only a joint distribution $q(X,T)$, but also, together with $p(X,Y)$, a joint distribution $q(X,Y,T)$ through the assumed Markov chain $T - X - Y$. Similarly, $q_\pi(T|X)$ defines a joint distribution $q_\pi(X,Y,T)$.

B.1. Factorisation for all parameter λ

Proposition 9 *For all $0 \leq \lambda \leq \Lambda$, every solution $q(T|X) \in IB(\lambda)$ factorises as $q(T|X) = \bar{q}(T|\mathcal{X}_J) \circ \pi_{\mathcal{X}}$.*

The proof will derive from the following lemma:

Lemma 10 *For every $q(T|X)$, we have:*

(i) $q_{\pi_\mathcal{X}}(Y;T) = q(Y,T)$, *so that in particular* $I_{q_{\pi_\mathcal{X}}}(Y;T) = I_q(Y;T)$.

(ii) $I_{q_{\pi_\mathcal{X}}}(X;T) \leq I_q(X;T)$, *where equality holds if and only if* $q_{\pi_\mathcal{X}}(T|X) = q(T|X)$.

Proof (i). For all $y \in \mathcal{Y}, t \in \mathcal{T}$,

$$\begin{aligned} q_{\pi_\mathcal{X}}(y,t) &= \sum_{x\in\mathcal{X}} p(y,x) q_{\pi_\mathcal{X}}(t|x) = \sum_{j=1}^{n} \sum_{x\in\mathcal{X}_j} p(y,x) q_{\pi_\mathcal{X}}(t|x) \\ &= \sum_{j=1}^{n} \sum_{x,x'\in\mathcal{X}_j} p(y,x)p(x') \frac{q(t|x')}{p(\mathcal{X}_j)} \\ &= \sum_{j=1}^{n} \sum_{x,x'\in\mathcal{X}_j} p(y,x')p(x) \frac{q(t|x')}{p(\mathcal{X}_j)} \qquad (21) \\ &= \sum_{j=1}^{n} \sum_{x'\in\mathcal{X}_j} p(y,x') q(t|x') \\ &= \sum_{x\in\mathcal{X}} p(y,x') q(t|x') = q(y,t), \end{aligned}$$

where the first and last lines use that $\{\mathcal{X}_j\}_j$ is a partition of $\mathcal{X}$; and line (21) uses the definition of the sets $\mathcal{X}_j$ through the equivalence relation $\sim_\mathcal{X}$ (see equation (2)): i.e., for $x, x' \in \mathcal{X}_j$, we have for all $y \in \mathcal{Y}$, $p(y|x) = p(y|x')$, which is equivalent to $p(y,x)p(x') = p(y,x')p(x)$.

(ii). We apply point (ii) in Lemma 6. ■

But Lemma 10 means that for the IB problem (1), if we replace the chanel $q(T|X) \in C(\mathcal{X}, \mathcal{T})$ by the corresponding $q_{\pi_\mathcal{X}}(T|X)$, then (i) the value for of the constraint function is unchanged, and (ii) the value of the target function does not increase, with equality if and only if $q_{\pi_\mathcal{X}}(T|X) = q(T|X)$. In particular, if $q(T|X)$ solves the IB problem, then we must have $q(T|X) = q_{\pi_\mathcal{X}}(T|X)$: i.e., $q(T|X) = \bar{q} \circ \pi_\mathcal{X}$.

B.2. Explicit form of solutions for $\lambda = \Lambda$ (point (i) in Theorem 2)

In this section, we prove point (i) in Theorem 2, i.e., that $\mathrm{IB}(\Lambda) = \{\gamma \circ \pi_\mathcal{X} : \quad \gamma \in C_{\mathrm{cong}}(\mathcal{X}_J, \mathcal{T})\}$.

We will here denote by $\mathcal{X}_t^q$ the "probabilistic pre-image" $\mathcal{A}_t^q$ from Appendix A (see equation (13)): i.e., for a channel $q = q(T|X)$ and all $t \in \mathcal{T}$,

$$\mathcal{X}_t^q := \{x \in \mathcal{X} : \quad q(t|x) > 0\}. \qquad (22)$$

Lemma 11 *We have* $I_q(T;Y) \leq \Lambda := I(X;Y)$, *and the following are equivalent:*

(i) $I_q(T;Y) = \Lambda$.

(ii) For all $t \in \mathrm{supp}(q(T))$, *there exists a partition element* $\mathcal{X}_j \in \bar{\mathcal{X}}$ *such that* $\mathcal{X}_t^q \subseteq \mathcal{X}_j$.

(iii) The channel $\bar{q}(T|\mathcal{X}_J) \in C(\bar{\mathcal{X}}, \mathcal{T})$ *defined in* (10) *is congruent.*

The equivalence of (i) and (ii) means that the constraint $I_q(T;Y) = I(X;Y)$ holds if and only if $p(Y|x)$ is constant on the pre-image $\mathcal{X}_t^q$ of every symbol t.

Proof $(i) \Leftrightarrow (ii)$. We have

$$\begin{aligned} I_q(T;Y) &= \sum_{y \in \mathrm{supp}(p(Y)), t \in \mathrm{supp}(q(T))} p(y) q(t|y) \log\left(\frac{q(t|y)}{q(t)}\right) \\ &= \sum_{y \in \mathrm{supp}(p(Y)), t \in \mathrm{supp}(q(T))} p(y) \left(\sum_x p(x|y) q(t|x)\right) \log\left(\frac{\sum_x p(x|y) q(t|x)}{\sum_x p(x) q(t|x)}\right). \end{aligned}$$

But for all $y \in \mathrm{supp}(p(Y))$ and $t \in \mathrm{supp}(q(T))$, from the log-sum inequality (Csiszár and Körner, 2011), with the convention $0 \log(\frac{0}{0}) := 0$,

$$\left(\sum_x p(x|y) q(t|x)\right) \log\left(\frac{\sum_x p(x|y) q(t|x)}{\sum_x p(x) q(t|x)}\right) \leq \sum_x p(x|y) q(t|x) \log\left(\frac{p(x|y) q(t|x)}{p(x) q(t|x)}\right). \tag{23}$$

So that, summing over y and t, we get $I_q(Y;T) \leq I(X;Y)$, with equality if and only if for all $y \in \mathrm{supp}(p(Y)), t \in \mathrm{supp}(q(T))$, it holds in (23). From the equality case of the log-sum inequality (Csiszár and Körner, 2011), the latter is equivalent to the existence of nonzero constants $(\alpha_{y,t})_{y \in \mathrm{supp}(p(Y)), t \in \mathrm{supp}(q(T))}$ such that

$$\forall x \in \mathcal{X}, \quad p(x) q(t|x) = \alpha_{y,t} p(x|y) q(t|x),$$

i.e., such that, for all $y \in \mathrm{supp}(p(Y)), t \in \mathrm{supp}(q(T))$, the quantity $\frac{p(x|y)}{p(x)}$ is constant on the subset of elements $x \in \mathcal{X}$ for which $q(t|x) > 0$. But the latter subset is precisely $\mathcal{X}_t^q$ (see definition (13)), and

$$\frac{p(x|y)}{p(x)} = \frac{1}{p(y)} p(y|x),$$

where $\frac{1}{p(y)}$ does not depend on x. Thus we proved that $I(Y;T) = I(X;Y)$ holds if and only if for all $t \in \mathrm{supp}(q(T))$, the distribution $p(Y|x)$ does not depend on $x \in \mathcal{X}_t^q$: i.e., if and only if for all $t \in \mathrm{supp}(q(T))$, there exists an $\mathcal{X}_j$ such that $\mathcal{X}_t^q \subseteq \mathcal{X}_j$.

$(ii) \Leftrightarrow (iii)$. Apply Lemma 7. ■

Combining the previous results directly yields that

$$\mathrm{IB}(\Lambda) \subseteq E := \{\gamma \circ \pi_{\mathcal{X}}, \quad \gamma \in C_{\mathrm{cong}}(\{1, \ldots, n\}, \mathcal{T})\}.$$

Indeed, fix a solution $q(T|X) \in \mathrm{IB}(\Lambda)$. Proposition 9 proves that $q(T|X) = \bar{q}(T|\mathcal{X}_J) \circ \pi_{\mathcal{X}}$. But because we must have $I_q(Y;T) = \Lambda := I(X;Y)$, Lemma 11 yields that $\bar{q}(T|\mathcal{X}_J)$ is here congruent.

Let us now prove the converse inclusion, i.e., that $E \subseteq \mathrm{IB}(\Lambda)$.

Lemma 12 *For all $q(T|X) \in E$, we have $I_q(T;Y) = I(X;Y)$ and $I_q(X;T) = H(\pi_{\mathcal{X}}(X))$.*

Proof From point (i) in Lemma 8, $q(T|X) = \gamma \circ \pi_{\mathcal{X}}$ implies $\gamma = \bar{q}(T|\mathcal{X}_J)$. But by definition of E, here γ is assumed congruent, thus $\bar{q}(T|\mathcal{X}_J)$ is congruent. So that from Lemma 11, $I_q(T;Y) = I(X;Y)$. Point (ii) in Lemma 8 yields $I_q(X;T) = H(\pi_{\mathcal{X}}(X))$. ■

Now, because the IB problem is defined as the minimisation of a continuous function on a compact domain, it has at least one solution, say $q_*(T|X)$, which we know belongs to E from the the inclusion $\mathrm{IB}(\Lambda) \subseteq E$ that we already proved. But Lemma 12 then implies that for all $q(T|X) \in E$, we have $I_q(T;Y) = I_{q_*}(T;Y)$ and $I_q(X;T) = I_{q_*}(X;T)$. Thus any $q(T|X) \in E$ must also be a solution, i.e., $q(T|X) \in \mathrm{IB}(\Lambda)$. This ends the proof of point (i) in Theorem 2.

B.3. End of the proof of Theorem 2

Proposition 9 ensures that for all λ and $q(T|X) \in \mathrm{IB}(\lambda)$, we have the factorisation $q(T|X) = \bar{q}(T|\mathcal{X}_J) \circ \pi_{\mathcal{X}}$, with $\bar{q}(T|\mathcal{X}_J)$ defined in (10). Thus, for all $\sigma \in \mathrm{Bij}(\mathcal{X})$,

$$\begin{aligned} \sigma \in G_{\mathrm{ci}} \quad &\Leftrightarrow \quad \forall x \in \mathcal{X}, \quad p(Y|x) = p(Y|\sigma \cdot x) \\ &\Leftrightarrow \quad \forall x \in \mathcal{X}, \quad x \sim_{\mathcal{X}} \sigma \cdot x \\ &\Leftrightarrow \quad \pi \circ \sigma = \pi \\ &\Rightarrow \quad \bar{q}(T|\mathcal{X}_J) \circ \pi \circ \sigma = \bar{q}(T|\mathcal{X}_J) \circ \pi \\ &\Leftrightarrow \quad q(T|X) \circ \sigma = q(T|X), \end{aligned} \tag{24}$$

which yields point (iii) of Theorem 2. Moreover, if we assume that $\lambda = \Lambda$, then from Lemma 11, here $\bar{q}(T|\mathcal{X}_J)$ is a congruent channel, i.e., there exists a function f such that $f \circ \bar{q}(T|\mathcal{X}_J)$ is the identity on $\bar{\mathcal{X}}$. Therefore the only implication in (24) becomes an equivalence as well, which yields point (ii) of Theorem 2.

Let us now prove point (iv) in Theorem 2. The statement is equivalent to proving that the equivalence relation defined by the partition in orbits under G_{ci}, which we denote here by $\sim_{\mathrm{ci}}$, coincides with the equivalence relation $\sim_{\mathcal{X}}$ defined in (2). Moreover, by definition of an orbit, $x \sim_{\mathrm{ci}} x'$ means that there exitst $\sigma \in \mathrm{Bij}(\mathcal{X})$ such that (i) $\sigma \in G_{\mathrm{ci}}$, i.e., $p(Y|\sigma \cdot x'') = p(Y|x'')$ for all $x'' \in \mathcal{X}$, and (ii) $x' = \sigma \cdot x$.

Thus $x \sim_{\mathrm{ci}} x'$ clearly implies $p(Y|x) = p(Y|x')$, i.e., $x \sim_{\mathcal{X}} x'$. Conversely, let us fix $x, x' \in \mathcal{X}$ such that $x \sim_{\mathcal{X}} x'$. We define σ as the transposition that permutes x and x', and fixes all the other elements of $\mathcal{X}$. It is straightforward to verify that σ satisfies points (i) and (ii) above, i.e., that we have $x \sim_{\mathrm{ci}} x'$.

Appendix C. Appendix for Section 3

C.1. On the projection on the exponential family

Let us recall that $\mathrm{cl}\,\mathcal{E}$ denotes the topological closure of the exponential family $\mathcal{E}$. Here, we denote by $\tilde{p} \in \mathrm{cl}\,\mathcal{E}$ the unique distribution (Ay et al., 2017) which achieves the minimum in $\inf_{r \in \mathrm{cl}\,\mathcal{E}} D(p||r)$; but we do not assume, a priori, that $\tilde{p}$ minimises $\inf_{r \in \mathrm{cl}\,\mathcal{E}} D(\kappa \cdot p||\kappa \cdot r)$. Note that we always have $\mathrm{supp}(p) \subseteq \mathrm{supp}(\tilde{p})$, because otherwise $D(p||\tilde{p}) = +\infty > \inf_{r \in \mathcal{E}} D(p||r)$. In particular, whenever $p(A)$ is full support, then $\tilde{p}(A)$ is full support as well. In the latter case, $\tilde{p}$ is thus both in $\mathrm{cl}\,\mathcal{E}$ and the interior of the simplex $\Delta_{\mathcal{A}}$, which implies that $\tilde{p} \in \mathcal{E}$.

Let us now prove that $\tilde{p}$ also minimises the latent space divergence, i.e., for all $\kappa \in C(\mathcal{A}, \mathcal{T})$, we automatically have $D(p||\tilde{p}) = \min_{r\in\mathcal{E}} D(\kappa \cdot p||\kappa \cdot \mathcal{E})$. Indeed, for all $r \in \mathcal{E}$ and with the convention $0 \log(\frac{0}{0}) := 0$,

$$
\begin{aligned}
D(\kappa \cdot p||\kappa \cdot \tilde{p}) - D(\kappa \cdot p||\kappa \cdot r) &= \sum_{t\in\mathcal{T}} (\kappa \cdot p)(t) \log\left(\frac{(\kappa \cdot r)(t)}{(\kappa \cdot \tilde{p})(t)}\right) \\
&= \sum_{t\in\mathcal{T}} \sum_{a\in\mathcal{A}} p(a)q(t|a) \log\left(\frac{\sum_{a\in\mathcal{A}} r(a)q(t|a)}{\sum_{a\in\mathcal{A}} \tilde{p}(a)q(t|a)}\right) \\
&\leq \sum_{t\in\mathcal{T}} \sum_{a\in\mathcal{A}} p(a)q(t|a) \log\left(\frac{r(a)q(t|a)}{\tilde{p}(a)q(t|a)}\right) \qquad (25) \\
&= \sum_{a\in\mathcal{A}} p(a) \log\left(\frac{r(a)}{\tilde{p}(a)}\right) \\
&= D(p||\tilde{p}) - D(p||r) \\
&\leq 0, \qquad (26)
\end{aligned}
$$

where line (25) uses the log-sum inequality (Csiszár and Körner, 2011), and line (26) that $D(p||\tilde{p}) = D(p||\mathcal{E})$. In particular, $\tilde{p}$ is the unique distribution in $\operatorname{cl}\mathcal{E}$ minimising both $\inf_{r\in\operatorname{cl}\mathcal{E}} D(p||r)$ and $\inf_{r\in\operatorname{cl}\mathcal{E}} D(\kappa \cdot p||\kappa \cdot r)$.

C.2. Proof of Theorem 3

We will use the notations and definitions from Section 3.1 and Appendix A, which are consistent. We also write $q(T, A)$, $q_\pi(A, T)$, $\tilde{q}(T, A)$ $\tilde{q}_\pi(T, A)$ the joint distributions defined resp. by $q(t, a) := p(a)q(t|a)$, $q_\pi(a, t) := p(a)q_\pi(t|a)$, $\tilde{q}(t, a) = \tilde{p}(a)q(t|a)$ and $\tilde{q}_\pi(t, a) = \tilde{p}(a)q_\pi(t|a)$. Note that for $\kappa = q(T|A)$, the quantity $D(\kappa\cdot p||\kappa\cdot\tilde{p})$ then becomes $D(q(T)||\tilde{q}(T))$, which will be a more convenient notation in the proofs below.

C.2.1. Factorisation for all parameter λ

Proposition 13 *For all $0 \leq \lambda \leq \Lambda$, every solution $q(T|A) \in dIB(\lambda)$ factorises as $q(T|A) = \bar{q}(T|\mathcal{A}_J) \circ \pi$.*

The proof will derive from the following lemma:

Lemma 14 *For every $q(T|A) \in C(\mathcal{A}, \mathcal{T})$, we have:*

(i) $q_\pi(T) = q(T)$ and $\tilde{q}_\pi(T) = \tilde{q}(T)$. In particular, $D(q_\pi(T)||\tilde{q}_\pi(T)) = D(q(T)||\tilde{q}(T))$.

(ii) $I_{q_\pi}(A; T) \leq I_q(A; T)$, where equality holds if and only if $q_\pi(T|A) = q(T|A)$.

Proof (i). $q_\pi(T) = q(T)$ is point (i) in Lemma 6. Moreover, for all $t \in \mathcal{T}$,

$$\begin{aligned}
\tilde{q}_\pi(t) &= \sum_{a\in\mathcal{A}} q_\pi(t|a)\tilde{p}(a) \\
&= \sum_j \sum_{a\in\mathcal{A}_j} q_\pi(t|a)\tilde{p}(a) = \sum_j \sum_{a\in\mathcal{A}_j} \frac{q(t,\mathcal{A}_j)}{p(\mathcal{A}_j)}\tilde{p}(a) \\
&= \sum_j \sum_{a,a'\in\mathcal{A}_j} \frac{1}{p(\mathcal{A}_j)} q(t|a')p(a')\tilde{p}(a) \\
&= \sum_j \sum_{a,a'\in\mathcal{A}_j} \frac{1}{p(\mathcal{A}_j)} q(t|a')p(a)\tilde{p}(a') \qquad (27)\\
&= \sum_j \sum_{a'\in\mathcal{A}_j} q(t|a')\tilde{p}(a') = \sum_{a'\in\mathcal{A}} q(t|a')\tilde{p}(a') \\
&= \tilde{q}(t)
\end{aligned}$$

where (27) uses the definition of the sets $\mathcal{A}_j$ through the equivalence relation $\sim$ defined in (6), i.e., $p(a)\tilde{p}(a') = p(a)\tilde{p}(a')$ for $a, a' \in \mathcal{A}_j$.

(ii). We apply point (ii) in Lemma 6. ■

But Lemma 14 means that for the dIB problem (3), if we replace the channel $q(T|A)$ by the corresponding $q_\pi(T|A)$, then (i) the value for of the constraint function is unchanged, and (ii) the value of the target function does not increase, with equality if and only if $q_\pi(T|A) = q(T|A)$. In particular, if $q(T|A)$ solves the dIB problem, then we must have $q(T|A) = q_\pi(T|A)$: i.e., $q(T|A) = \bar{q}(T|\mathcal{A}_J) \circ \pi$.

C.2.2. Explicit form of solutions for $\lambda = \Lambda$ (Theorem 3)

In this section, we prove Theorem 3, i.e., that $\mathrm{dIB}(\Lambda) = \left\{\gamma \circ \pi : \quad \gamma \in C_{\mathrm{cong}}(\bar{\mathcal{A}}, \mathcal{T})\right\}$. Recall that $\mathcal{A}_t^q$ is the "probabilistic pre-image of t through $q(T|A)$" (see equation (13)).

Lemma 15 *We have $D(q(T)||\tilde{q}(T)) \le \Lambda := D(p(A)||\tilde{p}(A))$, and the following are equivalent:*

(i) $D(q(T)||\tilde{q}(T)) = \Lambda$.

(ii) For all $t \in \mathrm{supp}(q(T))$, there exists a partition element $\mathcal{A}_j \in \bar{\mathcal{A}}$ such that $\mathcal{A}_t^q \subseteq \mathcal{A}_j$.

(iii) The channel $\bar{q}(T|\mathcal{A}_J) \in C(\bar{\mathcal{A}}, \mathcal{T})$ defined in (10) is congruent.

Proof $(i) \Leftrightarrow (ii)$. We have

$$D(q(T)||\tilde{q}(T)) = \sum_{t\in\mathrm{supp}(q(T))} \left(\sum_{a\in\mathcal{S}} q(t|a)p(a)\right) \log\left(\frac{\sum_{a\in\mathcal{S}} q(t|a)p(a)}{\sum_{a\in\mathcal{S}} q(t|a)\tilde{p}(a)}\right),$$

But for all $t \in \text{supp}(q(T))$, from the log-sum inequality (Csiszár and Körner, 2011), with the convention $0 \log(\frac{0}{0}) := 0$,

$$\left(\sum_{a \in \mathcal{S}} q(t|a)p(a)\right) \log\left(\frac{\sum_{a \in \mathcal{S}} q(t|a)p(a)}{\sum_{a \in \mathcal{S}} q(t|a)\tilde{p}(a)}\right) \leq \sum_{a \in \mathcal{S}} q(t|a)p(a) \log\left(\frac{q(t|a)p(a)}{q(t|a)\tilde{p}(a)}\right). \tag{28}$$

So that, summing over t, we get $D(q(T)||\tilde{q}(T)) \leq D(p(A)||\tilde{p}(A))$, with equality if and only if for all $t \in \mathcal{T}$, it holds in (28). From the equality case of the log-sum inequality (Csiszár and Körner, 2011), the latter is equivalent to the existence of nonzero constants $(\alpha_t)_{t \in \text{supp}(q(T))}$ such that

$$\forall a \in \mathcal{A} \quad q(t|a)p(a) = \alpha_t q(t|a)\tilde{p}(a),$$

i.e., such that, for all $t \in \text{supp}(q(T))$, we have $p(a) = \alpha_t \tilde{p}(a)$ for all $a \in \mathcal{A}$ such that $q(t|a) > 0$, i.e.,

$$\forall t \in \text{supp}(q(T)),\ \forall a \in \mathcal{A}_t^q,\ \frac{p(a)}{\tilde{p}(a)} = \alpha_t$$

In the above, note that the fraction $\frac{p(a)}{\tilde{p}(a)}$ does make sense, because here $\text{supp}(\tilde{p}) = \mathcal{A}$ (see Section C.1). Thus we proved that equality holds in (28) if and only if for all $t \in \text{supp}(q(T))$, the quotient $p(a)/\tilde{p}(a)$ is constant on $\mathcal{A}_t^q$, i.e., if and only if for all $t \in \text{supp}(q(T))$, there exists an $\mathcal{A}_j$ such that $\mathcal{A}_t^q \subseteq \mathcal{A}_j$.

$(ii) \Leftrightarrow (iii)$. Apply Lemma 7. ■

Combining the previous results directly yields that

$$\text{dIB}(\Lambda) \subseteq E := \left\{\gamma \circ \pi, \quad \gamma \in C_{\text{cong}}(\bar{\mathcal{A}}, \mathcal{T})\right\}.$$

Indeed, fix a solution $q(T|A) \in \text{dIB}(\Lambda)$. Proposition 13 proves that $q(T|A) = \bar{q}(T|\mathcal{A}_J) \circ \pi$. But because we must have $D(q(T)||\tilde{q}(T)) = \Lambda := D(p(A)||\tilde{p}(A))$, Lemma 15 yields that $\bar{q}(T|\mathcal{A}_J)$ is here congruent.

Let us now prove the converse inclusion, i.e., that $E \subseteq \text{dIB}(\Lambda)$.

Lemma 16 *For all $q(T|A) \in E$, we have $D(q(T)||\tilde{q}(T)) = D(p(A)||\tilde{p}(A))$ and $I_q(A;T) = H(\pi(A))$.*

Proof From point (i) in Lemma 8, $q(T|A) = \gamma \circ \pi$ implies $\gamma = \bar{q}(T|\mathcal{A}_J)$. But by definition of E, here γ is assumed congruent, thus so is $\bar{q}(T|\mathcal{A}_J)$. So that from Lemma 15, $D(q(T)||\tilde{q}(T)) = \Lambda = D(p(A)||\tilde{p}(A))$. Point (ii) in Lemma 8 yields $I_q(A;T) = H(\pi(A))$. ■

Now, because the dIB problem is defined as the minimisation of a continuous function on a compact domain, it has at least one solution, say $q_*(T|A)$, which we know belongs to E from the the inclusion $\text{dIB}(\Lambda) \subseteq E$ that we already proved. But Lemma 16 then implies that for all $q(T|A) \in E$, we have $D(q(T)||\tilde{q}(T)) = D(q_*(T)||\tilde{q}_*(T))$ and $I_q(A;T) = I_{q_*}(A;T)$. Thus any $q(T|A) \in E$ must also be a solution, i.e., $q(T|A) \in \text{dIB}(\Lambda)$. This ends the proof of point (i) in Theorem 3.

C.3. Geometric complexity perspective on $\mathbf{dIB_{ce}}$

Let us recall the definition of the exponential family $\mathcal{E}_{\text{ce}}$ from Section 3.2:

$$\mathcal{E}_{\text{ce}} := \Delta_{\mathcal{X}} \otimes \{\mathcal{U}(\mathcal{Y})\} := \{r(X)\mathcal{U}(\mathcal{Y}),\ r(X) \in \Delta_{\mathcal{X}}\},$$

It is straightforward to verify that

$$\mathcal{E}_{\text{ce}} = \{p(X,Y) \in \Delta_{\mathcal{X}\times\mathcal{Y}} :\ \forall x \in \mathcal{X}, \forall y, y' \in \mathcal{Y},\ p(x,y) = p(x,y')\},$$

i.e., $\mathcal{E}_{\text{ce}}$ is the hierarchical model of distributions on $\mathcal{X} \times \mathcal{Y}$ which actually depend only on $\mathcal{X}$. Following the information geometric approach to complexity from (Ay et al., 2011) thus leads to the interpretation that $D(p(X,Y)||\mathcal{E}_{\text{ce}})$ measures the "degree to which the system (X,Y) is more than just its part X". See Chapter 6 in (Ay et al., 2017) for a concise presentation of this approach to complexity, where point (2) of Example 6.1 in this reference corresponds to what we refer to as $\mathcal{E}_{\text{ce}}$.

C.4. The Divergence IB captures equivariances (proof of Theorem 4)

(i). The proof is almost identical to that of point (ii) of Theorem 2 (see Appendix B.3). Here $\text{dIB}(\Lambda)$ denotes the solutions to the specific dIB problem defined in Section 3.2, and π the corresponding projection defined by

$$(x,y) \sim (x',y') \quad \Leftrightarrow \quad p(x,y)p(x') = p(x',y')p(x).$$

Proposition 13 ensures that for all λ and $\kappa = q(T|A) \in \text{dIB}_{\text{ce}}(\lambda)$, we have the factorisation $\kappa = \bar{\kappa} \circ \pi$, with $\bar{\kappa} = \bar{q}(T|\mathcal{S}_J)$ defined in (10). Thus, for all $(\sigma,\tau) \in \text{Bij}(\mathcal{X}) \times \text{Bij}(\mathcal{Y})$,

$$\begin{aligned}
(\sigma,\tau) \in G_{\text{ce}} \quad &\Leftrightarrow \quad \forall (x,y) \in \mathcal{X}\times\mathcal{Y}, \quad p(y|x) = p(\tau\cdot y|\sigma\cdot x) \\
&\Leftrightarrow \quad \forall (x,y) \in \mathcal{X}\times\mathcal{Y}, \quad p(x,y)p(\sigma\cdot x) = p(\sigma\cdot x, \tau\cdot y)p(x) \\
&\Leftrightarrow \quad \forall (x,y) \in \mathcal{X}\times\mathcal{Y}, \quad (x,y) \sim (\sigma\cdot x, \tau\cdot y) \\
&\Leftrightarrow \quad \pi\circ(\sigma,\tau) = \pi \\
&\Rightarrow \quad \bar{\kappa}\circ\pi\circ(\sigma,\tau) = \bar{\kappa}\circ\pi \\
&\Leftrightarrow \quad \kappa\circ(\sigma,\tau) = \kappa,
\end{aligned} \tag{29}$$

where the first equivalence follows easily from the definition of equivariance (see Lemma 15 in (Charvin et al., 2023) for details). This yields point (ii). Moreover, if we assume that $\lambda = \Lambda$, then from Lemma 15, here $\bar{\kappa}$ is a congruent channel, i.e., there exists a function f such that $f \circ \bar{\kappa}$ is the identity on $\bar{\mathcal{A}}$. Thus the only implication in (29) becomes an equivalence as well, which yields point (i) of Theorem 4.

(iii). Here, the reasoning used for the proof of point (iii) in Theorem 2 does not work. Indeed the transposition $\Phi \in \text{Bij}(\mathcal{X}\times\mathcal{Y})$ that permutes two pairs (x,y) and (x',y') and fixes all the other ones does not have a split form $\sigma\otimes\tau$ for some $(\sigma,\tau) \in \text{Bij}(\mathcal{X}) \times \text{Bij}(\mathcal{Y})$.

Moreover, let $\mathcal{X} = \{1,2,3\}$ and $\mathcal{Y} = \{1,2\}$, with $p(X)$ uniform and $p(Y|X)$ defined through the row transition matrix

$$\begin{pmatrix} c & p_{12} \\ p_{21} & c \\ p_{31} & p_{32} \end{pmatrix}$$

where we choose c, p_{12}, p_{21}, p_{31}, and p_{32} pairwise distinct. It can be easily shown that this channel has no non-trivial equivariances, i.e., $G_{\mathrm{ce}} = \{e_{\mathcal{X}\times\mathcal{Y}}\}$, so that the projection on orbits π_{ce} is the identity of $\mathcal{X}\times\mathcal{Y}$. Yet the projection π defined by the the relation $\sim$ will here identify the two pairs (x, y) such that $p(y|x) = c$. Therefore $\pi_{\mathrm{ce}} \neq \pi$.

C.5. Relation to the Intertwining IB

Our work is heavily inspired from that in (Charvin et al., 2023); in this section we explicitly relate the two. The latter reference considered the Intertwining IB problem, namely,

$$\mathrm{IIB}(\lambda) \;:=\; \underset{\substack{\kappa\in C(\mathcal{X}\times\mathcal{Y},\mathcal{T}):\\ D(\kappa\cdot p(X,Y)||\kappa\cdot p(X)p(Y))\geq\lambda}}{\arg\min} \; I_\kappa(X,Y;T). \tag{30}$$

This problem is used to characterise equivariances under specific conditions: if (i) the distribution $p(X, Y)$ is discrete and full support, and (ii) $p(Y)$ is uniform, then the solution κ to (31) with $\lambda = I(X; Y)$ are such that a pair $(\sigma, \tau) \in \mathrm{Bij}(\mathcal{X}) \times \mathrm{Bij}(\mathcal{Y})$ is an equivariance if and only if $\kappa \circ (\sigma \otimes \tau) = \kappa$.

Thus, Section 3.2 in this work is an improvement on the latter result: here, we replace the Intertwining IB problem by the similar problem $\mathrm{dIB}_{\mathrm{ce}}$, from which we obtain the same characterisation as above, except that the assumption (ii) can now be dropped.

Moreover, it can readily be verified that problem (30) is a dIB problem with $\mathcal{E} = \Delta_{\mathcal{X}}\otimes\Delta_{\mathcal{Y}}$ and $C = C(\mathcal{X}\times\mathcal{Y}, \mathcal{T})$. In this sense, the present work is an extension and generalisation of the Intertwining IB framework. From this perspective, the pairs (σ, τ) such that $\kappa \circ (\sigma \otimes \tau) = \kappa$ for some κ solving (30) can be seen as a new kind of symmetries (exact of soft depending on the value of λ), which are in general distinct from equivariances, but which deserve further investigation.

C.6. The classic IB is a Divergence IB

Ref. (Charvin et al., 2023) proves that the classic IB can be formulated as an Intertwining IB with specific constraints on the shape of compression channels. More precisely, define $\mathcal{T}$ as $\mathcal{T} := \mathcal{T}_{\mathrm{IB}} \times \mathcal{Y}$ with $\mathcal{T}_{\mathrm{IB}} := \mathbb{N}$, [6] and consider the set

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

of channels that can compress the $\mathcal{X}$ coordinate but copy the $\mathcal{Y}$ coordinate. This leads to the problem

$$\mathrm{IIB}_{C_{\mathrm{IB}}}(\lambda) \;:=\; \underset{\substack{\kappa\in C_{\mathrm{IB}(X,Y)}:\\ D(\kappa(p(X,Y))||\kappa(p(X)p(Y)))=\lambda}}{\arg\min} \; I_\kappa(X,Y;T), \tag{31}$$

Then:

Proposition 17 ((Charvin et al., 2023), Prop. 5) *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 (31) if and only if $\kappa_{\mathcal{X}} = \kappa_{\mathcal{X}}(T_{IB}|X)$ solves the IB problem (1).*

6. This choice is formally equivalent to $\mathcal{T} := \mathbb{N}$, as there is a bijection between $\mathbb{N}$ and $\mathbb{N}\times\mathcal{Y}$.

In this sense, the classic IB is equivalent to the problem (31). Importantly, in can be easily verified that the latter which is a dIB with $C = C_{\mathrm{IB}(X,Y)}$ and $\mathcal{E} = \Delta_{\mathcal{X}} \otimes \Delta_{\mathcal{Y}}$.

However, for the sake of consistency with the results presented in this work, let us also prove that the classic IB is equivalent to a dIB with still $C = C_{\mathrm{IB}(X,Y)}$, but now

$$\mathcal{E} = \mathcal{E}_{\mathrm{ce}} := \{r(X)\mathcal{U}(\mathcal{Y}),\ r(X) \in \Delta_{\mathcal{X}}\},$$

which is the exponential family used in Section 3.2 to fully characterise channel equivariances.

As mentioned in Section 3.2, we have $D(p(X,Y)||\mathcal{E}_{\mathrm{ce}}) = D(p(X,Y)||p(X)\mathcal{U}(\mathcal{Y}))$. Moreover for $\kappa = \kappa_{\mathcal{X}} \otimes e_{\mathcal{Y}} \in C_{\mathrm{IB}}$, we have $\kappa \cdot p(X,Y) = q(T,Y)$, while $\kappa \cdot p(X)\mathcal{U}(\mathcal{Y}) = q(T)\mathcal{U}(\mathcal{Y})$ and $\kappa \cdot p(X)p(Y) = q(T)p(Y)$, where the joint distribution $q(T,X,Y)$ is defined using $\kappa = q(T|X)$ together with $p(X,Y)$ and the Markov chain $T - X - Y$. Thus

$$\begin{aligned} D(\kappa \cdot p(X,Y)||\kappa \cdot \mathcal{E}_{\mathrm{ce}}) &= D(q(T,Y)||q(T)\mathcal{U}(\mathcal{Y})) \\ &= D(q(T,Y)||q(T)p(Y)) + D(p(Y)||\mathcal{U}(\mathcal{Y})) \\ &= D(\kappa(p(X,Y))||\kappa(p(X)p(Y))) + D(p(Y)||\mathcal{U}(\mathcal{Y})). \end{aligned}$$

Therefore the constraint function in the dIB defined in (31), and the constraint function for the same problem but with $\mathcal{E}_{\mathrm{IIB}}$ replaced by $\mathcal{E}_{\mathrm{ce}}$, differ by a constant K that depends only on $p(Y)$, which is here fixed. In particular, the corresponding dIB problems are equivalent, in that for all $0 \le \lambda \le D(p(X,Y)||p(X)p(Y))$,

$$\mathrm{IIB}_{C_{\mathrm{IB}}}(\lambda) = \mathrm{dIB}_{\mathcal{E}_{\mathrm{ce}},C_{\mathrm{IB}}}(\lambda + K).$$

As we proved above that $\mathrm{dIB}_{\mathcal{E}_{\mathrm{IIB}},C_{\mathrm{IB}}}$ is equivalent to the classic IB, this proves that $\mathrm{dIB}_{\mathcal{E}_{\mathrm{ce}},C_{\mathrm{IB}}}$ is also equivalent to the classic IB (up to shifting the trade-off parameter λ by a constant K).

In other words, our framework captures channel invariances — which are a special case of channel equivariances with trivial action on the output space — by using the exponential family $\mathcal{E}_{\mathrm{ce}}$ that captures equivariances, and imposing the additional constraint C_{IB} of only compressing the input space but leaving the output space unchanged.

C.7. The Divergence IB captures distribution invariances (proof of Theorem 5)

The proof is almost identical to that of points (ii) and (iii) of Theorem 2 (see Appendix B.3). Here $\mathrm{dIB}(\Lambda)$ denotes the solutions to the specific dIB problem defined in Section 3.3, and π the corresponding projection defined by

$$a \sim a \quad \Leftrightarrow \quad p(a) = p(a') \tag{32}$$

Proposition 13 ensures that for all λ and $\kappa \in \mathrm{dIB}(\lambda)$, we have the factorisation $\kappa = \bar{\kappa} \circ \pi$, with $\bar{\kappa}$ defined in (10). Thus, for all $\Phi \in \mathrm{Bij}(\mathcal{A})$,

$$\begin{aligned} \Phi \in G_{\mathrm{di}} \quad &\Leftrightarrow \quad \forall a \in \mathcal{A}, \quad p(a) = p(\Phi \cdot a) \\ &\Leftrightarrow \quad \forall a \in \mathcal{A}, \quad a \sim \Phi \cdot a \\ &\Leftrightarrow \quad \pi \circ \Phi = \pi \\ &\Rightarrow \quad \bar{\kappa} \circ \pi \circ \Phi = \bar{\kappa} \circ \pi \\ &\Leftrightarrow \quad \kappa \circ \Phi = \kappa. \end{aligned} \tag{33}$$

This yields point (iii) of Theorem 3. Moreover, if we assume that $\lambda = \Lambda$, then from Lemma 15, here $\bar{\kappa}$ is a congruent channel, i.e., there exists a function f such that $f \circ \bar{\kappa}$ is the identity on $\bar{\mathcal{S}}$. Therefore the only implication in (33) becomes an equivalence as well, which yields point (i).

(iii). The statement is equivalent to proving that the equivalence relation defined by the partition in orbits under G_{di}, which we denote here by $\sim_{\mathrm{di}}$, coincides with the equivalence relation $\sim$ defined in (32). Moreover, by definition of an orbit, $a \sim_{\mathrm{di}} a'$ means that there exitst $\Phi \in G_{\mathrm{di}}$ such that (i) $\Phi \in G_{\mathrm{di}}$, i.e., $p(\Phi \cdot a'') = p(a'')$ for all $a'' \in \mathcal{A}$, and (ii) $a' = \Phi \cdot a$.

Thus $a \sim_{\mathrm{di}} a'$ clearly implies $p(a) = p(a')$, i.e., $a \sim a'$. Conversely, let us fix $a, a' \in \mathcal{A}$ such that $a \sim a'$. We define $\Phi \in \mathrm{Bij}(\mathcal{A})$ as the transposition that permutes a and a', and fixes all the other elements of $\mathcal{A}$. It is straightforward to verify that Φ satisfies points (i) and (ii) above, i.e., that we have $a \sim_{\mathrm{di}} a'$.

Appendix D. Appendix for section 4

In this appendix, the distribution $p(A)$ is allowed to not be full support, and we denote by $\mathcal{S}$ this support. In this case, there is still a unique distribution $\tilde{p}$ in the closure $\mathrm{cl}\,\mathcal{E}$ of $\mathcal{E}$ such that $D(p||\tilde{p}) = \inf_{r\in\mathcal{E}} D(p||\tilde{p})$ (Ay et al., 2017). We denote by $\tilde{\mathcal{S}}$ the support of $\tilde{p}$. Note that $\tilde{p} := \min_{r\in\mathcal{E}} D(p||r) < +\infty$ implies $\mathcal{S} \subseteq \tilde{\mathcal{S}}$. We also assume now that $\mathcal{T}$ is finite, and we define the dIB Lagrangian, on $C(\mathcal{A}, \mathcal{T})$, as

$$\mathcal{L}_\beta(q(T|A)) := I_q(A;T) - \beta D(q(T)||\tilde{q}(T)). \tag{34}$$

D.1. Minimisers on $\mathcal{S} \subseteq \mathcal{A}$ yield minimisers on $\mathcal{A}$

In this section, we reduce the minimisation of $\mathcal{L}_\beta$ on $C(\mathcal{A}, \mathcal{T})$ to a minimisation over channels defined only on the support $\mathcal{S}$ of $p = p(A)$. More precisely, we show that a minimiser of $\mathcal{L}_\beta$ can always be obtained the following way: choose a minimiser $\kappa \in C(\mathcal{S}, \mathcal{T})$ of the Lagrangian $\mathcal{L}_\beta$ restricted to $C(\mathcal{S}, \mathcal{T})$, and extend it to a channel in $C(\mathcal{A}, \mathcal{T})$ by sending $\mathcal{A} \setminus \mathcal{S}$ on a dummy symbol $t_0 \notin \mathrm{supp}(\kappa \cdot p)$. This allows us, in our numerical experiments, to use the BA algorithm described in Section D.2 below to find solutions in $C(\mathcal{S}, \mathcal{T})$, and then extend them to $C(\mathcal{A}, \mathcal{S})$ as described above.

Let $q(T|A) \in C(\mathcal{A}, \mathcal{T})$. We write $q_{\mathcal{S}}(T|A) \in C(\mathcal{S}, \mathcal{T})$ and $p_{\mathcal{S}}(A)$ the restrictions of $q(T|A)$, resp. $p(A)$, to $\mathcal{S} \subseteq \mathcal{A}$: i.e., $q_{\mathcal{S}}(t|a) := q(t|a)$ and $p_{\mathcal{S}}(a) := p(a)$ for all $a \in \mathcal{S}$, $t \in \mathcal{T}$ — note that these are abuses of notation, as the input alphabet of $q_{\mathcal{S}}(T|A)$ is actually only $\mathcal{S}$, and similarly $p_{\mathcal{S}}(A)$ is only defined on $\mathcal{S}$. Of course $p_{\mathcal{S}}$ is a probability on $\mathcal{S} := \mathrm{supp}(p(A))$. We extend all the notations relating to $q(T|A)$ in Section C to $q_{\mathcal{S}}(T|A)$; in particular, for $a \in \mathcal{S}, t \in \mathcal{T}$,

$$q_{\mathcal{S}}(t) := \sum_{a\in\mathcal{S}} q_{\mathcal{S}}(t,a) := \sum_{a\in\mathcal{S}} p_{\mathcal{S}}(a) q_{\mathcal{S}}(t|a) = \sum_{a\in\mathcal{S}} p(a) q(t|a) = q(t), \tag{35}$$

$$\tilde{q}_{\mathcal{S}}(t) := \sum_{a\in\mathcal{S}} \tilde{q}_{\mathcal{S}}(t,a) := \sum_{a\in\mathcal{S}} \tilde{p}_{\mathcal{S}}(a) q_{\mathcal{S}}(t|a) := \sum_{a\in\mathcal{S}} \tilde{p}(a) q(t|a),$$

or

$$
\begin{aligned}
\mathcal{L}_{\beta,\mathcal{S}}(q_{\mathcal{S}}(T|A)) &:= I_{q_{\mathcal{S}}}(A;T) - \beta D(q_{\mathcal{S}}(T)||\tilde{q}_{\mathcal{S}}(T)) \\
&:= \sum_{a\in\mathcal{S},t\in\mathrm{supp}(q_{\mathcal{S}}(T))} p_{\mathcal{S}}(a)q_{\mathcal{S}}(t|a)\log\left(\frac{q_{\mathcal{S}}(t|a)}{q_{\mathcal{S}}(t)}\right) - \beta \sum_{t\in\mathrm{supp}(q_{\mathcal{S}}(T))} q_{\mathcal{S}}(t)\log\left(\frac{q_{\mathcal{S}}(t)}{\tilde{q}_{\mathcal{S}}(t)}\right).
\end{aligned}
\tag{36}
$$

We also denote by $\tilde{\mathcal{S}} \subseteq \mathcal{A}$ the support of $\tilde{p} = \tilde{p}(A)$.

Proposition 18 *Let $q(T|A) \in C(\mathcal{A}, \mathcal{T})$. Then $q(T|A)$ is a global minimum of $\mathcal{L}_\beta$ if and only If*

(i) $q_{\mathcal{S}}(T|A)$ is a global minimum of $\mathcal{L}_{\beta,\mathcal{S}}$,

(ii) For all $t \in \mathrm{supp}(q(T))$ and $a \in \tilde{\mathcal{S}} \setminus \mathcal{S}$, we have $q(t|a) = 0$.

In particular, if $q_{\mathcal{S}}(T|A)$ is a global minimum $\mathcal{L}_{\beta,\mathcal{S}}$, we obtain a global minimum of $\mathcal{L}_\beta$ with the extension $q'(T|A) \in C(\mathcal{A}, \mathcal{T})$ of $q_{\mathcal{S}}(T|A)$ defined through

$$
q'(T|a) := \begin{cases} q_{\mathcal{S}}(T|a) & \textit{if } a \in \mathcal{S}, \\ \delta_{t_0} & \textit{if } a \in \mathcal{A}\setminus\mathcal{S}, \end{cases} \tag{37}
$$

where we chose $t_0 \in \mathcal{T} \setminus \mathrm{supp}(q(T))$.

Before proving this result, let us recall that $q(t) = \sum_{a\in\mathcal{S}} p(a)q(t|a)$, so that $\mathrm{supp}(q(T)) = \mathrm{supp}(q_{\mathcal{S}}(T))$ can be seen as the "probabilistic image of $\mathcal{S}$ through the channel $q(T|A)$", and does not depend on the values of $q(t|a)$ for $a \in \tilde{\mathcal{S}} \setminus \mathcal{S}$. Thus the condition (ii) in Proposition 18 means that $q(T|A)$ sends the elements of $\mathcal{S}$ and $\tilde{\mathcal{S}} \setminus \mathcal{S}$ on distinct subsets of bottleneck symbols in $\mathcal{T}$. Moreover, intuitively, the channel $q'(T|A)$ extends $q_{\mathcal{S}}(T|A)$ by sending all the elements a outside $\mathcal{S}$ on a "dummy" symbol t_0 which lies outside the image $\mathrm{supp}(q_{\mathcal{S}}(T))$ of $\mathcal{S}$ through $q_{\mathcal{S}}(T|A)$.

Proof We have

$$
\begin{aligned}
D(q(T)||\tilde{q}(T)) &= \sum_{t\in\mathrm{supp}(q(T))} q(t)\log\left(\frac{q(t)}{\tilde{q}(t)}\right)\\
&= \sum_{t\in\mathrm{supp}(q(T)),a\in\mathcal{A}} q(t|a)p(a)\log\left(\frac{\sum_{a\in\mathcal{A}} q(t|a)p(a)}{\sum_{a\in\mathcal{A}} q(t|a)\tilde{p}(a)}\right)\\
&= \sum_{t\in\mathrm{supp}(q(T)),a\in\mathcal{S}} q(t|a)p(a)\log\left(\frac{\sum_{a\in\mathcal{S}} q(t|a)p(a)}{\sum_{a\in\mathcal{S}} q(t|a)\tilde{p}(a)+\sum_{a\in\tilde{\mathcal{S}}\setminus\mathcal{S}} q(t|a)\tilde{p}(a)}\right)\\
&\le \sum_{t\in\mathrm{supp}(q(T)),a\in\mathcal{S}} q(t|a)p(a)\log\left(\frac{\sum_{a\in\mathcal{S}} q(t|a)p(a)}{\sum_{a\in\mathcal{S}} q(t|a)\tilde{p}(a)}\right) \qquad (38)\\
&= \sum_{t\in\mathrm{supp}(q(T)),a\in\mathcal{S}} q'(t|a)p(a)\log\left(\frac{\sum_{a\in\mathcal{S}} q'(t|a)p(a)}{\sum_{a\in\mathcal{S}} q'(t|a)\tilde{p}(a)}\right)\\
&= \sum_{t\in\mathrm{supp}(q(T)),a\in\mathcal{A}} q'(t|a)p(a)\log\left(\frac{\sum_{a\in\mathcal{A}} q'(t|a)p(a)}{\sum_{a\in\mathcal{A}} q'(t|a)\tilde{p}(a)}\right) \qquad (39)\\
&= \sum_{t\in\mathcal{T},a\in\mathcal{A}} q'(t|a)p(a)\log\left(\frac{\sum_{a\in\mathcal{A}} q'(t|a)p(a)}{\sum_{a\in\mathcal{A}} q'(t|a)\tilde{p}(a)}\right) \qquad (40)\\
&= D(q'(T)||\tilde{q}'(T)),
\end{aligned}
$$

where we defined with the marginals $q'(T) := \sum_{t\in\mathcal{T}} q'(t|a)p(a)$ and $\tilde{q}'(T) := \sum_{t\in\mathcal{T}} q'(t|a)\tilde{p}(a)$. Note that (39) uses $q'(t|a) = 0$ for $a \in \mathcal{A}\setminus\mathcal{S}$, $t \in \mathrm{supp}(q(T))$, and (40) uses

$$
\begin{aligned}
\sum_{a\in\mathcal{A}} q'(t_0|a)p(a)\log\left(\frac{\sum_{a\in\mathcal{A}} q'(t_0|a)p(a)}{\sum_{a\in\mathcal{A}} q'(t_0|a)\tilde{p}(a)}\right) &= \sum_{a\in\mathcal{A}\setminus\mathcal{S}} q'(t_0|a)\times 0\log\left(\frac{\sum_{a\in\mathcal{A}\setminus\mathcal{S}} q'(t_0|a)\times 0}{\sum_{a\in\mathcal{A}} q'(t_0|a)\tilde{p}(a)}\right)\\
&= 0.
\end{aligned}
$$

Moreover, the r.h.s. of (38) coincides with $D(q_{\mathcal{S}}(T)||q_{\mathcal{S}}(T))$. On the other hand it is straightforward to verify that $I_q(A;T) = I_{q'}(A;T) = I_{q_{\mathcal{S}}}(A;T)$. Thus

$$
\mathcal{L}_\beta(q(T|A)) \ge \mathcal{L}_\beta(q'(T|A)) = \mathcal{L}_{\beta,\mathcal{S}}(q_{\mathcal{S}}(T|A)), \qquad (41)
$$

and equality is achieved in (41) if and only if it is achieved in (38). The latter is equivalent to $\sum_{a\in\tilde{\mathcal{S}}\setminus\mathcal{S}} q(t|a)\tilde{p}(a) = 0$ for all $t \in \mathrm{supp}(q(T))$, i.e., to $q(t|a) = 0$ for all $t \in \mathrm{supp}(q(T))$ and $a \in \tilde{\mathcal{S}}\setminus\mathcal{S}$, i.e., to point (ii) in Proposition 18.

Assume now that $q(T|A)$ minimises $\mathcal{L}_\beta$. Then equation (41) and its equality case clearly imply point (ii) in Proposition 18. Moreover, if $q_{\mathcal{S},1}(T|A)$ is another channel in $C(\mathcal{S},\mathcal{T})$, we can extend it to a channel $q'_1(T|A)$ similarly as in (37), which yields

$$
\mathcal{L}_{\beta,\mathcal{S}}(q_{\mathcal{S},1}(T|A)) = \mathcal{L}_\beta(q'_1(T|A)) \ge \mathcal{L}_\beta(q(T|A)) = \mathcal{L}_{\beta,\mathcal{S}}(q_{\mathcal{S}}(T|A)),
$$

whence point (i) in Proposition 18.

Conversely, assume that points (i) and (ii) hold. Fix an arbitrary distribution $q_1(T|A) \in C(\mathcal{A}, \mathcal{T})$ and write $q_{\mathcal{S},1}(T|A) \in C(\mathcal{S}, \mathcal{T})$, resp. $q_1'(T|A) \in C(\mathcal{A}, \mathcal{T})$, the restriction of $q_1(T|A)$ to $\mathcal{S}$, resp. the corresponding channel defnied similarly as in (37). Then

$$\begin{aligned}\mathcal{L}_\beta(q_1(T|A)) &\geq \mathcal{L}_\beta(q_1'(T|A)) = \mathcal{L}_{\beta,\mathcal{S}}(q_{\mathcal{S},1}(T|A)) \\ &\geq \mathcal{L}_{\beta,\mathcal{S}}(q_{\mathcal{S}}(T|A)) = \mathcal{L}_\beta(q'(T|A)) \\ &\geq \mathcal{L}_\beta(q(T|A)),\end{aligned}$$

so that $q(T|A)$ is indeed a global minimum of $\mathcal{L}_\beta$. ■

D.2. Self-consistent equation and Blahut-Arimoto algorithm

Here we describe a Blahut-Arimoto (BA) iterative algorithm to compute the minimisers of the dIB Lagrangian (34). Following Proposition 18, we aim at a minimiser $q_{\mathcal{S}}(T|A)$ of the Lagrangian $\mathcal{L}_{\beta,\mathcal{S}}$ restricted to $\mathcal{S} := \operatorname{supp}(p(A))$ (see equation (36)), and will then extend it to the $q'(T|A)$ defined in (37). To alleviate notations, in this section we will directly write $q(T|A)$ instead of $q_{\mathcal{S}}(T|A)$. As we will see, our algorithm does not provably converge to a global minimum of the dIB Lagrangian, but it has the same guarantees as the BA algorithm for the classic IB (Tishby et al., 2000): namely, the values of the Lagrangian decrease at each step and converge to a fixed value, and the limit of a corresponding convergent sequence $(\kappa_i)_{i\in\mathbb{N}}$ must satisfy equation (9).

D.2.1. Critical points are characterised by a self-consistent equation

Taking into account the constraints $\sum_{t\in\mathcal{T}} q(t|a) = 1$ for all $a \in \mathcal{S}$, but not the inequality constraints $q(t|a) \geq 0$ for all $a \in \mathcal{S}, t \in \mathcal{T}$, we obtain the extended Lagrangian

$$\mathcal{L}_{\beta,\mu}(q(T|A)) := I_q(A;T) - \beta D(q(T)||\tilde{q}(T)) + \sum_{a\in\mathcal{S},t\in\mathcal{T}} \mu_a q(t|a). \tag{42}$$

We derive $\mathcal{L}_{\beta,\mu}$ on the open set

$$\mathcal{Q}_+ := \{(q(t|a))_{a\in\mathcal{S},t\in\mathcal{T}} \,:\, \forall a \in \mathcal{S}, \forall t \in \mathcal{T},\, q(t|a) > 0\} = (\mathbb{R}_+)^{|\mathcal{S}||\mathcal{T}|}.$$

First, $q(t) := \sum_{x'} p(a')q(t|a')$ and $\tilde{q}(t) := \sum_{x'} \tilde{p}(a')q(t|a')$, so that

$$\begin{aligned}\partial_{q(t|a)} q(t) &= p(a), \\ \partial_{q(t|a)} \tilde{q}(t) &= \tilde{p}(a).\end{aligned}$$

Moreover, note that $q(T)$ and $\tilde{q}(T)$ are strictly positive for $(q(t|a))_{a,t} \in \mathcal{Q}_+$. Thus we can write

$$\begin{aligned}
\partial_{q(t|a)} I_q(A;T) &= \partial_{q(t|a)} \sum_{a',t'} p(a') q(t'|a') \log\left(\frac{q(t|a')}{q(t)}\right) \\
&= p(a) \log\left(\frac{q(t|a)}{q(t)}\right) + \sum_{a'} p(a') q(t|a') \frac{q(t)}{q(t|a')} \frac{q(t)\delta_{a'=a} - p(a) q(t|a')}{q(t)^2} \\
&= p(a) \log\left(\frac{q(t|a)}{q(t)}\right) + \sum_{a'} p(a') \left(\delta_{a'=a} - \frac{p(a) q(t|a')}{q(t)}\right) \\
&= p(a) \log\left(\frac{q(t|a)}{q(t)}\right) + p(a) - p(a) \frac{q(t)}{q(t)} \\
&= p(a) \log\left(\frac{q(t|a)}{q(t)}\right),
\end{aligned}$$

and

$$\begin{aligned}
\partial_{q(t|a)} D(q(T)||\tilde{q}(T)) &= \partial_{q(t|a)} \sum_{a',t'} p(a') q(t'|a') \log\left(\frac{q(t)}{\tilde{q}(t)}\right) \\
&= p(a) \log\left(\frac{q(t)}{\tilde{q}(t)}\right) + \sum_{a'} p(a') q(t|a') \frac{\tilde{q}(t)}{q(t)} \frac{p(a)\tilde{q}(t) - \tilde{p}(a) q(t)}{\tilde{q}(t)^2} \\
&= p(a) \log\left(\frac{q(t)}{\tilde{q}(t)}\right) + \left(\sum_{a'} p(a') q(t|a')\right) \left(\frac{p(a)}{q(t)} - \frac{\tilde{p}(a)}{\tilde{q}(t)}\right) \\
&= p(a) \log\left(\frac{q(t)}{\tilde{q}(t)}\right) + q(t) \left(\frac{p(a)}{q(t)} - \frac{\tilde{p}(a)}{\tilde{q}(t)}\right) \\
&= p(a) \log\left(\frac{q(t)}{\tilde{q}(t)}\right) + p(a) - \frac{q(t)}{\tilde{q}(t)} \tilde{p}(a).
\end{aligned}$$

Therefore

$$\partial_{q(t|a)} \mathcal{L}_{\beta,\mu}(q(T|A)) = p(a) \log\left(\frac{q(t|a)}{q(t)}\right) \; - \; \beta \left(p(a) \log\left(\frac{q(t)}{\tilde{q}(t)}\right) + p(a) - \frac{q(t)}{\tilde{q}(t)} \tilde{p}(a)\right) + \mu_a.$$

Now, recall that here the input set of $q(T|A)$ is $\mathcal{S} = \mathrm{supp}(p(A))$. Hence, we can absorb $p(a)$ into the constant μ_a, and get that a necessary condition for local minimisers of the dIB Lagrangian $\mathcal{L}_\beta$ on $\mathcal{Q}_+$ is the existence of constants $(\mu_a)_a \in \mathbb{R}^{|\mathcal{S}|}$ such that

$$\log\left(\frac{q(t|a)}{q(t)}\right) \; - \; \beta \left(\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + 1 - \frac{q(t)\tilde{p}(a)}{\tilde{q}(t) p(a)}\right) + \mu_a = 0$$

i.e., such that

$$q(t|a) = q(t) \exp\left[\beta \left(\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + 1 - \frac{q(t)\tilde{p}(a)}{\tilde{q}(t) p(a)}\right) + \mu_a\right].$$

Thus we proved that local minimisers of the dIB Lagrangian $\mathcal{L}_\beta$ over the set of channels $q(T|A) \in C(\mathcal{S}, \mathcal{T})$ with strictly positive entries satisfy the necessary condition

$$q(t|a) = \frac{1}{Z(a,\beta)} q(t) \exp\left[-\beta\left(\frac{q(t)\tilde{p}(a)}{\tilde{q}(t)p(a)} - \log\left(\frac{q(t)\tilde{p}(a)}{\tilde{q}(t)p(a)}\right) - 1\right)\right], \tag{43}$$

where $Z(a,\beta)$ is a positive normaliser. Note that in (43), we added the factor $\frac{\tilde{p}(a)}{p(a)}$ in the logarithm. This equivalent reformulation is more suited to the implementation of the Blahut-Arimoto algorithm described below. Indeed, in this form, the expression in the exponential is always non-positive (as shown by the study of the function $x \mapsto x - \log(x) - 1$), which avoids overflow for large β.

Note that *a priori*, there might also be local minimisers of $\mathcal{L}_\beta$ on the border of $C(\mathcal{S}, \mathcal{T})$. For the sake of completeness, let us outline an argument showing that this is actually not the case. The computations above show that, deriving $\mathcal{L}_\beta(q(T|A))$ as a function on $\mathcal{Q}_+$, we get

$$\partial_{q(t|a)} \mathcal{L}_\beta(q(T|A)) = p(a)\left[\log\left(\frac{q(t|a)}{q(t)}\right) - \beta\left(\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + 1 - \frac{q(t)\tilde{p}(a)}{\tilde{q}(t)p(a)}\right)\right].$$

In particular, for $q(T|A) \in C(\mathcal{S}, \mathcal{T})$ strictly positive but with at least one coordinate approaching 0, the directional derivative w.r.t this coordinate diverges to $-\infty$. Indeed, $D(p||\tilde{p}) < +\infty$ implies $\mathcal{S} = \operatorname{supp}(p(A)) \subseteq \operatorname{supp}(\tilde{p}(A))$, while each $q(T|a)$ is a probability, with $p(A)$ and $\tilde{p}(A)$ fixed; so that there are strictly positive constants k and K such that $k \leq q(t) \leq K$ and $k \leq \tilde{q}(t) \leq K$ for all $q(T|A) \in \mathcal{Q}_+$. Thus the term $\log\left(\frac{q(t)}{\tilde{q}(t)}\right) + 1 - \frac{q(t)\tilde{p}(a)}{\tilde{q}(t)p(a)}$ remains bounded as well. But again because $q(t) \geq k$, on the other hand $\log\left(\frac{q(t|a)}{q(t)}\right)$ diverges to $-\infty$ when $q(t|a)$ goes to 0.

Using classic arguments, we can then use the divergence to $-\infty$ of the gradient close to the border, along with the continuity of $\mathcal{L}_\beta$ on the whole closed set $C(\mathcal{S}, \mathcal{T})$, to prove that $q(T|A)$ cannot be a local minimum of $\mathcal{L}_\beta$ over $C(\mathcal{S}, \mathcal{T})$ if it has a coordinate $q(t|a)$ equal to 0, i.e., if it is on the border of $C(\mathcal{S}, \mathcal{T})$.

D.2.2. Blahut-Arimoto algorithm

Here, we denote by $C_+(\mathcal{S}, \mathcal{T})$ the subset of $C(\mathcal{S}, \mathcal{T})$ made of channels with only positive entries, by $\Delta_{\mathcal{T},+}$ the open simplex of full-support probabilities on $\mathcal{T}$, and by $\mathbb{R}_+$ the positive real numbers. We define, for $q(T|A) \in C(\mathcal{S}, \mathcal{T})$, a probability $r(T) \in \Delta_{\mathcal{T}}$ on $\mathcal{T}$, and some $r(T) \in (\mathbb{R}_+)^{|\mathcal{T}|}$,

$$\begin{aligned} &F(q(T|A), r(T), m(T)) \\ &\quad := \sum_{a,t} p(a)q(t|a)\log\left(\frac{q(t|a)}{r(t)}\right) - \beta \sum_{a,t} p(a)q(t|a)\left(\log\left(m(t)\frac{\tilde{p}(a)}{p(a)}\right) - m(t)\frac{\tilde{p}(a)}{p(a)} + 1\right). \end{aligned}$$

The function F is thus defined on the open and convex set

$$\operatorname{Dom}_F := C_+(\mathcal{S}, \mathcal{T}) \times \Delta_{\mathcal{T},+} \times (\mathbb{R}_+)^{|\mathcal{T}|}.$$

The next proposition defines the Blahut-Arimoto (BA) algorithm adapted to our problem, and describes its properties.

Proposition 19 *The function F is convex in each of its coordinates. Moreover, for $q_i(T|A) \in C_+(\mathcal{S}, \mathcal{T})$, defining*

$$\begin{aligned} r_{i+1}(t) &:= \sum_a p(a) q_i(t|a), \\ m_{i+1}(t) &:= \frac{\sum_a p(a) q_i(t|a)}{\sum_a \tilde{p}(a) q_i(t|a)}, \\ q_{i+1}(t|a) &:= \frac{1}{Z(a,\beta)} r_{i+1}(t) \exp\left[-\beta\left(m_{i+1}(t)\frac{\tilde{p}(a)}{p(a)} - \log\left(m_{i+1}(t)\frac{\tilde{p}(a)}{p(a)}\right) - 1\right)\right], \end{aligned} \tag{44}$$

where $Z(a,\beta)$ is a positive normaliser, we have:

(i) All quantities in (44) are well-defined, and $(q_{i+1}(T|A), r_{i+1}(T), m_{i+1}(T)) \in \mathrm{Dom}_F$.

(ii) $F(q_i(T|A), r_i(T), m_i(T)) = \mathcal{L}_\beta(q_i(T|A)) + K$, where the Lagrangian $\mathcal{L}_\beta$ is defined in (34) and K is a constant that does not depend on i.

(iii) At each update of $q_i(T|A)$, $r_i(T)$ and $m_i(T)$, the function F is minimised w.r.t. the corresponding coordinate. In particular,

$$F(q_{i+1}(T|A), r_{i+1}(T), m_{i+1}(T)) \leq F(q_i(T|A), r_i(T), m_i(T)).$$

Before proving it, let us first draw the consequences of Proposition 19. Define some $q_0(T|A) \in C_+(\mathcal{S}, \mathcal{T})$, and the corresponding sequence $(q_i(T|A), r_i(T), m_i(T))_{i\geq 1}$ from (44). From point (i), the sequence is included in Dom_F, and from point (ii), we have, for all i,

$$\mathcal{L}_\beta(q_i(T|A)) = F((q_i(T|A), r_i(T), m_i(T))) - K.$$

From point (iii), this yields a non-increasing sequence of images $(\mathcal{L}_\beta(q_i(T|A)))_i$. As $\mathcal{L}_\beta$ is bounded from below, this implies that this sequence converges. Moreover, as the closure $\overline{C_+(\mathcal{S}, \mathcal{T})} = C(\mathcal{S}, \mathcal{T})$ of $C_+(\mathcal{S}, \mathcal{T})$ is compact, we can, up to extracting a subsequence, assume that $(q_i(T|A))_i$ converges to a point $q_*(T|A) \in C(\mathcal{S}, \mathcal{T})$.[7] From the definition of $(q_i(T|A))_i$ through (44) and from the continuity of this iterative equation, we obtain that the limit $q_*(T|A)$ satisfies the fixed-point equation (43). Hence we proved the claims made the beginning of Appendix D.2 about this BA algorithm. Note that even though F is convex in each coordinate, we did not prove that F is convex as a whole. Thus we cannot apply the classic BA arguments (Yeung, 2008) to prove that the sequence $(\mathcal{L}_\beta(q_i(T|A)))_i$ converges to a global minimum of $\mathcal{L}_\beta$. However, the statements proved here match exactly the corresponding statements proven for the BA algorithm in the classic IB case (Tishby et al., 2000).

Proof [Proof of Proposition 19]

The convexity of F in each coordinate is straightforward. Point (i) comes from the fact that $q(T|A) \in C(\mathcal{S}, \mathcal{T})$, where $\mathcal{S}$ is the support of $p(A)$, which contains that of $\tilde{p}(A)$ (because $D(p||\tilde{p}) < +\infty$). Point (ii) is a direct computation. Let us now prove point (iii).

7. In practice, in numerical implementations, we always observed the convergence of $(q_i(T|A))_i$, without any subsequence extraction.

For fixed $(r(T), m(T))$, we know that the function $F(\cdot, r(T), m(T))$ is convex on $C_+(\mathcal{S}, \mathcal{T})$, so that the minimum is achieved at points $q(T|A)$ such that $\nabla_{q(T|A)} F(q(T|A), r(T), m(T) = 0$. A direct computation shows that the latter is equivalent to, for all $a \in \mathcal{S}, t \in \mathcal{T}$,

$$q(t|a) = \frac{1}{Z(a,\beta)} r(t) \exp\left[-\beta \left(m(t)\frac{\tilde{p}(a)}{p(a)} - \log\left(m(t)\frac{\tilde{p}(a)}{p(a)}\right) - 1\right)\right], \tag{45}$$

where $Z(a, \beta)$ is a positive normaliser. Moreover, it is standard (Yeung, 2008) to prove that, for fixed $(q(T|A), m(T)) \in C_+(\mathcal{S}, \mathcal{T}) \times (\mathbb{R}_+)^{|\mathcal{T}|}$, the minimimum of F w.r.t to $r(T)$ is achieved for

$$r(T) = q(T) := \sum_a p(a) q(T|a) \tag{46}$$

Eventually, for fixed $(q(T|A), r(T)) \in C_+(\mathcal{S}, \mathcal{T}) \times \Delta_{\mathcal{T},+}$ the minimum of F w.r.t. $m(T)$ is, again by convexity, achieved if and only if the corresponding gradient vanishes. But we have

$$\begin{aligned}
\partial_{m(t)} F(q(T|A), r(T), m(T)) &= \sum_a p(a) q(t|a) \left(\frac{1}{m(t)} - \frac{\tilde{p}(a)}{p(a)}\right) \\
&= \frac{q(t)}{m(t)} - \tilde{q}(t),
\end{aligned}$$

so that the gradient w.r.t $m(T)$ cancels if and only if for all $t \in \mathcal{T}$,

$$m(t) = \frac{q(t)}{\tilde{q}(t)} = \frac{\sum_a p(a) q(t|a)}{\sum_a \tilde{p}(a) q(t|a)}. \tag{47}$$

This proves point (iii). ■

D.3. Details on effective cardinality

(Zaslavsky and Tishby, 2019) defines a concept of effective cardinality for the Lagrangian formulation of the classic IB. Here, we adapt this concept to the dIB framework in its primal formulation, i.e., problem (3), and in a way which also encompasses the case $\mathrm{supp}(p(A)) \subsetneq \mathcal{A}$. For $\kappa = q(T|A) \in C(\mathcal{A}, \mathcal{T})$, consider the "probabilistic image of $\mathcal{A}$ through κ", i.e.,

$$\kappa \cdot \mathcal{A} := \{t \in \mathcal{T} \,:\, \exists a \in \mathcal{A} \,:\, q(t|a) > 0\}.$$

Note that this definition depends only on κ and not on $p(A)$. We then define the cardinality of κ as $K(\kappa) := |\kappa \cdot \mathcal{A}|$. However, for $\kappa \in \mathrm{dIB}(\lambda)$, the number $K(\kappa)$ does not necessarily carry any meaningful information about the dIB problem itself: e.g., it can be easily verified that composing any $\kappa \in \mathrm{dIB}(\lambda)$ with a congruent channel γ (which can arbitrarily increase the cardinality) still yields a solution $\gamma \circ \kappa \in \mathrm{dIB}(\lambda)$. This motivates the definition of the minimum number of symbols t necessary to describe the output of a bottleneck encoder κ. Formally:

Definition 20 *The* effective cardinality *of a dIB solution $\kappa \in dIB(\lambda)$ is*

$$k(\kappa) := \min_{\gamma \in C(\mathcal{T}):\ \gamma\circ\kappa \in dIB(\lambda)} K(\gamma \circ \kappa),$$

i.e., it is the minimum bottleneck cardinality obtained from a post-processing of κ that still produces a dIB solution for the same parameter λ.

Let us fix an arbitrary $\kappa = q(T|A) \in \mathrm{dIB}(\lambda)$, write $q(A,T) := p(A)q(T|A)$, and assume first that $p(A)$ is full support. It can be shown, using the log-sum inequality, that $k(\kappa)$ is the cardinality of the partition $\bar{\mathcal{T}}$ of $\mathrm{supp}(q(T))$ defined by the equivalence relation $t \sim t' \Leftrightarrow q(A|t) = q(A|t')$.

For the non full support case, denote by $\tilde{p}(A)$ the unique distribution satisfying $D(p||\tilde{p}) = D(p||\mathcal{E})$ (Ay et al., 2017), and note that $D(p||\mathcal{E}) < \infty$ implies $\mathcal{S} \subseteq \tilde{\mathcal{S}}$, where $\mathcal{S} := \mathrm{supp}(p(A))$ and $\tilde{\mathcal{S}} := \mathrm{supp}(\tilde{p}(A))$. It can be easily verified that the value of $q(T|a)$ for $a \in \tilde{\mathcal{S}}^c$ affects neither the target nor the constraint function of the dIB problem (3). However, a direct consequence of Proposition 18 is that if $\tilde{\mathcal{S}} \setminus \mathcal{S} \neq \emptyset$, the image $\kappa \cdot \mathcal{A}$ of $\mathcal{A}$ must contain at least one symbol $t_0 \notin \mathrm{supp}(q(T))$, on which to send the elements of $\tilde{\mathcal{S}} \setminus \mathcal{S}$ (for clarity, let us recall that $q(t) := \sum_{a\in\mathcal{S}} p(a)q(t|a)$, so that $\mathrm{supp}(q(T))$ is the "probabilistic image of $\mathcal{S}$ through $q(T|A)$", and does not depend on $q(T|a)$ for $a \in \tilde{\mathcal{S}} \setminus \mathcal{S}$). It can be easily verified that the above implies that here the effective cardinality becomes $|\bar{\mathcal{T}}| + 1$. Note that this is the situation we encounter in our numerical experiments (Section 4.1).

We use the above to numerically compute the effective cardinality. Note that the choice of the threshold for rounding $|q(t|a) - q(t|a')|$ to 0 is here important. We choose 10^{-3}.

D.4. Computable form of $D^p(\kappa||C_G)$

Here we provide more details on the divergence introduced in Section 3.4, and prove that it can be computed directly as a divergence between two channels. Let $\mathcal{S} := \mathrm{supp}(p(A))$. For two channels κ, ν in either $C(\mathcal{A},\mathcal{T})$ or $C(\mathcal{S},\mathcal{T})$, we define their KL divergence $D^p(\kappa||\nu)$ with respect to $p = p(A)$, as (Ay et al., 2017)

$$D^p(\kappa||\nu) := \sum_{a\in\mathcal{S}} p(a) D(\kappa(T|a)||\nu(T|a)).$$

We also define, for a group G acting on $\mathcal{A}$, the set of input-symmetric channels w.r.t. G, i.e.,

$$C_G := \{\nu \in C(\mathcal{A},\mathcal{T}) : \quad \forall \Phi \in G, \quad \nu \circ \Phi = \nu\ \},$$

and the corresponding divergence of some $\kappa \in C(\mathcal{A},\mathcal{T})$ from C_G with respect to p as (Ay, 2015)

$$D^p(\kappa||C_G) := \min_{\nu\in C_G} D^p(\kappa||\nu).$$

For all purposes relevant to this article's scopes, we have $D^p(\kappa||C_G)$ if and only if $\kappa \circ \Phi = \kappa$ for all $\Phi \in G$. More precisely:

Assume first that $p(A)$ is full support. Then, from the continuity of the KL divergence and the fact that C_G is a closed subset of $C(\mathcal{A},\mathcal{T})$, we have $D(\kappa||C_G) = 0$ if and only if $\kappa \in C_G$, i.e., $\kappa \circ \Phi = \kappa$ for all $\Phi \in G$.

Let us now drop the full support assumption on $p(A)$, but assume instead that (i) the group G leaves $\mathcal{S}$ invariant, and (ii) the channel $\kappa = q(T|A)$ is as $q'(T|A)$ in equation (37), i.e., it sends $\mathcal{S}^c$ on a single symbol outside the image of $\mathcal{S}$ through κ. From point (i), the action of G on $\mathcal{A}$ induces a action on $\mathcal{S}$, and a corresponding set $C_{G_\mathcal{S}}$. Denote by $\kappa_\mathcal{S}$ the restriction of a channel $\kappa \in C(\mathcal{A}, \mathcal{T})$ to $\mathcal{S}$. Using point (ii), it can be easily verified that for all $\kappa \in C(\mathcal{A}, \mathcal{T})$, we have $\kappa \in C_G$ if and only if $\kappa_\mathcal{S} \in C_{G_\mathcal{S}}$, and that $D(\kappa||C_G) = D(\kappa_\mathcal{S}||C_{G_\mathcal{S}})$. From that we can conclude, using the full support case described above, that here we also have $D^p(\kappa||C_G)$ if and only if $\kappa \circ \Phi = \kappa$ for all $\Phi \in G$.

Note that points (i) and (ii) are satisfied in our numerical experiments in Section 4.1, and that they are also automatically satisfied if $p(A)$ is full support.

Let us now provide a form of $D^p(\kappa||C_G)$ which is easier to compute.

Proposition 21 *Fix $p(A) \in \Delta_\mathcal{A}$, a finite group G acting on $\mathcal{A}$ and leaving $\mathcal{S}$ invariant, and $\kappa = q(T|A) \in C(\mathcal{A}, \mathcal{T})$. Then*

$$D^p(\kappa||C_G) = D^p(\kappa||\kappa_G)$$

where $\kappa_G = q_G(T|A) \in C(\mathcal{S}, \mathcal{T})$ is defined through, for $a \in \mathcal{S}$ and $t \in \mathcal{T}$,

$$q_G(t|a) := q(t|[a]) := \frac{\sum_{a' \in [a]} p(a') q(t|a')}{p([a])},$$

with $[a]$ the orbit of a under G.

Intuitively, κ_G is the average of the channel κ over the group G acting on its input, computed using the distribution p on the input.

Proof It is enough to prove that for all $\nu \in C_G$,

$$D^p(\kappa||\kappa_G) \le D^p(\kappa||\nu).$$

For $a \in \mathcal{S}$, we have $[a] \subseteq \mathcal{S}$, and $q_G(T|a')$ is well-defined and constant for $a' \in [a]$. Moreover, for $\nu = r(T|A) \in C_G$, it is straightforward to verify that $r(T|a')$ is also constant for $a' \in [a]$, so that

$$\sum_{a' \in [a]} r(t|a') p(a') = r(t|a) p([a]).$$

Thus, for $\nu = r(T|A) \in C_G$, and $a_1, \ldots, a_n$ a system of representatives of the orbits included in $\mathcal{S}$,

$$\begin{aligned}
D^p(\kappa||\nu) - D^p(\kappa||\kappa_G) &= \sum_{a \in \mathcal{S}, t \in \mathcal{T}} p(a) q(t|a) \log \left(\frac{q_G(t|a)}{r(t|a)} \right) \\
&= \sum_{i=1}^n \sum_t \log \left(\frac{q_G(t|a_i)}{r(t|a_i)} \right) \sum_{a \in [a_i]} p(a) q(t|a) \\
&= \sum_{i=1}^n \sum_t \log \left(\frac{\sum_{a \in [a_i]} q(t|a) p(a)}{\sum_{a \in [a_i]} r(t|a) p(a)} \right) \sum_{a \in [a_i]} p(a) q(t|a) \\
&= D(q_1||q_2) \ge 0,
\end{aligned}$$

where q_1 and q_2 are distributions defined on $\mathcal{S}/G \times \mathcal{T}$, through

$$q_1([a_i], t) = \sum_{a \in [a_i]} p(a) q(t|a),$$
$$q_2([a_i], t) = \sum_{a \in [a_i]} p(a) t(t|a).$$

$\blacksquare$