entropy

MDPI

Article

Exact and Soft Successive Refinement of the Information Bottleneck

Hippolyte Charvin *, Nicola Catenacci Volpi and Daniel Polani

School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK; n.catenacci-volpi@herts.ac.uk (N.C.V.); d.polani@herts.ac.uk (D.P.)
* Correspondence: h.charvin@herts.ac.uk

Abstract: The information bottleneck (IB) framework formalises the essential requirement for efficient information processing systems to achieve an optimal balance between the complexity of their representation and the amount of information extracted about relevant features. However, since the representation complexity affordable by real-world systems may vary in time, the processing cost of updating the representations should also be taken into account. A crucial question is thus the extent to which adaptive systems can *leverage the information content of already existing IB-optimal representations for producing new ones*, which target the same relevant features but at a different granularity. We investigate the information-theoretic optimal limits of this process by studying and extending, within the IB framework, the notion of *successive refinement*, which describes the ideal situation where no information needs to be discarded for adapting an IB-optimal representation's granularity. Thanks in particular to a new geometric characterisation, we analytically derive the successive refinability of some specific IB problems (for binary variables, for jointly Gaussian variables, and for the relevancy variable being a deterministic function of the source variable), and provide a linear-programming-based tool to numerically investigate, in the discrete case, the successive refinement of the IB. We then soften this notion into a *quantification* of the loss of information optimality induced by several-stage processing through an existing measure of unique information. Simple numerical experiments suggest that this quantity is typically low, though not entirely negligible. These results could have important implications for (i) the structure and efficiency of incremental learning in biological and artificial agents, (ii) the comparison of IB-optimal observation channels in statistical decision problems, and (iii) the IB theory of deep neural networks.

Keywords: information bottleneck; successive refinement; unique information; incremental learning; coarse-graining; Blackwell order; deep learning

Citation: Charvin, H.; Catenacci Volpi, N.; Polani, D. Successive Refinement of the Information Bottleneck. *Entropy* **2023**, *25*, 1355. https://doi.org/10.3390/e25091355

Academic Editors: Jan Lewandowsky and Gerhard Bauch

Received: 25 July 2023
Revised: 8 September 2023
Accepted: 13 September 2023
Published: 19 September 2023

1. Introduction

1.1. Conceptualisation and Organisation Outline

Consider the problem, for an information-processing system, of extracting relevant information about a target variable Y within a correlated source variable X, under constraints on the cost of the information processing needed to do so—yielding a compressed representation T. This situation can be formalised in an information-theoretic language, where the information-processing cost is measured with the mutual information $I(X;T)$ between the source X and the representation T of it, while the relevancy about Y of the information extracted by T is measured by $I(Y;T)$. The problem thus becomes that of maximising the relevant information $I(Y;T)$ under bounded information-processing cost $I(X;T)$, i.e., we are interested in the *information bottleneck* (IB) problem [1,2], which, in primal form, can be formulated as

$$\underset{\substack{q(T|X):\\ T-X-Y,\ I(X;T)\leq\lambda}}{\arg\max} \quad I(Y;T). \tag{1}$$

Here, the trade-off parameter λ controls the bound on the permitted information-processing cost and thus, intuitively, the resulting representation's granularity. The Markov chain condition $T - X - Y$ ensures that any information that the bottleneck T extracts about the relevancy variable Y can only come from the source X. The solutions to (1) for varying λ trace the so-called *information curve*, i.e., the λ-parameterised curve

$$\big(I_\lambda(X;T), I_\lambda(Y;T)\big)_{\lambda \geq 0} \subseteq \mathbb{R}^2, \tag{2}$$

where $I_\lambda(X;T)$ and $I_\lambda(X;T)$ are defined by a bottleneck T of parameter λ (see the black curve in the first figure in Section 2 below). This curve indicates the informationally optimal bounds on the feasible trade-offs between relevancy $I(Y;T)$ and complexity $I(X;T)$ of the representation T. In this sense, the IB method provides a fundamental understanding of the *informationally optimal limits* of information-processing systems.

These limits are crucial for both understanding and building adaptive behaviour. For instance, choosing X to be an agent's past and Y to be its future leads it to extract the most relevant features of its environment [3–6]. More generally, the IB point of view on modelling embodied agents' representations has been leveraged for unifying efficient and predictive coding principles in theoretical neuroscience—at the level of single neurons [3,7–9] and neuronal populations [9–13]—but also for studying sensor evolution [14–16], the emergence of common concepts [17] and of spatial categories [18], the evolution of human language [19–21], or for implementing informationally efficient control in artificial agents [22–24]. This line of research brings increasing support to the hypothesis that, particularly for evolutionary reasons, biological agents are often poised close to optimality in the IB sense. It also provides a framework for both measuring and improving artificial agents' performance.

However, one aspect of the IB framework conflicts with a crucial feature of real-world systems: the informationally optimal limits that it describes only consider a given representation T *taken in isolation* from any other one in the system. This point of view *a priori* disregards the *relationship between representations*, which is crucial in real-world information-processing systems. Thus, it is crucial to consider the following question: does the relationship between a set of internal representations $T_1, \ldots, T_n$ impact their individual information optimality? In this paper, we are mostly interested in a specific kind of relationship: when $T_1, \ldots, T_n$ are successively produced in this order, and each new T_i builds on both the previous representation T_{i-1} and new information from the fixed source X to extract information about the fixed relevancy Y. This scenario formalises the *incorporation of information into already learned representations*—as is the case in developmental learning, or, more generally, any kind of learning process that goes through identifiable successive steps.

More precisely, consider an informationally bounded agent that extracts information about a relevant variable Y within an environment X. If the agent is informationally optimal, given an affordable complexity cost λ_1, it must maximise the relevant information that it extracts from the environment—resulting in a bottleneck representation T_1, i.e., a solution to (1) with parameter λ_1. Then, assume that at, a later stage, the complexity cost that the agent can afford increases to $\lambda_2 > \lambda_1$, while the goal is still to extract information about the same relevant feature Y within the same environment X. To keep being informationally optimal, the agent should thus update its representation so it becomes a bottleneck of parameter λ_2. Given this setting, the question we ask is: to which extent can the content learned into T_1 be leveraged for the production of T_2? It is indeed not intuitively clear that T_2 should keep all the information from T_1. An informal example is the fact that most pedagogical curricula teach knowledge via successive approximations, where, at a more advanced level, the content learned at the beginner level must sometimes be *unlearned* to successively proceed further, even though it was perfectly reasonable—in our language, informationally optimal—to deliver the first beginner sketch to students that would never progress to learn the expert level.

This question has been formalised, in the rate-distortion literature, with the notion of *successive refinement* (SR) [25–29], which, in short, refers to the situation where several-stage processing does not incur any loss of information optimality. More precisely, in the context outlined above, there is successive refinement if the processing cost of first producing a coarse bottleneck T_1 of parameter λ_1 and then refining it to a finer bottleneck T_2 of parameter $\lambda_2 > \lambda_1$ is no larger than the processing cost of directly producing a bottleneck T_2 of parameter λ_2 without any intermediary bottleneck T_1 (see Section 2.1 and Appendix B.2 for formal definitions). The aim of this work is to push the understanding of successive refinement in the IB framework [30–32] further, as well as to expand the analysis to a *quantification* of the lack of SR, in cases where the latter does not hold exactly. We start by leveraging general results in existing IB literature [33,34] to prove that successive refinement always holds for jointly Gaussian (X, Y), and when Y is a deterministic function of X. However, it is seems crucial, for further progress on more general scenarios, to design specifically tailored mathematical and numerical tools. In this regard, we provide two main contributions.

First, we present a simple geometric characterisation of SR, in terms of convex hulls of the decoder symbol-wise conditional probabilities $q(X|t)$, for t varying in the bottleneck alphabet $\mathcal{T}$. This characterisation is proven in the discrete case under an additional but mild assumption of injectivity of the decoder $q(X|T)$. This new point of view fits well with an ongoing convexity approach to the IB problem [35–39] and might thus help develop a new geometric perspective on the successive refinement of the IB. As an example, we use this geometric characterisation to prove that SR always holds for binary source X and binary relevancy Y. Moreover, this characterisation makes it straightforward to numerically assess, with a linear program checking convex hull inclusions, whether or not two discrete bottlenecks T_1 and T_2 achieve successive refinement. As we demonstrate with minimal numerical examples, this can help in investigating the SR structure of any given IB problem, i.e., how successive refinement depends on the particular combination of trade-off parameters λ_1 and λ_2.

Second, we soften [18] the traditional notion of successive refinement and study the *extent to which* several-stage processing incurs a loss of information optimality. More precisely, we propose to measure soft successive refinement with the *unique information* [40] (UI) that the coarser bottleneck T_1 holds about the source X, as compared to the finer one T_2. Explicitly, this UI is defined as the minimal value of $I_q(X; T_1|T_2)$ over all distributions $q := q(X, T_1, T_2)$ whose marginals $q(X, T_1)$ and $q(X, T_2)$ coincide with the corresponding bottleneck distributions (see Section 3.1 for details). As a first exploration of soft SR's qualitative features, we investigate the landscapes of unique information over trade-off parameters, for again some simple example distributions $p(X, Y)$. These landscapes seem to unveil a rich structure, which was largely hidden by the traditional notion of SR, that only distinguished between SR being present or absent. Among the general features suggested by these experiments, the most significant are that (i) soft SR seems strongly influenced by the trajectories of the decoders $q_\lambda(X|T)$ over λ; (ii) the UI often goes through sharp variations at the bifurcations [41–44] undergone by the bottlenecks (in a fashion compatible with the presence of discontinuities of either the UI itself, or its differential, with regard to trade-off parameters); and (iii) the loss of information optimality seems always small—more precisely, the global bound on the UI was observed to be typically one or two orders of magnitude lower than the system's globally processed information (see Section 3.2 for formal statements). These three conclusions are phenomenological and limited to our minimal examples, but they shed light on the kind of structure that can be investigated by further research. They also suggest the relevance that developing this theoretical framework might have for the scientific question that motivates it. In particular, the link with IB bifurcations and the overall small loss of information optimality would, if generalisable, have interesting consequences for the structure and efficiency of incremental learning.

As a side contribution, we draw along the paper formal equivalences between our framework and other notions proposed in the literature, thus making the formal framework also relevant to decision problems [40,45] and to the information-theoretic approach to deep learning [46]. This flexibility of interpretation stems from the fact that even though our formal framework crucially depends on the order of the bottleneck representations' trade-off parameters, it does not depend on the order in which these representations are produced. Thus, a sequence of bottlenecks can be equally well interpreted as produced from coarsest to finest—as is the case for the information incorporation interpretation outlined above—or from finest to coarsest—as is the case in feed-forward processing. This conceptual unity sheds light on the common formal structure shared by these diverse phenomena.

In the next Section 1.2, we review related work. After having established notations and recalled some general notions in Section 1.3, we formally introduce the notion of the successive refinement of the IB in Section 2.1, where we also prove successive refinability in the case of Gaussian vectors and deterministic channel $p(Y|X)$. We then present the convex hull characterisation in Section 2.2, before using it to prove successive refinement for the case of binary source and relevancy variables. The following Section 2.3 leverages the convex hull characterisation to gather some first insights from minimal experiments. These experiments suggest an intuition for defining soft successive refinement, which we formalise in Section 3.1 through a measure of unique information [40], where we provide theoretical motivations for our choice. This new measure is explored in Section 3.2 with additional numerical experiments that highlight the general features described above. The alternative interpretations of both exact and soft SR, in terms of decision problems and feed-forward deep neural networks, are developed in Sections 4.1 and 4.2, respectively. We then describe the limitations and potential future work in Section 5, and conclude in Section 6.

1.2. Related Work

The notion of successive refinement has been long studied in the rate-distortion literature [25–29]. However, classic rate-distortion theory [47] usually considers distortion functions defined on the random variables' *alphabets*, whereas the IB framework can be regarded as a rate-distortion problem only if one allows the distortion to be defined on the space of probability *distributions* [48]. Successive refinement thus needed to be adapted to the IB framework, which was achieved starting from various perspectives.

In [30,31], successive refinement is formulated within the IB framework. Then, Ref. [32] goes further by considering the informationally optimal limits of several-stage processing in general, without comparing it to single-stage processing. In both these works, the problem is initially defined in asymptotic coding terms, and only then given a single-letter characterisation. On the contrary, we will directly define successive refinement from a single-letter perspective. It turns out that our single-letter definition and the operational multi-letter definition from [30,31] are equivalent. The two latter works—as well as [32]—thus provide our single-letter definition with an operational interpretation that also formalises the intuition of an informationally optimal incorporation of information (see Proposition 1 and Appendix B.2).

Another notion named "successive refinement" as well can be found in [46]. This work, instead of modelling information incorporation, rather considers the successive processing of data along a feed-forward pipeline—which encompasses the example of deep neural networks. Fortunately, the "successive refinement" defined in [46] happens to encompass the notion we develop here; more precisely, in [46], the relevancy variable is allowed to vary across processing stages, but if we choose it to be always the same, then "successive refinement" as defined in [46] and "successive refinement" as defined here are formally equivalent (see Section 4.2). In other words, the situation considered in this paper is a particular case of [46], so our results, methods, and phenomenological insights are directly relevant to [46]. For instance, our proof of SR for binary X and Y (see Proposition 5)

is a generalisation of Lemma 1 in [46], which proves SR when X is a Bernoulli variable of parameter $\frac{1}{2}$ and $p(Y|X)$ is a binary symmetric channel.

More generally speaking, the link between successive refinement and the IB theory of deep learning [49–56] has been noted since the inception of the latter research agenda [49], and, besides in [46], it was also further developed in [57]. Section 4.2 makes clear in which sense our results are relevant to this line of research. In particular, our minimal experiments suggest (if they are scalable to the much richer deep learning setting) that trained deep neural networks should lie close to IB bifurcations: i.e., if X is the network's input, Y the feature to be learnt and $L_1, \ldots, L_n$ the network's successive layers, the points $(I(X;L_i), I(Y,L_i))$ should lie close to points of the information curve corresponding to IB bifurcations. This feature was already suggested in [49,50], but for reasons not explicitly related to successive refinement. Note that while the phenomenon of IB bifurcations has been studied from a variety of perspectives (see, e.g., [41–44]), here, we adopt that of [43], which frames IB bifurcations as parameter values where the minimal number of symbols required to represent a bottleneck increases.

In [58], successive refinability is proved for discrete source X and relevancy $Y = X$. Our Proposition 3 generalises this result to either discrete or continuous source X, with relevancy Y being an arbitrary function of X, with a similar argument as that in [58].

In [33], links between the IB framework and renormalisation group theory are exhibited. Even though the questions addressed in the latter work are thus distinct from those addressed here, the Gaussian IB's *semigroup structure* defined and proven in [33] implies the successive refinability of Gaussian vectors (see Proposition 2, and see Appendix 2 for more details on the semigroup structure). This generalises Lemma 3 in [46], which proves SR when X and Y are jointly Gaussian, but each one-dimensional (see Section 4.2 for the relevance of [46] to our framework).

The geometric approach in which we propose to study the successive refinement of the IB is closely related to the convexity approach to the IB [35–39], which frames the IB problem as that of finding the lower convex hull of a well-chosen function. This formulation happens to fit neatly with our convex hull characterisation of successive refinement; we use it to apply the characterisation to proving successive refinability in the case of a binary source and relevancy. Moreover, it is worth noting that our convex hull characterisation makes successive refinement tightly related to the notion of *input-degradedness* [59], through which additional operational interpretations can be given to successive refinement, particularly in terms of randomised games.

The loss of information optimality induced by several-stage processing has already been studied in [60] (see next paragraph), but a quantification of it based on *soft Markovianity* was, to the best of our knowledge, only considered in [18]. Here, we take inspiration in the latter work to quantify soft successive refinement, but we explicitly address the problem that joint distributions over distinct bottlenecks are not uniquely defined. This leads us to use the *unique information* defined in [40] within the context of partial information decomposition [61–64] as our measure of soft SR. This unique information has tight links with the Blackwell order [45,65], which allows us in Section 4.1 to provide a second alternative interpretation of (exact and soft) successive refinement in terms of decision problems.

Ref. [60] proves the near-successive refinability of rate-distortion problems when the distortion measure is the squared error. However, the latter work's approach is different from ours in two respects. First, the distortion measures are different: in particular, as mentioned above, the IB distortion is defined over the space of probability distributions on symbols, unlike the squared error, which is defined on the space of symbols itself. Second, Ref. [60] quantifies the lack of SR as the respective differences between sequences of optimal rates (for given distortion sequences) of a several-stage processing system and the corresponding optimal rates (for the same distortions) of a single-stage processing system. Here, we quantify the lack of SR with a single quantity: the unique information defined by bottlenecks with different granularities. We are, at this stage, not aware of a link

between this value of unique information and differences in one-stage and several-stage optimal rates.

1.3. Technical Preliminaries

In this section, we fix the notations and conventions that we will use along the paper and recall some general notions that we will need.

1.3.1. Notations and Conventions

The random variables are denoted by capital letters, e.g., X, their alphabets by calligraphic ones, e.g., $\mathcal{X}$, and their symbols by lower-case letters, e.g., x. Sometimes, we will mix upper- and lower-case notations to denote a family where some symbols vary, while others are fixed, e.g., $q(X|t) := \big(q(x|t)\big)_{x\in\mathcal{X}}$, or $q(x|T) := \big(q(x|t)\big)_{t\in\mathcal{T}}$. Throughout the whole paper, X is the fixed source and Y the fixed relevancy of the IB problem. The variable T defined by the solution $q(T|X)$ to the primal IB problem (1) is called a *primal* bottleneck. We use the same symbol T for *Lagrangian* bottlenecks, i.e., variables defined by solutions $q(T|X)$ to the Lagrangian bottleneck problem (see Equation (3) below). By "bottleneck" without further specification, we refer to either a primal or Lagrangian bottleneck. The fixed source-relevancy distribution is denoted $p(X,Y)$, and any distribution involving at least one bottleneck is denoted with the letter q, e.g., $q(X,Y,T)$. When it is necessary to make the trade-off parameter explicit, we index the corresponding objects by λ, e.g., $q_\lambda(T|X)$ or $I_\lambda(Y;T)$. Unless explicitly stated otherwise, the source X, relevancy Y, and any considered bottleneck T are defined as either all discrete or all continuous. Probability simplices, and sometimes some of their subsets are written using the generic symbol Δ; for instance, the source simplex is denoted by $\Delta_{\mathcal{X}}$.

Without loss of generality, we always restrict X, Y, and the bottleneck T to their respective supports so that, in particular, all the conditional distributions are unambiguously well-defined, both in the discrete and the continuous case.

We will denote by I_Y the function from $\mathbb{R}_+$ to $\mathbb{R}_+$ defined by $I_Y(\lambda) := I(Y;T)$, where T is a solution to the primal IB problem (1) for the parameter λ. The *information curve*, defined above in Equation (2), is thus also the graph of the function I_Y.

1.3.2. General Facts and Notions

The following properties of the IB framework will be useful [35,37]:

- A bottleneck must saturate the information constraint, i.e., solutions T to (1) must satisfy $I_\lambda(X;T) = \lambda$. In other words, the primal trade-off parameter is the complexity cost of the corresponding bottleneck.
- The function $I_Y : \lambda \mapsto I_\lambda(T;Y)$ is constant for $\lambda \geq H(X)$. We will thus always assume, without loss of generality, that $\lambda \in [0, H(X)]$.
- In the discrete case, choosing a bottleneck cardinality $|\mathcal{T}| = |\mathcal{X}| + 1$ is enough to obtain optimal solutions. Thus, we always assume, without loss of generality, that $|\mathcal{T}| \leq |\mathcal{X}| + 1$, where $|\mathcal{T}| < |\mathcal{X}| + 1$ might occur if needed to make T full support .

To compute bottleneck solutions, instead of directly solving the primal problem (1), following common practice, we will solve its Lagrangian relaxation [66]:

$$\underset{q(T|X)\,:\,T-X-Y,}{\arg\min} \quad I(X;T) \;-\; \beta\, I(Y;T), \tag{3}$$

where the complexity-relevancy trade-off is now parameterised by $\beta \geq 0$, which corresponds to the inverse of the information curve's slope [41]. As the information curve is known to be concave, the Lagrangian parameter β is an increasing function of the primal parameter $\lambda = I(X;T)$. Moreover, we can, without loss of generality, assume that $\beta \geq 1$ [43]. (Note that when the information curve is not strictly concave, the Lagrangian formulation does not allow one to obtain all the solutions to the primal problem [39,67]. However,

in our simple numerical experiments, we always obtained strictly concave information curves.)

We will also need the following concepts [43]:

Definition 1. *Let T be a (primal or Lagrangian) discrete bottleneck. The* effective cardinality *$k = k(T)$ is the number of distinct pointwise conditional probabilities $q(X|t)$ for varying t.*

Definition 2. *A discrete (primal or Lagrangian) bottleneck T is a* canonical bottleneck, *or is in* canonical form, *if all the pointwise conditional probabilities $q(X|t)$ are distinct, i.e., equivalently, if $|\mathcal{T}| = k(T)$, where $k(T)$ is the effective cardinality of T.*

Our definition of effective cardinality, even though slightly different from the original one in [43], is equivalent to the latter for Lagrangian bottlenecks. And, importantly, every (primal or Lagrangian) bottleneck can be reduced to its canonical form by merging the symbols with identical $q(X|t)$ (see Appendix A.1 for more details). We will be particularly interested in the *change* of effective cardinality, which has been identified in [43] as characterising the bottleneck phase-transitions, or *bifurcations*.

In Figure 1, we present examples of bottleneck conditional distributions $q(X|T)$, visualised as the family of points $\{q(X|t),\, t \in \mathcal{T}\}$ on the source simplex $\Delta_{\mathcal{X}}$, where, here, $|\mathcal{X}| = 3$, and the bottleneck is computed with $|\mathcal{T}| = 3$ in both examples. However, in Figure 1 (left), there are only two distinct $q(X|t)$, so there must be two equal pointwise probabilities $q(X|t_1)$ and $q(X|t_2)$; thus, $k = 2$ and the canonical form of T is obtained by merging t_1 and t_2. On the contrary, in Figure 1 (right), there are three distinct $q(X|t)$, so, here, $k = 3$ and the bottleneck is already in canonical form.

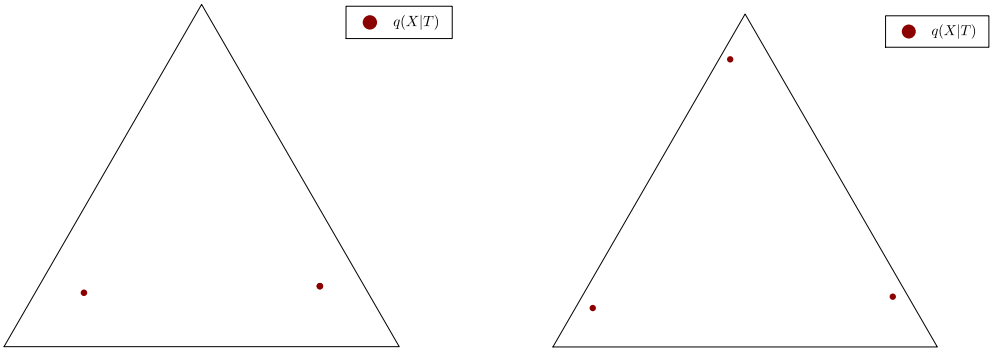

Figure 1. Examples of distributions $q(X|T)$, visualised as families of points $\{q(X|t),\, t \in \mathcal{T}\}$ on the source simplex $\Delta_{\mathcal{X}}$, where, here, $|\mathcal{X}| = 3$. Each of the triangle's vertices represents the Dirac probability of some $x \in \mathcal{X}$. The bottleneck's effective cardinality is $k = 2$ on the left and $k = 3$ on the right.

Eventually, the notions of *consistency* and *extension* will be crucial to us.

Definition 3. *Let $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ be a Cartesian product of (continuous or discrete) alphabets. For $C = \{c_1, \ldots, c_r\} \subseteq \{1, \ldots, m\}$ a subset of coordinates, we write*

$$\bigtimes_{c \in C} \mathcal{A}_c := \mathcal{A}_{c_1} \times \cdots \times \mathcal{A}_{c_r}.$$

For each $1 \leq i \leq n$, we consider a subset of coordinates C_i and a probability distribution q_i over $\bigtimes_{c \in C_i} \mathcal{A}_c$. The distributions $q_1, \ldots, q_n$ are said to be consistent *if, for every $1 \leq i, j \leq n$, the respective marginals of q_i and q_j on their common coordinates $\bigtimes_{c \in C_i \cap C_j} \mathcal{A}_c$ are equal.*

For instance, if T_1 and T_2 are two bottlenecks, they define consistent distributions $q_1(X, Y, T_1)$ and $q_2(X, Y, T_2)$ because, by definition, their respective marginals on their common coordinates $\mathcal{X} \times \mathcal{Y}$ are $q_1(X, Y) = q_2(X, Y) = p(X, Y)$.

Definition 4. *Let $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ be a Cartesian product of (continuous or discrete) alphabets, and $q_1, \ldots, q_n$ be consistent probability distributions over distinct but potentially overlapping coordinates of $\mathcal{A}$. A distribution q over the whole $\mathcal{A}$ is called an* extension *of the family of distributions $\{q_1, \ldots, q_n\}$ if it is consistent with each q_i.*

Consider bottlenecks $T_1, \ldots, T_n$ of same source X and relevancy Y for resp. parameters $\lambda_1, \ldots, \lambda_n$. They define a consistent family of distributions $\{q_{\lambda_i}(X, T_i),\ 1 \leq i \leq n\}$. One of the central mathematical objects of this work is the set of their extensions into *joint* distributions $q(X, T_1, \ldots, T_n)$:

Notation 1. *For given bottlenecks $T_1, \ldots, T_n$ of respective parameters $\lambda_1, \ldots, \lambda_n$, we denote by $\Delta_{\lambda_1,\ldots,\lambda_n}$ the set of extensions $q(X, T_1, \ldots, T_n)$ of the family of distributions $\{q_{\lambda_i}(X, T_i),\ 1 \leq i \leq n\}$.*

In general, for a fixed family of bottlenecks, there is a multitude of possible ways to extend them into a joint distribution; indeed, $\Delta_{\lambda_1,\ldots,\lambda_n}$ traces a polytope on the simplex $\Delta_{\mathcal{X}\times\mathcal{T}_1\times\cdots\times\mathcal{T}_n}$ of joint distributions (see Appendix A in [40]). This feature is the formal version of our previous statement that the IB framework does not entirely specify the *relationship* between representations $T_1, \ldots, T_n$: it only constrains it through the set $\Delta_{\lambda_1,\ldots,\lambda_n}$. Questions about possible relationships between IB representations are thus questions about properties of the set $\Delta_{\lambda_1,\ldots,\lambda_n}$.

2. Exact Successive Refinement of the IB

2.1. Formal Framework and First Results

Here, we formally describe, within the IB framework, the rate-distortion-theoretic notion of *successive refinement* (SR) [25–27,29]. We propose a purely single-letter definition (i.e., we only consider single source, relevancy, and bottleneck variables), which makes the presentation simpler but still conveys the intuition of information incorporation. After having presented the notion of SR in the IB framework, we describe its Markov chain characterisation (see Proposition 1), which mirrors the characterisation of SR for classic rate-distortion problems [26], and makes our formulation equivalent to previous multi-letter operational definitions, which also formalise the intuition of information incorporation [30–32]. We then leverage this characterisation to prove SR in the case of Gaussian vectors and deterministic channel $p(Y|X)$.

Intuitively, there is successive refinement when a finer bottleneck T_2 does not discard any of the information extracted by a coarser bottleneck T_1. This can be imposed by requiring that $T_2 = (T_1, S_2)$ for some variable S_2, which encodes the "supplement" of information that "refines" T_1 into T_2. In the general case:

Definition 5. *Let $0 < \lambda_1 < \cdots < \lambda_n$, and a discrete or continuous $p(X, Y)$ be given. There is* successive refinement *(SR) for parameters $(\lambda_1, \ldots, \lambda_n)$ if there exist variables $(T_1, S_2, S_3, \ldots, S_n)$ such that*

- *T_1 is a bottleneck with parameter λ_1;*
- *For every $2 \leq i \leq n$, the variable $T_i := (T_{i-1}, S_i)$ is a bottleneck with parameter λ_i.*

Note that even though it does not appear explicitly in this definition, the relevancy variable Y is indeed crucial to it, as it defines what a bottleneck is (see Equation (1)). If the conditions of Definition 5 hold, we will also say that the IB problem defined by $p(X, Y)$ is $(\lambda_1, \ldots, \lambda_n)$-refinable. If bottlenecks $T_1, \ldots, T_n$ satisfy the definition's conditions, we will say that they *achieve* successive refinement, or, simply, that there is successive refinement between these bottlenecks. If there is successive refinement for all combina-

tions $0 < \lambda_1 < \cdots < \lambda_n$ of trade-off parameters, we will say that the corresponding IB problem is successively refinable. Eventually, when it will be needed in later sections to contrast this notion with that of soft successive refinement, we will refer to it as *exact* successive refinement.

For instance, let $0 < \lambda_1 < \lambda_2$ and $\mathcal{X} = \mathcal{Y} = \{0,1\}$. We consider $Y := X \oplus Z$, where $\oplus$ denotes the modulo-2 addition, and X and Z are Bernouilli variables with parameters $\frac{1}{2}$ and a, respectively, for an arbitrary $0 \leq a \leq \frac{1}{2}$. In this case, it is proven in Lemma 1 of [46] that, for well-chosen binary variables S_1 and S_2, we have that X, S_1, and S_2 are mutually independent, and the variables $X \oplus S_1$ and $X \oplus S_1 \oplus S_2$ are bottlenecks of resp. parameters λ_1 and λ_2. Moreover, using the independence of S_2 with $(X, X \oplus S_1)$ and the assumed Markov chain $Y - X - X \oplus S_1 \oplus S_2$, a straightforward computation shows that to get a bottleneck of parameter λ_2, the variable $X \oplus S_1 \oplus S_2$ can be replaced by $(X \oplus S_1, S_2)$. Thus, here, the IB problem is (λ_1, λ_2)-refinable, where successive refinement is achieved by $T_1 := X \oplus S_1$ and $T_2 = (T_1, S_2)$.

It is helpful to visualise SR on the information plane, i.e., that on which lies the information curve. Indeed, successive refinement can be understood in terms of specific translations on the information plane: those resulting from concatenating an already existing variable T_{i-1} with a new variable S_i—let us call them "accumulative translations" because they result from a processing that does not discard any of the information already collected. Let us focus on the case $n = 2$ and first note that, whether or not (T_1, S_2) is a bottleneck, we have

$$I(X; T_1, S_2) = I(X; T_1) + I(X; S_2|T_1),$$

and, similarly,

$$I(Y; T_1, S_2) = I(Y; T_1) + I(Y; S_2|T_1).$$

In other words, the measure of both the complexity cost and relevance for (T_1, S_2) can be decomposed into the same measures first for T_1 and then for the "supplement" of information S_2, conditionally on the "already collected" information T_1. In Figure 2 (left and right), we first fix a coarse bottleneck T_1, understood here as a point $\big(I(X;T_1), I(Y;T_1)\big)$ on the information curve. Once T_1 is known, we supplement it with a new variable S_2, which incurs both an additional complexity cost $I(X; S_2|T_1)$ and an additional relevant information gain $I(Y; S_2|T_1)$. The question of successive refinement is that of whether the additional complexity cost can be leveraged enough for the resulting relevant information gain to take (T_1, S_2) "up to the information curve", i.e., to be such that $\big(I(X;T_1,S_2), I(Y;T_1,S_2)\big)$ is on the information curve. This is the case in Figure 2, right, and not the case in Figure 2, left. In short, there is successive refinement between two points on the information curve if and only if there exists an "accumulative translation" from the coarser one to the finer one.

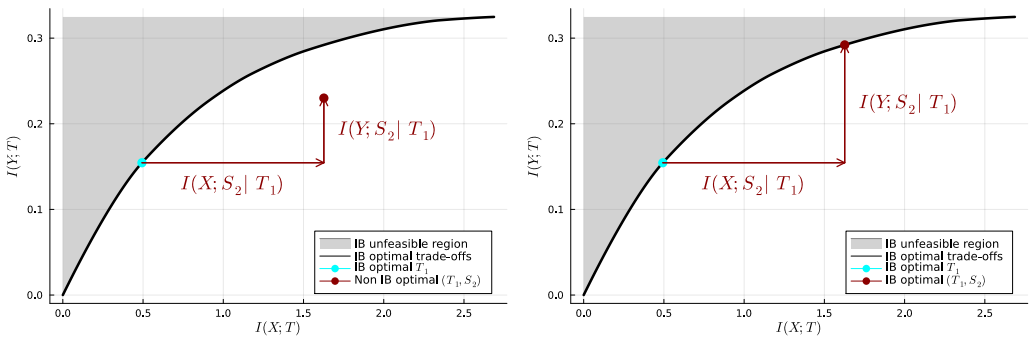

Figure 2. Successive refinement visualised on the information plane. On the left, adding the information from the variable S_2 (the supplement variable) is not efficient enough to achieve successive refinement. On the right, it is. See main text for details (the values of $I(X; S_2|T_1)$ and $I(Y; S_2|T_1)$ have been chosen arbitrarily to illustrate each case).

Let us now describe a more formal characterisation, where point (ii) will mirror the characterisation of SR for classic rate-distortion problems [26].

Proposition 1. *Let $0 < \lambda_1 < \cdots < \lambda_n$. The following are equivalent:*

(i) *There is successive refinement for parameters $(\lambda_1, \ldots, \lambda_n)$;*

(ii) *There exist bottlenecks $T_1, \ldots, T_n$, of common source X and relevancy Y, with respective parameters $\lambda_1, \ldots, \lambda_n$, and an extension $q(X, T_1, \ldots, T_n)$ of the $q_i := q_i(X, T_i)$, such that, under q, we have the Markov chain*

$$X - T_n - \cdots - T_1. \tag{4}$$

(iii) *There exist bottlenecks $T_1, \ldots, T_n$, of common source X and relevancy Y, with respective parameters $\lambda_1, \ldots, \lambda_n$, and an extension $q(Y, X, T_1, \ldots, T_n)$ of the $q_i := q_i(Y, X, T_i)$, such that, under q, we have the Markov chain*

$$Y - X - T_n - \cdots - T_1. \tag{5}$$

Proof. See Appendix B.1. It is relatively straightforward because we started directly from a single-letter definition. □

Proposition 1 was already known to be a characterisation of SR of the IB [30–32]. However, as the latter references start from an operational problem in terms of asymptotic rates and distortions for multi-letter systems, here, Proposition 1 shows that our single-letter Definition 5 is equivalent to the operational definitions in [30–32]. See Appendix B.2 for more details.

Remark 1. *Crucially, the order of the indexing in (4) and (5) depends only on the order of the trade-off parameters $\lambda_1 < \cdots < \lambda_n$, and not on the order in which the bottlenecks T_i are produced, which is just the interpretation we started from. In particular, Proposition 1 makes equally legitimate the interpretation of bottlenecks produced from the finest one to the coarsest one, each new bottleneck thus implementing a further coarsening of the source X. This alternative interpretation renders successive refinement relevant to feed-forward processing, including in particular the Blackwell order (see Section 4.1) and deep neural networks (see Section 4.2). For ease of presentation, though, we will stick to the information incorporation interpretation along most of the paper.*

Moreover, from Proposition 1, we can leverage existing IB literature to prove the successive refinability of two specific settings. (For an explicit definition of what we mean, in Proposition 2, by successive refinement in the case of the *Lagrangian* IB problem, see Appendix B.3.)

Proposition 2. *If X, Y are jointly Gaussian vectors, then the Lagrangian IB problem defined by $p(X, Y)$ is $(\lambda_1, \ldots, \lambda_n)$-refinable for all $\lambda_1 < \cdots < \lambda_n$.*

This result is a direct consequence of a property named a *semigroup structure*, and is proven for the Gaussian IB framework in [33], which relates the latter framework with renormalisation group theory. The semigroup structure denotes, in short, the situation where iterating the operation of coarse graining a variable by computing a bottleneck—where, at each iteration, the previous bottleneck becomes the source of the next IB problem—still outputs a bottleneck for the original problem. This semigroup structure is a stronger property than successive refinement and, as it is satisfied in the Gaussian case, this implies the successive refinability of Gaussian vectors (see Appendix B.3 for more details). Beyond Proposition 2, this relationship between successive refinement and the semigroup structure hints at potentially interesting links between the composition of coarse-graining operators and successive refinement. In this respect, note that our numerical results below (see Sections 2.3 and 3.2) suggest that, for non-Gaussian vectors, successive refinement does not

always hold and thus, *a fortiori*, that the semigroup structure might not always be satisfied in the IB framework—or at least not perfectly.

Eventually, in the case of deterministic channel $p(Y|X)$, an explicit solution to the IB problem (1) is known [34]: $T = Y$ with probability α, and $T = e$ with probability $1 - \alpha$, for e a dummy symbol, and some well-chosen $0 < \alpha < 1$. This specific solution allows one to address successive refinement for the deterministic case:

Proposition 3. *Let X be a discrete or continuous variable, and Y be a deterministic function of X. Then, the IB problem defined by $p(X, Y)$ is successively refinable for all trade-off parameters $\lambda_1 < \cdots < \lambda_n$.*

Proof. See Appendix B.4. A proof was already proposed, from an asymptotic coding perspective, for discrete X and $Y = X$, in [58]. We use a similar argument here. □

Note, though, that the solution used here to prove successive refinement is, as noted in [34], not very interesting: it is nothing more than an increasingly noisy version of Y. It is not clear whether or not there exists more interesting bottleneck solutions in the deterministic case, and if so, whether these other solutions are successively refinable. Proposition 3 will in any case be useful for our own purposes: we will use it to set aside the deterministic case in the proof of SR for binary X and Y (Proposition 5 below).

Until now, we used existing results from the IB literature that, even though not originally aimed at it, happen to yield interesting consequences for the problem of the successive refinement of the IB. However, it seems crucial, for further progress on the latter topic, to design specifically tailored mathematical and numerical tools. This is the purpose of the following sections of this paper; in particular, in the next section, we present a simple geometric characterisation of the IB's successive refinability.

2.2. The Convex Hull Characterisation and the Case $|\mathcal{X}| = |\mathcal{Y}| = 2$

In this section, we present our convex hull characterisation of successive refinement. We then show its relevance both to numerical computations—thanks to a linear program for checking the condition—and to proving new mathematical results—which we exemplify by proving, thanks to this new characterisation, the successive refinability of binary variables. Here, as in our subsequent numerical experiments in Section 2.3, we will focus on discrete variables and $n = 2$ processing stages, even though our results are thought of as a first step towards a generalisation to continuous variables and an arbitrary number of processing stages.

The convexity approach that we propose hinges upon changing the perspective on the IB problem (1) from an optimisation over the encoder channels $q(T|X)$ to an optimisation over the *decoder* channels $q(X|T)$; indeed, (1) can be equivalently presented as the "reversed" optimisation problem

$$\underset{\substack{(q(T),q(X|T)):\\ \sum_t q(t)q(X|t)=p(X)\\ T-X-Y,\ I(X;T)\leq\lambda}}{\arg\max} \quad I(Y;T). \tag{6}$$

Formulations (1) and (6) yield the same solutions because, through the Markov chain $T - X - Y$, the joint distribution $q(X, Y, T)$ is equivalently determined by specifying some $q(T|X)$ or specifying some pair $\big(q(T), q(X|T)\big)$ that satisfies the consistency condition $\sum_t q(t)q(X|t) = p(X)$. This condition says that the source distribution $p(X)$ must be retrievable as a convex combination of the $q(X|t)$, where the weights are given by the $q(t)$.

Moreover, this formulation leads to a crucial intuition concerning the relationship between successive refinement and the set $\mathcal{H}_T := Hull\{q(X|t), t \in \mathcal{T}\}$, where, for a set $E \subseteq \mathbb{R}^n$, we denote by $Hull(E)$ the convex hull of E, i.e., the set of points obtained as convex combinations of points in E. First, note that, for a bottleneck T, the set $\mathcal{H}_T$ is reduced to a single point if and only if T is independent from the source X. Conversely, $\mathcal{H}_T$ coincides

with the whole source simplex $\Delta_{\mathcal{X}}$ if and only if T captures all the information from the source, i.e., if $I(X;T) = H(X)$. Generalising these extreme cases suggests the intuition that $\mathcal{H}_T$ describes the *information content* held by the bottleneck T about the source X. Now, let us recall that successive refinement from a coarse bottleneck T_1 to a finer bottleneck T_2 means intuitively that T_2 can be obtained without discarding any of the information extracted by T_1 about the source X; in other words, that the information content of T_1 about the source X *is included* in that of T_2. Combining this latter intuition with the one about $\mathcal{H}_T$ being the information content of a bottleneck T suggests the following characterisation of successive refinement:

$$Hull\{q(X|t_1), t_1 \in \mathcal{T}_1\} \subseteq Hull\{q(X|t_2), t_2 \in \mathcal{T}_2\}, \tag{7}$$

where T_1 and T_2 are bottlenecks of parameters $\lambda_1 < \lambda_2$, respectively. This condition is visualised in Figure 3. The characterisation indeed holds, at least for the discrete case and under a mild assumption of injectivity of the finer bottleneck's decoder:

Proposition 4. *Let $0 < \lambda_1 < \lambda_2$, and assume that $p(X,Y)$ is discrete.*

If there is successive refinement for parameters (λ_1, λ_2), then there exist bottlenecks T_1, T_2 of parameters λ_1, λ_2, respectively, such that the convex hull condition (7) is satisfied.

Conversely, if there exist bottlenecks T_1, T_2 of parameters λ_1, λ_2, respectively, such that the convex hull condition (7) holds and such that the decoder $q(X|T_2)$, seen as a probability transition matrix, is injective, then there is successive refinement for parameters (λ_1, λ_2). Moreover in this latter case, if T_1, T_2 are bottlenecks that achieve successive refinement, the extension $\tilde{q}(X, T_1, T_2)$ of $q(X, T_1)$ and $q(X, T_2)$ such that $X - T_2 - T_1$ holds is uniquely defined.

Proof. See Appendix B.5. The idea consists in translating the Markov chain characterisation $X - T_2 - T_1$ into the convex hull condition (7). The direct sense is straightforward. For the converse direction, observe that, even though as soon as (7) is satisfied it provides a joint distribution $\tilde{q}(X, T_1, T_2)$ that satisfies the Markov chain $X - T_2 - T_1$, it is not clear whether this distribution is consistent with $q(X, T_1)$. The potential problem stems from the fact that $\tilde{q}$ must be such that the channel $\tilde{q}(T_2|T_1)$ maps the marginal $q(T_1)$ to the marginal $q(T_2)$. The injectivity assumption, however, provides a sufficient condition for it to be the case. This assumption happens to also imply the uniqueness of the extension, among all those that satisfy the Markov chain $X - T_2 - T_1$. □

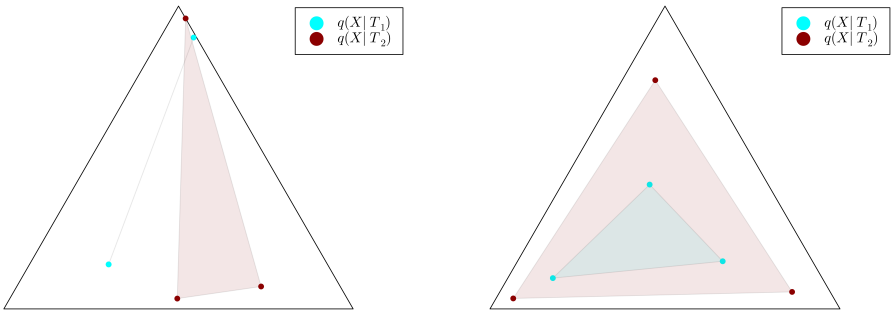

Figure 3. Illustration of the convex hull condition. The black triangle represents the source simplex $\Delta_{\mathcal{X}}$ with, here, $|\mathcal{X}| = 3$, and the pointwise bottleneck decoder probabilities $\{q(X|t),\, t \in \mathcal{T}\}$ are represented on it (in cyan for the coarser bottleneck T_1 and in red for the finer one T_2). The convex hull of the respective families of points are shaded with the corresponding color. On the left, the condition is not satisfied; on the right, it is.

Even though the injectivity assumption might seem restrictive, in practice, in our numerical experiments below (see Sections 2.3 and 3.2), we always found that the decoder channel $q(X|T_2)$ could be chosen as injective by reducing it to its effective cardinality (see Section 1.3)—a process that leaves the convex hull condition (7) unchanged because it leaves

the points $q(X|t_2)$ unchanged. See also Appendix D for a conjecture that, if true, would simplify our convex hull characterisation in the case of a strictly concave information curve.

Remark 2. *The convex hull condition happens to be equivalent to the* input-degradedness *preorder on channels (see Proposition 1 in [59]). Even though we will not develop this point further when considering alternative interpretations of SR (Section 4), it is worth noting that, through input-degradedness, SR can be given additional operational interpretations, particularly in terms of randomised games (see Section IV-C in [59]).*

Our new characterisation provides a simple way of checking whether or not two bottlenecks T_1 and T_2 achieve SR. Recall that the Markov chain characterisation (Proposition 1, point (ii)) shows that SR is a feature of the space Δ_{q_1,q_2} of all extensions $q(X,T_1,T_2)$ of individual bottleneck distributions $q_1(X,T_1)$ and $q_2(X,T_2)$. While this set might, *a priori*, be difficult to study directly, our characterisation (7) reduces the problem to a simple geometric property relating only two explicitly given conditional distributions: $q(X|T_1)$ and $q(X|T_2)$. Moreover, note that (7) is equivalent to

$$\forall t_1 \in \mathcal{T}_1, \quad q(X|t_1) \in Hull\{q(X|t_2), t_2 \in \mathcal{T}_2\},$$

and that checking whether a point is in the convex hull of a finite set of other points can be cast as a linear programming problem [68]. As a consequence, one can bound the time complexity of checking condition (7) as $O(|\mathcal{X}|\,K)$, where K is the time complexity bound of a linear program with $2|\mathcal{X}|+2$ variables and $3|\mathcal{X}|+2$ constraints. As a consequence, using the bound on K proved in [69], the time complexity of checking (7) is no worse that $\tilde{O}(|\mathcal{X}|^{\omega+1}\log(\frac{|\mathcal{X}|}{\delta}))$, where $\omega \approx 2.38$ corresponds to the complexity of matrix multiplication, δ is the relative accuracy, and the $\tilde{O}(\cdot)$ notation hides polylogarithmic factors (see Appendix B.6 for details).

We deem this convex hull characterisation to be important for theory as well. Indeed, it reduces the question of successive refinement to a question about the structure of the trajectories, on the source probability simplex $\Delta_{\mathcal{X}}$, of the points $q_\lambda(X|t)$ for varying λ. Thus, any theoretical progress on the description of these bottleneck trajectories might lead to theoretical progress on the side of successive refinement. As a first step in this direction, we show that this geometric point of view helps to solve the question of SR in the case of a binary source and relevancy (This result generalises the already known fact that there is always successive refinement when X is a Bernoulli variable of parameter $\frac{1}{2}$ and $p(Y|X)$ is a binary symmetric channel (see Lemma 1 in [46] and see Section 4.2 for explanations on why the latter work's framework encompasses ours). Moreover, a potential generalisation of our result to an arbitrary number of processing stages is left to future work).

Proposition 5. *If $|\mathcal{X}| = |\mathcal{Y}| = 2$, then, for any discrete distribution $p(X,Y)$ and any trade-off parameters $\lambda_1 < \lambda_2$, the IB problem defined by $p(X,Y)$ is (λ_1,λ_2)-successively refinable.*

Proof. Let us here outline the proof presented in Appendix B.7. The case of deterministic $p(Y|X)$ was already dealt with in Proposition 3, so we can assume that $p(Y|X)$ is not deterministic. In this case, the IB problem with $|\mathcal{X}| = |\mathcal{Y}| = 2$ and $n = 2$ has been extensively studied in [35]. In short, the latter approach leverages the fact that a pair $(q(T), q(X|T))$ is a solution to the IB problem (6) if the convex combination of the points $F_\beta(q(X|t))$, with weights given by $q(T)$, achieves the lower convex envelope of the function F_β, where F_β is a well-chosen function on the source simplex $\Delta_{\mathcal{X}}$ and β is the information curve's inverse slope. This work, along with considerations from [37], which uses the same convexity approach, yields in particular that (i) the points $q_\beta(X|t)$ are the extreme points of a non-empty open segment uniquely defined by β, and (ii) this latter segments grows as a function of the inverse slope β and thus, by concavity, as a function of λ. This implies that the convex hull condition is always satisfied for $\lambda_1 < \lambda_2$. As point (i) also implies

that, here, $q_{\lambda_2}(X|T_2)$ must be injective, Theorem 4 allows us to conclude the successive refinability for $n = 2$ processing stages. □

The proof of Proposition 5 exemplifies how our convex hull characterisation interlocks well with the convexity approach to the IB [35–39]. In this sense, our characterisation brings a new theoretical tool to the study of the successive refinement of the IB.

2.3. Numerical Results on Minimal Examples

In this section, we leverage our new convex hull characterisation to investigate successive refinement on minimal numerical examples, i.e., with discrete and low-cardinality distributions $p(X, Y)$. Our experiments suggest that, in general, successive refinement does not always hold exactly. However, they also highlight two other features: first, it seems that successive refinement is often shaped by IB bifurcations [41–44]. Second, even though successive refinement is often not satisfied exactly, visualisations suggest that it is often "close" to being satisfied. The formalisation of this latter intuition will be the topic of the next section.

We consider the Lagrangian form (3) of the IB problem (see Section 1.3). We compute solutions to it with the Blahut–Arimoto (BA) algorithm [1], combined with reverse deterministic annealing [19,70], starting from $\beta \approx \infty$ (i.e., in practice, $\beta \gg 1$) at the IB solution $T = X$ (we noticed that regular deterministic annealing sometimes yielded sub-optimal solutions because they followed sub-optimal branches at IB bifurcations [1,71], which was not the case for reverse annealing). We always obtained that $I(X;T)$ was a strictly increasing function of the Lagrangian parameter β, so it makes sense to index the solutions by $\lambda = I(X;T)$ rather than β; for instance, in this section and Section 3.2, we will write $q_\lambda(T|X)$ for our algorithm's output for a β such that $I(X;T) = \lambda$.

In all our numerical experiments, after reducing a bottleneck T to its canonical form (see Section 1.3), the decoder channel $q_\lambda(X|T)$ was injective. Therefore, thanks to Theorem 4, the convex hull condition (7) being satisfied here does imply successive refinement. In the remainder of the paper, we will thus use the convex hull condition as a proxy for numerically assessing successive refinement (see Appendix D for more details on what we mean by "proxy"). This condition can be investigated in two ways. First, for two distinct trade-off parameters $\lambda_1 < \lambda_2$, we can compute whether the convex hull condition (7) holds or not with the linear program described in Appendix B.6. Second, for $|\mathcal{X}| \leq 3$, we can visualise the whole trajectories, for varying λ, of the points $q_\lambda(X|t)$ on the source simplex $\Delta_\mathcal{X}$. As we will see, this yields interesting qualitative insights.

As a sanity check for our algorithm, we compute bottleneck solutions for binary X and Y, which we proved in Proposition 5 to be successively refinable for all trade-off parameters. We used the linear program to check the convex hull condition numerically for all pairs $\lambda_1 < \lambda_2$ and for distributions $p(X, Y)$ uniformly sampled on the joint probability simplex $\Delta_{\mathcal{X}\times\mathcal{Y}}$. We find that the convex hull condition is indeed always numerically satisfied.

Then, we study the case $|\mathcal{X}| = |\mathcal{Y}| = 3$, once again uniformly sampling example distributions $p(X, Y)$ on $\Delta_{\mathcal{X}\times\mathcal{Y}}$. Figures 4–6 show, for representative examples, visualisations of the trajectories over λ of the $q_\lambda(X|t)$ (left)—which we will refer to as the *bottleneck trajectories*—along with the corresponding computations of the convex hull condition as a function of λ_1 and $\lambda_2 \geq \lambda_1$ (right)—which we will refer to as the *SR patterns* (The correspodning $p(Y|X)$ are plotted in Appendix E, and $p(X)$ is shown in Figures 4–6 (left). The explicit $p(X, Y)$ corresponding to each of these paper's figures can be found at: https://gitlab.com/uh-adapsys/successive-refinement-ib/(accessed on 12 September 2023).

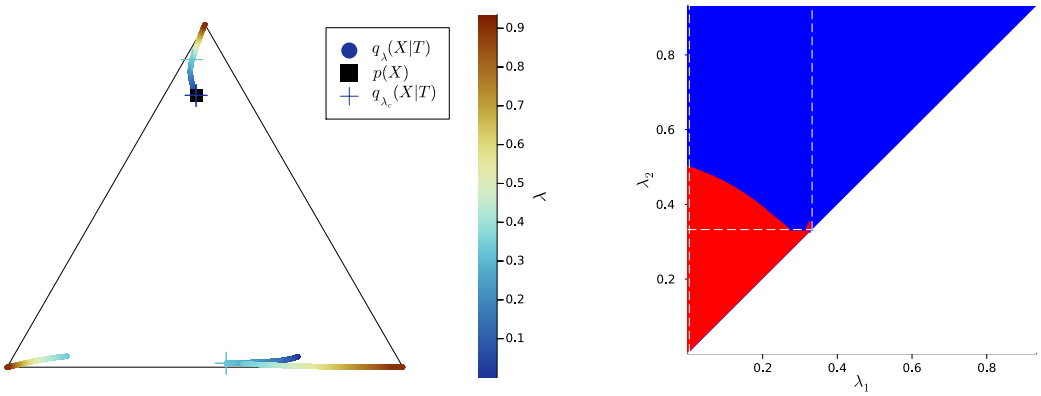

Figure 4. Left: bottleneck trajectories for an example distribution $p(X,Y)$ such that $|\mathcal{X}| = |\mathcal{Y}| = 3$, i.e., trajectory of $q_\lambda(X|T)$, represented as the family of points $\{q_\lambda(X|t),\, t \in \mathcal{T}\}$ on the source simplex $\Delta_\mathcal{X}$, as a function of $\lambda = I(X;T)$ (crosses: value of $q_{\lambda_c}(X|T)$ just before a symbol split at a critical parameter λ_c, where the crosses' color corresponds to the value of λ_c). The conditional distribution $q_\lambda(X|T)$ is defined by the single point $p(X)$ when $\lambda = 0$ (dark blue cross on the black square), or by two distinct points between the first and second symbol splits (dark blue to cyan), or by three distinct points after the second symbol split (cyan to red). Note the discontinuity of $q_\lambda(X|T)$ at each symbol split (without the discontinuity, the trajectory around a symbol split would look like a branching). Right: corresponding SR pattern, i.e., corresponding output for the convex hull condition (blue: satisfied; red: not satisfied; dashed white lines: critical values $\lambda_c(i)$ of either λ_1 or λ_2). For instance, the critical value $\lambda_c(2) \approx 0.33$ corresponds, on the bottleneck trajectories (left), to the symbol split from two to three symbols (cyan crosses). Note that $\lambda_c(1) \approx 0$. The respective $p(Y|X)$ corresponding to this figure and to Figures 5 and 6 are plotted in Appendix E.

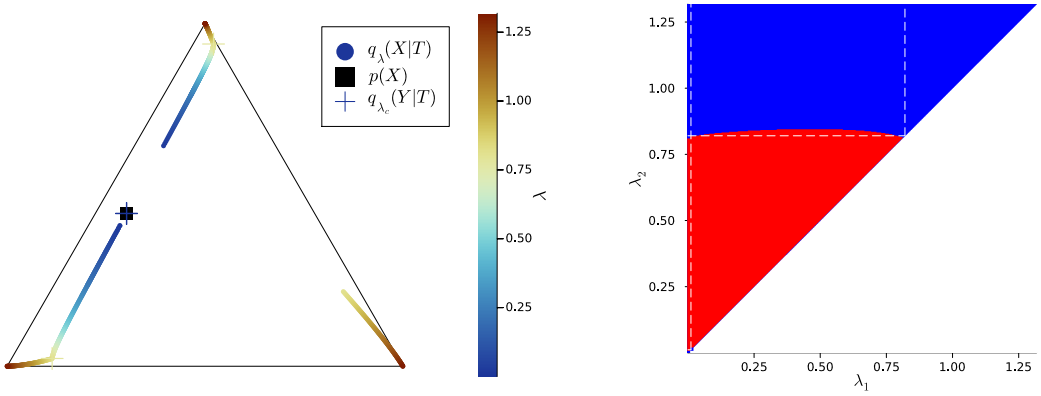

Figure 5. Same as Figure 4, with a different example distribution $p(X,Y)$ such that $|\mathcal{X}| = |\mathcal{Y}| = 3$.

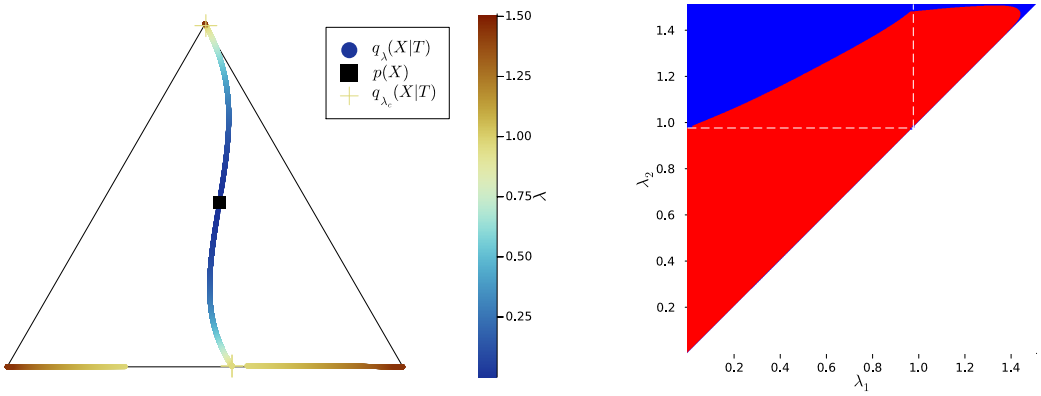

Figure 6. Same as Figure 4, with a different example distribution $p(X,Y)$ such that $|\mathcal{X}| = |\mathcal{Y}| = 3$.

Let us first give a general description of the bottleneck trajectories. For $\lambda \approx 0$, the $q_\lambda(X|t)$ all coincide with the source distribution $p(X)$. This should be the case, as, for $0 = \lambda = I(X;T)$, the bottleneck T is independent of X. Then, when λ increases, the trajectories seem piecewise continuous, where each discontinuity corresponds to a symbol split, i.e., a change in effective cardinality (see Section 1.3). We mark with a cross, for each $t \in \mathcal{T}$, the $q_\lambda(X|t) = q_{\lambda_c}(X|t)$ located just before such a change in effective cardinality.

In the examples of Figures 4–6, as $|\mathcal{X}| = 3$, there are two symbol splits, corresponding to that from one to two and two to three symbols, respectively. Eventually, for large λ, the last continuous segment of bottleneck trajectories corresponds to effective cardinality $k(T_\lambda) = |\mathcal{X}|$, and, for the maximal λ, each corner of the source simplex $\Delta_\mathcal{X}$ is reached by $q(X|t)$ for some $t \in \mathcal{T}$. This means that for maximum λ, there is a deterministic bijective relationship between T and X. The latter is expected: for maximum λ, bottlenecks are minimal sufficient statistics of X for Y [72]; where for $p(X,Y)$ sampled uniformly on the simplex, these minimal sufficient statistics are, with probability 1, just permutations of X.

Definition 6. *In the following, we refer to the piece of trajectory where the bottleneck's effective cardinality $k = k(T_\lambda)$ is equal to the integer i as the "*segment $k = i$*", i.e., it is the segment where $q_\lambda(X|T)$ corresponds to exactly i distinct points on the source simplex $\Delta_\mathcal{X}$; for instance, in Figure 4, the segment $k = 2$ corresponds to the first piece of trajectory spanning colors from dark blue to cyan.*

Notation 2. *We denote by $\lambda_c(i)$ the trade-off parameter's critical value corresponding to the i-th change in effective cardinality, i.e., the symbol split from i to $i+1$ symbols. Here, we will only need to consider the critical values $\lambda_c(1) = 0$ and $\lambda_c(2)$, corresponding to the splits from one to two and two to three symbols, respectively.*

Let us now come back to the question of successive refinement: for which parameters $\lambda_1 < \lambda_2$ is the convex hull condition satisfied? The right-hand sides of Figures 4–6 provide the answers corresponding to trajectories on the respective left-hand sides—where blue and red mean that the condition is and is not satisfied, respectively. Moreover, we highlight with dashed white vertical and horizontal lines the critical parameter values $\lambda_1 = \lambda_c(i)$ and $\lambda_2 = \lambda_c(i)$, respectively, at which the symbol split occurs (see Appendix B.8 for details on the computation of these symbols splits). Note that we always have $\lambda_c(1) \approx 0$, which is expected, as a bottleneck T corresponding to some $\lambda = I(X;T) > 0$ must necessarily define at least two distinct $q_\lambda(X|t)$.

First, in these examples as in most non-reported examples, the convex hull condition (right) breaks as long as $\lambda_2 < \lambda_c(2)$, i.e., as long as the finer bottleneck's effective cardinality is at most $k = 2$. This can also be read from the bottleneck trajectories (left): if the condition was satisfied for all $\lambda_1 < \lambda_2 < \lambda_c(2)$, for instance, then the segment $k = 2$ would be a line segment. This is clearly not the case in Figures 4 and 6, and even though visually it virtually seems to be the case in Figure 5, the segment $k = 2$ happens to be very slightly curved, which is enough to break the convex hull condition. In other words, for $\lambda_1 < \lambda_2 < \lambda_c(i)$, several-stage processing seems to induce, in these examples, a nonzero loss of information optimality.

Then, for $\lambda_2 > \lambda_c(2)$, even though there is no single general pattern, the trajectory's structure at the bifurcation seems to impact successive refinement. Indeed, at the bifurcation at $\lambda_c(2)$, the set $Hull\{q_{\lambda_2}(X|t), t \in \mathcal{T}\}$ opens up along a new, third dimension, and keeps widening when λ_2 increases. This allows it to (gradually in Figures 4 and 6, or virtually straight away in Figure 5) encompass the segment $k = 2$ because it "overcomes" the curvature of this piece of trajectory. For instance, in Figure 4, because the segment $k = 2$ is strongly curved, the convex hull condition gets satisfied for all $\lambda_1 < \lambda_c(2)$ only if λ_2 is significantly larger than $\lambda_c(2)$. On the contrary, because in Figure 5, the segment $k = 2$ is virtually not curved, it is almost as soon as $\lambda_2 > \lambda_c(2)$ that the convex hull condition is satisfied for all $\lambda_1 < \lambda_c(2)$.

In Figure 6, the lack of successive refinement for $\lambda_2 > \lambda_c(2)$ does not seem to be due to the same phenomenon as the one just described. Generally speaking, we observed a whole variety of SR patterns (see Appendix F for more examples), and our aim here is not to try to interpret all of them. However, despite this diversity, the SR patterns that we studied typically shared a common qualitative feature: the bifurcation structure of the bottleneck trajectories seemingly participates in shaping these SR patterns. Mostly, it seems typically necessary, for SR to hold, that the larger parameter λ_2 has crossed the bifurcation value $\lambda_c(2)$, because the non-zero curvature of the segment $k = 2$ can only be "overcome" by opening the set $Hull\{q_{\lambda_2}(X|t), t \in \mathcal{T}\}$ along a new dimension, through the symbol split at $\lambda_2 = \lambda_c(2)$. This phenomenon will be explored in more details in Section 3.2.

Besides this relationship between SR and the structure of bottleneck bifurcations, this numerical study suggests a generalisation of the notion of successive refinement. Indeed, in Figure 5 for instance, even though the right-hand side asserts that successive refinement does not hold for $\lambda_1 < \lambda_2 < \lambda_c(2)$, the virtually linear piece of trajectory on the left-hand side suggests that this is "almost" the case. In the next section, we formalise this intuition.

3. Soft Successive Refinement of the IB

The minimal experiments from Section 2.3 suggest the intuition that even though successive refinement might not always hold exactly, when broken, it might still be "close" to being satisfied. More generally speaking, let us recall that we are trying here to understand the informationally optimal limits of several-stage information processing. As our numerical experiments suggest that the IB problem is not always successively refinable, it is desirable to *quantify* the lack of successive refinement—i.e., the lack of informational optimality induced by several-stage processing. These considerations lead to the notion of *soft successive refinement* [18], which we define and motivate in this section. As we will see, this generalisation of exact SR does not depend on the specific structure of the IB setting; rather, it can also be used as a generalisation of exact SR for *any* rate-distortion scenario.

3.1. Formalism

Let us first focus on the case $n = 2$: we thus want to quantify the amount of information captured by a coarse bottleneck T_1 and then discarded by a finer bottleneck T_2. Let us recall that, from Proposition 1, bottlenecks T_1 and T_2 achieve successive refinement if there exists an extension $q(X, T_1, T_2)$ of $q_1(X, T_1)$ and $q_2(X, T_2)$ such that, under q, we have the Markov chain $X - T_2 - T_1$, which is equivalent to $I_q(X; T_1|T_2) = 0$. It thus seems natural to quantify soft successive refinement with the conditional mutual information $I_q(X; T_1|T_2)$. However, the IB method does not entirely define the relationship between distinct bottlenecks; formally, there is a whole polytope $\Delta_{q_1,q_2} \subseteq \Delta_{\mathcal{X}\times\mathcal{T}_1\times\mathcal{T}_2}$ of possible extensions $q(X, T_1, T_2)$ of $q_1(X, T_1)$ and $q_2(X, T_2)$ (see Section 1.3). Among these possible extensions, it seems natural to search for those that minimise the violation of the SR condition $I_q(X; T_1|T_2) = 0$. This leads us to use the *unique information* [40]

$$UI(X : T_1 \setminus T_2) := \min_{q \in \Delta_{q_1,q_2}} I_q(X; T_1|T_2). \tag{8}$$

This quantity was already defined in [40] in the context of partial information decomposition [61–64], and it happens to be relevant to us for several reasons.

First of all, it depends only on the distributions $q_1(X, T_1)$ and $q_2(X, T_2)$, which are indeed the only distributions provided by the IB framework. Second, from Proposition 1, there is successive refinement if and only if there are two bottlenecks T_1 and T_2 such that $UI_{q_1,q_2}(X : T_1 \setminus T_2) = 0$. Third, it is thoroughly argued in [40] that (8) is a good measure of the information that only T_1, and not T_2, has about X, which is an interpretation that coincides neatly with the intuition that we want to operationalise here. Eventually, Proposition 6 below, which first requires some definitions, provides an information-geometric justification.

Definition 7. *For Δ a probability simplex and $E_1, E_2 \subseteq \Delta$, we define*

$$D_{KL}(E_1||E_2) := \inf_{r_1 \in E_1, r_2 \in E_2} D_{KL}(r_1||r_2),$$

where D_{KL} is the Kullback–Leibler divergence: $D_{KL}(r_1||r_2) := \sum_{a \in \mathcal{A}} r_1(a) \log\left(\frac{r_1(a)}{r_2(a)}\right)$, if the probability distributions r_1 and r_2 are defined on the discrete alphabet $\mathcal{A}$.

Definition 8. *The* successive refinement set $\Delta_{SR,n} \subseteq \Delta_{\mathcal{X} \times \mathcal{T}_1 \times \cdots \times \mathcal{T}_n}$ *is the set of distributions r on $\mathcal{X} \times \mathcal{T}_1 \times \cdots \times \mathcal{T}_n$ such that, under r, the Markov chain $X - T_n - \cdots - T_1$ holds.*

Note that $\Delta_{SR,n}$ does not require its elements to be extensions of any fixed bottleneck distributions $q_i(X, T_i)$ but imposes the Markov chain that characterises SR (see Proposition 1). SR is achieved for bottlenecks $q_1(X, T_1), \ldots, q_n(X, T_n)$ if and only if the successive refinement set $\Delta_{SR,n}$ and the extension set $\Delta_{q_1,\ldots,q_n}$ share a common distribution $q \in \Delta_{SR,n} \cap \Delta_{q_1,\ldots,q_n}$. In general (for $n = 2$), the following proposition can easily be derived:

Proposition 6. *For fixed distributions $q_1 = q_1(X, T_1)$, $q_2 = q_2(X, T_2)$, we have*

$$UI(X : T_1 \setminus T_2) = D_{KL}(\Delta_{q_1,q_2}||\Delta_{SR,2}). \tag{9}$$

Proof. See Appendix C.1. □

In this sense, $UI(X : T_1 \setminus T_2)$ quantifies "how far" the joint distributions extending the bottlenecks T_1 and T_2 are from making the successive refinement condition $X - T_2 - T_1$ hold true, where the "distance" is understood as a minimised Kullback–Leibler divergence.

Our new measure of soft SR is continuous:

Proposition 7 ([73], Property P.7)**.** *The unique information $UI(X : T_1 \setminus T_2)$ is a continuous function of the probabilities $q_1(X, T_1)$ and $q_2(X, T_2)$.*

Remark 3. *In particular, if $UI(X : T_1 \setminus T_2)$ has a discontinuity as a function of the parameter λ_1 or λ_2, which define the bottleneck distribution $q_{\lambda_1}(X, T_1)$ or $q_{\lambda_2}(X, T_2)$, respectively, then this can only be a consequence of a discontinuity of the probability $q_{\lambda_1}(X, T_1)$ as a function of λ_1 or $q_{\lambda_2}(X, T_2)$ as a function of λ_2, itself, respectively. This consideration will be useful for analysing our numerical experiments in Section 3.2.*

Moreover, the formulation (9) of unique information suggests a natural generalisation to an arbitrary number of processing stages:

Definition 9. *Let $T_1, \ldots, T_n$ be bottlenecks with respective parameters $\lambda_1 < \cdots < \lambda_n$, and $q_i(X, T_i)$ their respective individual distributions. One can quantify* soft successive refinement, *or, equivalently, the* lack of successive refinement, *through the divergence $D_{KL}(\Delta_{q_1,\ldots,q_n}||\Delta_{SR,n})$.*

While [74] proposes a provably convergent algorithm to compute $UI(X : T_1 \setminus T_2)$, to the best of our knowledge, there currently exists no provably convergent algorithm to compute $D_{KL}(\Delta_{q_1,\ldots,q_n}||\Delta_{SR,n})$ for $n > 2$. Our numerical investigations (see Section 3.2) will stick to the case $n = 2$, but this generalisation makes soft SR in particular, at least conceptually for now, more relevant to deep learning (see Section 4.2).

For the sake of completeness, let us point out that for each λ, there is a whole set of solutions $q_\lambda(T|X)$—or, equivalently, $q_\lambda(X, T)$—to the IB problem (1). Thus, the unique information, which is defined as a function of specific bottleneck distributions $q_1(X, T_1)$ and $q_2(X, T_2)$, could *a priori* not be uniquely defined by the corresponding trade-off parameters λ_1 and λ_2. This subtlety is further explained in Appendix D, where we also formulate a conjecture that would prove that, at least in the case of a strictly concave information curve, the trade-off parameters do uniquely define the unique information.

3.2. Numerical Results on Minimal Examples

A provably convergent algorithm that computes, in the discrete case, the unique information (8), was provided in [74]. In this section, we use the authors' implementation of this algorithm (https://github.com/infodeco/computeUI, accessed on 12 September 2023) to qualitatively investigate, on minimal examples, the landscapes of unique information (UI) and their relationship to the bottleneck trajectories on the simplex.

In Figures 7–9 (left), we plot again the same bottlenecks trajectories as in Figures 4–6 (left), but compare them this time with the unique information $UI(X : T_1 \setminus T_2)$, plotted as a function of λ_1 and λ_2 (right). We also plot, in Figures 10–12, some representative examples of the exact SR patterns (left) and UI landscapes (right) for slightly larger source and relevancy cardinalities, where $p(X, Y)$ is, as above, uniformly sampled — the explicit distributions $p(X, Y)$ corresponding to Figures 10–12 can be found at https://gitlab.com/uh-adapsys/successive-refinement-ib/. (see Appendix F for additional examples of comparison of the UI landscapes with bottleneck trajectories, and with the exact SR patterns.) Once again, we highlight with dashed white vertical and horizontal lines the critical parameter values $\lambda_1 = \lambda_c(i)$ and $\lambda_2 = \lambda_c(i)$, respectively, where, as expected, $\lambda_c(1) \approx 0$. We will first describe, for a fixed $p(X, Y)$, the relative variations in unique information as a function of λ_1 and λ_2. Then, we will compare the absolute values of unique information to the information globally processed by the system.

For all Figures from Figures 7–9, the UI landscape partly mirrors the respective exact SR pattern of Figures 4–6 (right). However, within the region where these latter figures answered a binary "no" to the question of exact SR, Figures 7–9 reveal a sharply uneven variation in the violation of SR, where, for important ranges of trade-off parameters, the unique information is negligible comparative to others. For instance, even though Figure 5 (right) seems to indicate that SR does not hold for $\lambda_1 < \lambda_2 < \lambda_c(2)$, the corresponding UI in Figure 8 (right) is virtually zero on a large part of this set of parameters, while still peaking for λ_2 close to $\lambda_c(2)$. This richer structure of the unique information landscape is further evidenced by Figures 10–12.

Moreover, the unique information landscapes seem shaped by the bottleneck trajectories. Most importantly, the influence of IB bifurcations on SR can be seen even more clearly with soft than with exact SR. In particular, in Figures 10–12, it seems that along the lines where one of the trade-off parameters crosses a critical value, the UI often goes through discontinuities, or at least sharp variations in either λ_1, λ_2, or both directions. In particular, even though patterns widely vary across different example distributions $p(X, Y)$, unique information can significantly *drop* when λ_2 crosses a critical value from below—a feature observed in both shown and non-shown examples. As we know that the unique information is continuous, the apparent discontinuity should be one of the bottleneck probability $q_{\lambda_2}(X, T_2)$ itself (see Proposition 7 and Remark 3). This is consistent with the observation from Section 2.3 that, at symbol splits, the trajectory of $q_\lambda(X|T)$ often seems to go through a discontinuity. Further, the fact that the sharp variation in UI is a *decrease* in this quantity (in increasing order of λ_2) is intuitively consistent with the fact that the bottleneck trajectory's discontinuity often induces a sudden "widening" (in increasing order of λ) of

$$\mathcal{H}_T := Hull\{q(X|t),\, t \in \mathcal{T}\}.$$

Indeed, for fixed λ_1, when λ_2 crosses a critical value from below, the corresponding symbol split means that $\mathcal{H}_{T_2}$ "widens" by opening up a new dimension, so it "more easily" encompasses $\mathcal{H}_{T_1}$, yielding as a consequence a drop in unique information. Recalling our intuition (see Section 2.2) that $\mathcal{H}_T$ describes the information content that a bottleneck T contains about the source X, the feature just described can be interpreted in the following way: the IB bifurcations seem to induce a sudden "expansion" (in increasing order of λ) of the information content carried by the bottleneck about the source, which makes the latter's content more easily contain the information content of coarser bottlenecks.

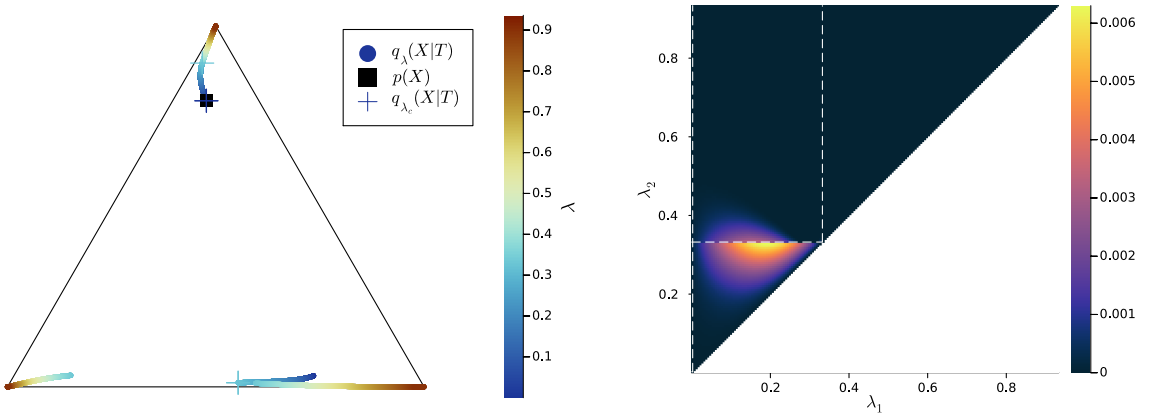

Figure 7. Left: example trajectory of $q_\lambda(X|T)$ as a function of $\lambda = I(X;T)$ (crosses: value of $q_{\lambda_c}(X|T)$ just before a symbol split at a critical parameter λ_c). Right: corresponding unique information, in bits (color), expressed as a function of the pair of trade-off parameters (white dashed lines indicate critical values $\lambda_c(i)$ of either λ_1 or λ_2.). For instance, the critical value $\lambda_c(2) \approx 0.33$ (right) corresponds, on the bottleneck trajectories (left), to the symbol split from two to three symbols (cyan crosses). The respective $p(Y|X)$ corresponding to this figure and to Figures 8 and 9 are plotted in Appendix E.

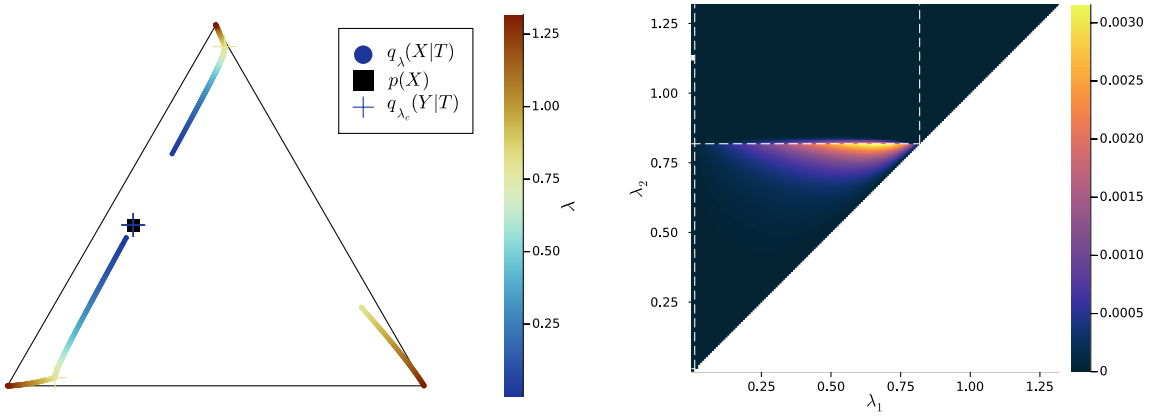

Figure 8. Same as Figure 7, where the example distribution $p(X,Y)$ is that of Figure 5.

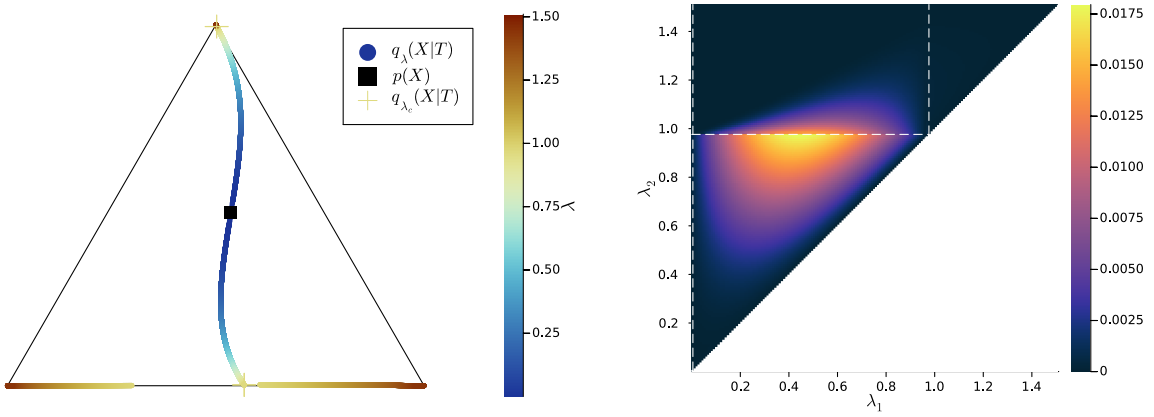

Figure 9. Same as Figure 7, where the example distribution $p(X,Y)$ is that of Figure 6.

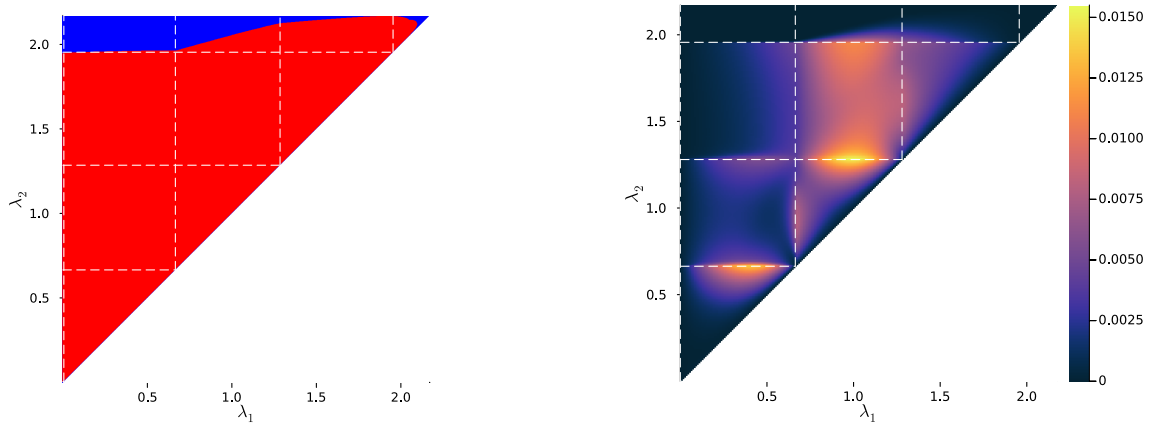

Figure 10. New example of an exact SR pattern and the corresponding UI landscape over trade-off parameters $\lambda_1 < \lambda_2$, where, here, $|\mathcal{X}| = 5$ and $|\mathcal{Y}| = 3$. Left: exact SR pattern, i.e., output for the convex hull condition (blue: satisfied, red: not satisfied). Right: corresponding UI landscape, in bits (color). White dashed lines indicate critical values $\lambda_c(i)$ of either λ_1 or λ_2. Note that (i) the binary notion of exact SR (left) filters out most of the structure unveiled by UI (right), (ii) the UI landscape seems highly impacted by IB bifurcations, and (iii) the UI is in any case always small, even though not entirely negligible. See main text for more details.

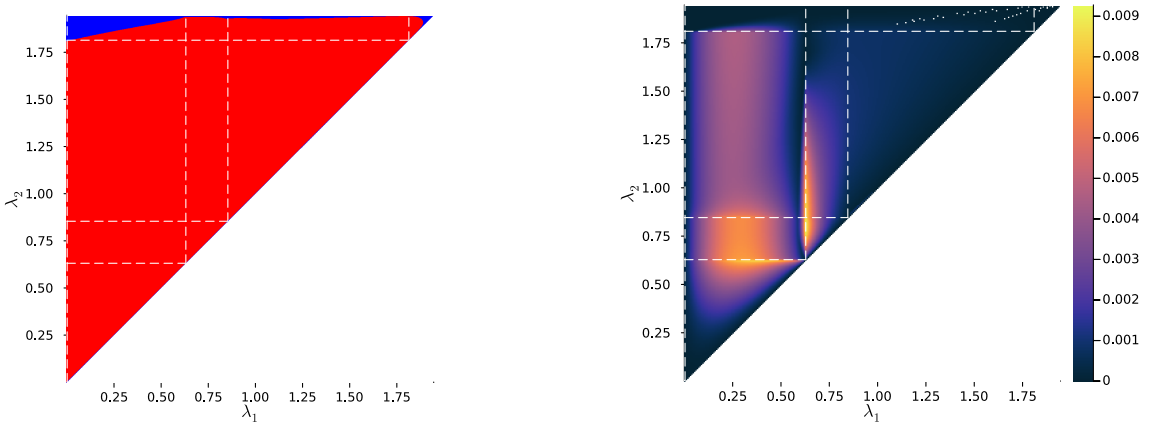

Figure 11. Same as Figure 10, with a new example distribution $p(X, Y)$, where, here, $|\mathcal{X}| = 5$ and $|\mathcal{Y}| = 3$. Besides the white orthogonal dashed lines, other white dots correspond to values of (λ_1, λ_2) for which the algorithm did not converge (see main text for a comment on this lack of convergence).

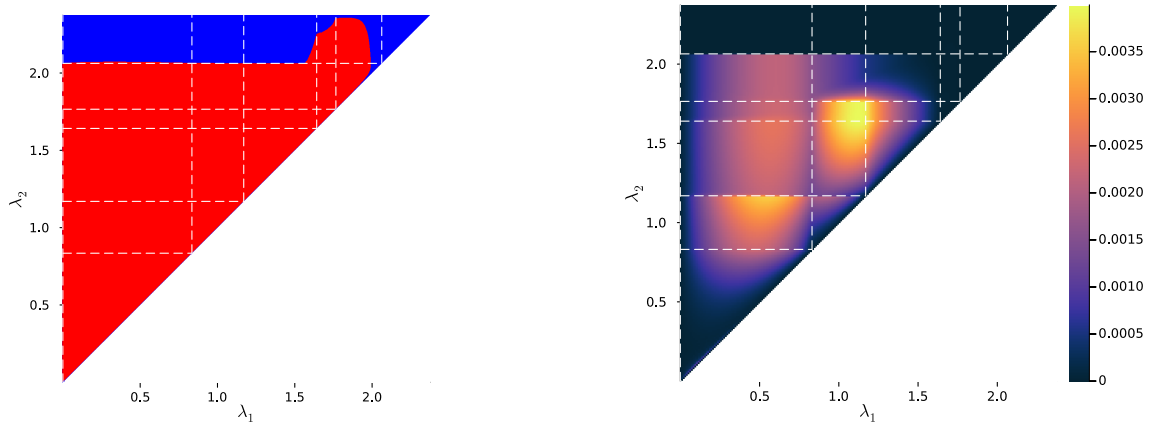

Figure 12. Same as Figure 10, with a new example distribution $p(X, Y)$, where, here, $|\mathcal{X}| = 7$ and $|\mathcal{Y}| = 5$.

Note, however, that these simple numerical results do not allow one to discriminate between the interpretation of the UI's sharp variations at bifurcations as a discontinuity with regard to trade-off parameters, or a discontinuity of the UI's *differential*. For instance, if the derivative with regard to λ_2 discontinuously takes a value close to $-\infty$ for λ_2 slightly larger than some λ_c, then the UI graph can seem discontinuous at finite numerical resolution, even if, formally, only the UI's differential is so. On the other hand, as an example, bifurcations can be characterised precisely as points of discontinuities of the derivatives, with regard to the trade-off parameter, of $I(T;X)$ and $I(T;Y)$ [43,75], even though the functions themselves are continuous [2,75]. A more involved analysis distinguishing discontinuities of UI from those of its differential is left to future work. In any case, the interpretation as a discontinuity of the differential rests on a weaker assumption, which is still sufficient for explaining the numerical results.

More generally, these results suggest that for a several-stage processing that is IB-optimal at each stage, to minimise the information discarded along stages, the trade-off parameters should rather lie close to well-chosen IB bifurcations. If this happens to be a general feature of the IB framework, it would have implications for incremental learning. Indeed, coming back to the modelling of embodied agents (see Section 1), for instance, it would mean that organisms that are poised close to information optimality by evolution should have a very specific structure of developmental learning, where the stages of learning should be discrete and determined by the right trade-off parameters.

Eventually, a last crucial feature was also satisfied on these minimal examples: whatever the structure of bottleneck trajectories, the maximal UI was significantly lower than the mutual information $I(X;T_1,T_2)$ between the external source X and the system's internal representations (T_1,T_2). More precisely, for an extension $q(X,T_1,T_2)$ of $q_{\lambda_1} := q_{\lambda_1}(X,T_1)$ and $q_{\lambda_2} := q_{\lambda_2}(X,T_2)$ that achieves the minimum in (8), let us define

$$\sigma(q_{\lambda_1},q_{\lambda_2}) := \frac{UI_{q_{\lambda_1},q_{\lambda_2}}(X:T_1\setminus T_2)}{I_q(X;T_1,T_2)}.$$

Note that decomposing $I_q(X;T_1,T_2)$, where $q \in \Delta_{q_1,q_2}$, with the chain rule for mutual information shows that this quantity only depends on q_{λ_1} and q_{λ_2}: thus here, $\sigma(q_{\lambda_1},q_{\lambda_2})$ is indeed well-defined by q_{λ_1} and q_{λ_2}. The maximum ratio over all trade-off parameters $\lambda_1 < \lambda_2$ was typically of the order of 1% in our minimal experiments; for instance, it was 1.89%, 0.39%, 1.82%, 2.03%, 1.34%, and 0.31% for the IB problems corresponding to Figures 7–12, respectively. Among all the (shown and non-shown) studied examples, it never exceeded 5.4%, and we did not notice an increase in this maximum ratio when the source or relevancy cardinalities were increased (the largest cardinalities that we experimented with were $|\mathcal{X}| = 20$, $|\mathcal{Y}| = 10$). In short, even though several-stage processing might incur a non-negligible loss of information optimality in the IB sense, these results suggest that this loss could often be significantly limited. Of course, here as in Section 2.3, on the one hand, the numerical results are purely phenomenological, and, on the other, it is at this stage far from being clear that the qualitative insights brought by these minimal experiments generalise well to more complex situations. However, they exhibit the potentially crucial qualitative features of exact and soft successive refinement in the IB framework, which can be targeted by further theoretical research.

4. Alternative Interpretations: Decision Problems and Deep Learning

The notion of successive refinement presented in this work builds on the intuition of the optimal incorporation of information. However, alternative interpretations can be given to the very same mathematical notion. First, thanks to the Sherman–Stein–Blackwell theorem [45,65], the rate-distortion-theoretic notion of SR can be shown to be equivalent to a specific order relation between the encoder of the finer bottleneck $q(T_2|X)$ and that of the coarser one $q(T_1|X)$, namely the *Blackwell order*. This point of view turns SR into an operational *decision-theoretic statement*; in short, there is SR when, for *any* task and *any* source

distribution $p(X)$, the optimal performance is better (or at least as good) when decisions are based on the output of $q(T_2|X)$ than when they are based on the output of $q(T_1|X)$. Second, the Markov chain (4) characterising successive refinement makes it directly relevant [46] to the IB analysis of deep neural networks [49–56]. In the next two sections, we make these connections explicit and relate them to this paper's investigations.

4.1. Successive Refinement, Decision Problems, and Orders on Encoder Channels

Here, we show that exact and soft successive refinement can be, in the discrete case at least, understood in terms of optimally solving decision problems on arbitrary tasks, through orders on the encoder channels $q(T|X)$ (or more precisely, pre-orders: i.e., we will consider binary relations that are reflexive and transitive). We will rely on [45], where these orders were considered.

Let us first make clear what we mean here by a decision problem. Consider a state variable X over a finite set $\mathcal{X}$, another finite set $\mathcal{A}$ of possible actions, and a reward function $u = u(x, a)$ that depends on both the value x of the state X, and the chosen action $a \in \mathcal{A}$. The agent's observation is not the state X itself, but only the output T of X through some stochastic channel $\kappa := p(T|X)$ (where we assume here that the observation space $\mathcal{T}$ is finite). To each observation-dependent policy $\pi = \pi(A|T)$ corresponds an expected reward

$$\mathbb{E}_\pi(u(X, A)) := \sum_t p(t)\mathbb{E}_{(X,A)\sim p(X|t)\pi(A|t)}(u(X, A)),$$

where $p(X|t)$ is determined from $\kappa := p(T|X)$, $p(X)$ through the Bayes rule, and $p(X|t)$ $\pi(A|t)$ denotes the product measure of $p(X|t)$ and $\pi(A|t)$. Solving the decision problem $(p(X), \mathcal{A}, u)$ for the observation channel κ means choosing a policy that yields an optimal expected reward

$$\mathcal{R}(p(X), \kappa, u) := \max_\pi \mathbb{E}_\pi(u(X, A)).$$

For instance, any Markov decision process can be seen as a decision problem as defined above (for discrete time and finite state-space, number of possible actions at each state, and horizon). In this case, $\mathcal{X}$ and $\mathcal{T}$ are the spaces of state trajectories and observation trajectories, respectively, that an agent can go through along one episode; $\mathcal{A}$ is the space of action sequences that can be chosen along the episode; and u is the cumulative reward, i.e., the (potentially discounted) sum of rewards obtained at each time-step in the episode. (See, e.g., [76] for more details on the terminology used in this example.)

We can now define the following order [45]:

Definition 10. *For two channels κ and μ, we write $\kappa \sqsupseteq_{\mathcal{X}} \mu$, if, for any decision problem $(p(X), \mathcal{A}, u)$, we have*

$$\mathcal{R}(p(X), \kappa, u) \geq \mathcal{R}(p(X), \mu, u).$$

In short, $\kappa \sqsupseteq_{\mathcal{X}} \mu$ means that, for any conceivable task based on any data distribution $p(X)$ over the fixed data space $\mathcal{X}$, the observation channel κ can yield a performance at least as good as that of the observation channel μ—if combined with a well-chosen policy. The second order is the *Blackwell order* [65]:

Definition 11. *For two channels κ and μ, we write $\kappa \sqsupseteq'_{\mathcal{X}} \mu$ if there exists a channel η such that $\mu = \eta \circ \kappa$, where "$\circ$" denotes the composition of channels, i.e., such that $M_\mu = M_\eta M_\kappa$, where M_μ, M_η, and M_κ are the column transition matrices corresponding to μ, η, and κ.*

It turns out that successive refinement can be characterised by either of these two orders, thanks to the Sherman–Stein–Blackwell theorem [45,65]. In other words, SR, which is *a priori* not a decision-theoretic statement, turns into one through its equivalence with the Blackwell order:

Proposition 8. *Let $0 < \lambda_1 < \lambda_2$. The following are equivalent:*

(i) *There is successive refinement for parameters (λ_1, λ_2).*
(ii) *There are bottlenecks T_1, T_2 of respective parameters λ_1, λ_2 such that*

$$q(T_2|X) \sqsupseteq_{\mathcal{X}} q(T_1|X).$$

(iii) *There are bottlenecks T_1, T_2 of respective parameters λ_1, λ_2 such that*

$$q(T_2|X) \sqsupseteq'_{\mathcal{X}} q(T_1|X).$$

Proof. Using the Markov chain characterisation (point (ii) in Proposition 1), the result is nothing more than a reformulation of Theorem 4 in [45] in the language of the present paper. Note that, to use this theorem, we need to assume that the source X is fully supported, but this is indeed an assumption that we are using along the whole paper because it does not incur any loss of generality (see Section 1.3). □

Let us highlight the intuitive meaning of Proposition 8. Point (ii) means that there is SR when the coarse representation T_1 can be retrieved by post-processing the finer representation T_2—which has implications in terms of feed-forward processing (see Section 4.2).

Now, the equivalence of SR with point (iii) relies on the mathematically deep part of the Sherman–Stein–Blackwell theorem [45], and provides a new operational meaning to SR. Namely, there is SR when, for *any* distribution $p(X)$ on the source, and *any* reward function, the optimal performance is at least as good when the decisions are based on the output of $q(T_2|X)$, seen as an observation channel, than when they are based on the output of $q(T_1|X)$. Let us stress that the fact that $q(T_2|X)$ defines a finer bottleneck than $q(T_1|X)$ crucially depends on $p(X, Y)$, i.e., on the specific source distribution $p(X)$, and on how the latter relates to the specific relevancy variable through $p(Y|X)$. Proposition 8 shows that SR describes a much more "universal" relation between the channels $q(T_1|X)$ and $q(T_2|X)$.

For example, assume that evolution poises the sensors of a given biological organism at optimality in the IB sense [10,16], i.e., if X is the environment, S some sensor's output (e.g., a retina's ganglion cells activation), and Y a behaviourally relevant feature (e.g., the edibility of food), then S is a bottleneck for $p(X, Y)$. Successive refinement here means that if the sensor S_2 is finer than S_1 as a bottleneck for the fixed feature Y relevant to a particular task, then S_2 will afford to the organism—if combined with the right decision making—better performances than S_1 on any other task, for any other input distribution $p(X)$. In other words, S_2 is then "universally better" than S_1, which is a very strong (and somewhat unexpected) generalisation.

Eventually, the unique information that we chose as our measure of soft SR has initially been thought precisely as measuring the deviation from the order "$\sqsupseteq_{\mathcal{X}}$" (see arguments in [45]). Unique information can thus, for instance, be understood as quantifying the deviation from a finer IB-optimal sensor to be "universally better" than a coarser one.

4.2. Successive Refinement and Deep Learning

As suggested by Remark 1 and Proposition 8-(ii), successive refinement can be equally well understood in terms of feed-forward processing, an interpretation which is particularly relevant to deep neural networks. Indeed, while the information bottleneck theory of deep learning [49–51] is still under debate [52–56], our results can be connected to some of this theory's specific claims concerning the benefits of hidden and output layers' IB-optimality.

Let $L_1, \ldots, L_n$ denote the successive layers of a feed-forward deep neural network (DNN), which is fed with an input X and attempts to extract, within it, information about a target variable Y,thus satisfying the Markov chain [49]

$$Y - X - L_1 - \cdots - L_n. \tag{10}$$

One of the claims of the IB theory of DNNs [49–51] is that, once converged, a DNN's hidden and output layers lie close to the information curve of the IB problem defined by $p(X, Y)$, with each new layer corresponding to a coarser trade-off parameter. The performance and generalisation abilities of DNNs would rely on this IB-optimality of networks after training. While these claims have been challenged [52,77], the identified caveats have sparked a still ongoing line of research [54–56], which suggests that more nuanced versions of the initial claims might still hold. Most importantly for us here, numerical results suggest that layer-by-layer training with the IB Lagrangian as the loss function induces a performance on par with end-to-end training with cross-entropy loss [54], while recent theoretical work proved that the IB trade-off optimises a bound on the generalisation error [56]. In other words, the IB method seems to be relevant at least as a normative, if not descriptive, framework for DNNs. Thus, an interesting informationally optimal limit to compare a given DNN to is a sequence of variables $L_1, \dots, L_n$ that

(i) Satisfy the Markov chain (10); and
(ii) Are each bottlenecks with source X and relevancy Y, for respective trade-off parameters $\lambda_1 > \cdots > \lambda_n$.

However, it is not clear that variables satisfying those conditions even exist; actually, it is the case if and only if the IB problem is $(\lambda_n, \dots, \lambda_1)$-successively refinable. Indeed, points (i) and (ii) are exactly the conditions of point (iii) in Proposition 1, with $T_i := L_{n-i}$, and the order of trade-off parameters reversed as well. In this sense, the notion of exact successive refinement is relevant to deep learning; in particular—as suggested by the numerical results from Section 2.3—it might well be the case that there is successive refinement only for well-chosen combinations of trade-off parameters. In this case, an IB-optimal DNN should be designed and trained in such a way that its successive layers implement a compression corresponding to these well-chosen trade-off parameters.

Remark 4. *The single-letter formulation above mirrors, in large part, the asymptotic coding version of [46]. More precisely, Ref. [46] defines in asymptotic coding terms a feed-forward processing pipeline where each layer tries to extract, from the input coming from the previous layer, information about a potentially distinct relevancy Y_i. Theorem 2 in [46] shows that, for constant relevancy $Y_i := Y$, the notion of "successive refinement" defined there by the authors happens to be equivalent to points (i) and (ii) above, and thus to our notion of "successive refinement". In particular, the deep learning interpretation presented in this section also has an operational formulation in terms of asymptotic coding.*

Now, if exact SR describes the situation where each layer of a DNN can potentially reach the information curve, is our notion of soft SR also relevant to deep learning? Note that, here,

- We know that the variables $L_1, \dots, L_n$ must satisfy $X - L_1 - \cdots - L_n$, i.e., we know that the joint distribution $q := q(X, L_n, \dots, L_1)$ must be in Δ_{SR};
- And we want to know "how close" we can choose this joint distribution q to one whose marginals $q(X, L_1), \dots, q(X, L_n)$ coincide with bottleneck distributions $q_1 := q_1(X, T_1), \dots, q_n := q_n(X, T_n)$, respectively, of parameters $\lambda_1 > \cdots > \lambda_n$, respectively, i.e., we want to know how close we can choose q to the set $\Delta_{q_1,\dots,q_n}$.

Thus, the quantity $D_{KL}(\Delta_{q_1,\dots,q_n} || \Delta_{SR,n})$ can also be interpreted as a measure of the deficiency of a DNN's layers from all those simultaneously being bottlenecks. Note, however, that, in previous sections, we knew that any joint distribution $q(X, T_1, \dots, T_n)$ had to be in the extension set $\Delta_{q_1,\dots,q_n}$, and wanted to know "how close" to the successive refinement set $\Delta_{SR,n}$, in the KL sense, we could choose it. On the contrary, in the case of DNNs, we know that any $q(X, T_1, \dots, T_n)$ must be in $\Delta_{SR,n}$—because the bottlenecks correspond to a DNN's layer—and want to know "how close" to $\Delta_{q_1,\dots,q_n}$ we can choose it.

From this perspective, the numerical results of Section 3.2 suggest interesting properties, or at least desirable features, of DNNs. First, if the fact that the UI is typically low

generalises well from our minimal investigation to the much richer deep learning setting, this would imply that even in situations where a DNN's successive layers cannot all lie exactly along the information curve, they might still be able to remain reasonably close to it. Second, the fact that UI (or its differential) seems to go through a discontinuity close to well-chosen bifurcations—such that the UI sharply drops when λ_2 crosses the bifurcation from below—suggests that, for each layer of the DNN to be individually as IB-optimal as possible, their corresponding trade-off parameters should each lie close to these IB bifurcations. This resonates with previous considerations suggesting that DNNs' hidden layers should [49] or might indeed do [50] lie at IB bifurcations.

5. Limitations and Future Work

Our convex hull characterisation intertwines the question of exact SR with the more fundamental question of the structure of decoder curves

$$\big\{\big(\lambda \mapsto q_\lambda(X|t)\big),\ t \in \mathcal{T}\big\} \tag{11}$$

on the source simplex $\Delta_{\mathcal{X}}$, a question for which the convexity approach to the IB problem [35–39] seems promising. In short, this approach reformulates the IB problem to that of finding the lower convex envelope of a well-chosen function F_β, defined on the source simplex $\Delta_{\mathcal{X}}$, and parameterised by the information curve's inverse slope β (see Appendix B.7). More precisely, bottlenecks are essentially characterised by the fact that the lower convex envelope must be achieved by convex combinations of the points $F_\beta(q(X|t))$; this approach thus provides analytical tools for proving key properties of the set of trajectories (11), which would then have consequences for SR through the convex hull condition. Despite the limited scope of the result itself, the proof of Proposition 5 gives an example of such a fruitful interaction, thus suggesting a way forward for further theoretical progress. As a first step, one could try to use the convexity approach to the IB to prove our Conjecture 1 about the unicity, up to permutations and injectivity of $q(X|T)$, for canonical bottlenecks and the strictly concave information curve. This would both simplify our convex hull characterisation of SR for the case of the strictly concave information curve (see Appendix D) and provide in itself a crucial property of the curves (11). Generally speaking, leveraging, through our convex hull characterisation, the convexity approach to the IB problem might allow one to (i) identify new wholly refinable IB problems, but also (ii) produce general methods to identify, for a given distribution $p(X,Y)$, the combination of parameters for which exact SR holds.

It must be stressed that even though we motivate the successive refinement of the IB by diverse scientific questions in Sections 2 and 4, in this work, we do not model any concrete system. Rather, our minimal numerical experiments target the qualitative exploration of the formalised problem. Our results might in turn be relevant for future modelling work (see the last paragraph of this section), but the most pressing aspect is to first develop further the theoretical and computational framework. In particular, it seems important to describe formally the apparent discontinuity of UI (or its differential) as a function of the trade-off parameters λ_1 and λ_2 at IB bifurcations (through that of the $q_\lambda(X,T)$ as functions of λ); to describe more formally why the UI tends to peak and then drop close to IB bifurcations; to provide global bounds on UI in general or as functions of the source and relevancy distribution $p(X,Y)$; or to make formal the informal relationship between the "extent to which" the convex hull condition is broken, and variations in UI. Another interesting contribution would be to provide an asymptotic coding interpretation to unique information; indeed, the deviation from successive refinement is more classically quantified as a difference between asymptotic rates or distortions (see, e.g., [60]), and it is not clear whether or not this interpretation can be made for UI. Numerically speaking, one could design algorithms allowing for the computation of UI for continuous $p(X,Y)$ and/or more than two processing stages. Indeed, the algorithm from [74] only encompasses the case of discrete variables and two processing stages. One could, for instance, take inspiration

from [74] to formulate the quantity $D_{KL}(\Delta_{q_1,\dots,q_n}||\Delta_{SR,n})$ as a double minimisation problem over separate parameters, allowing for an alternating optimisation algorithm.

The deep learning interpretation of (exact and soft) SR depends crucially on some aspects of the ongoing debate on the IB theory of deep learning [49–52,54–56]. In this regard, it would be interesting to directly measure the unique information between different layers of a DNN or determine whether or not having the layers lying close to IB bifurcations does induce better performance or generalisation capabilities.

Let us point out that our framework considers that the source of information X and the target variable Y are the same along all processing stages. More general frameworks could allow for variations in either the source of information (as in the case in temporal series) or the target variable (as is the case in transfer learning). Frameworks for both these kinds of extensions have already been proposed [46,78], and it would be interesting to study if, in these cases as well, the specific nature of the IB problem imprints the informationally optimal limits of several-stage processing.

Eventually, we deem the interpretation in terms of the incorporation of information to be particularly relevant to modelling adaptive behaviour. For instance, for a given developmental or skill-learning problem on a given task, our framework could help in distinguishing situations where the choice of the successive representations' complexity along incrementally learning the task does not matter (i.e., when there is successive refinement) from situations where these complexities must be minutely weighed, so as to avoid as much as possible the "waste" of cognitive work along the way (i.e., when the unique information is not negligible and unevenly distributed). In the latter case, our framework, once mature, might precisely describe those sequences of representations' complexity that minimise the "waste" of cognitive work from one learning stage to another, thereby potentially identifying key stages of skill or developmental learning. Future work should keep in mind the horizon of identifying such qualitative features and producing measures capturing the relevant phenomena for experimental research in these areas.

6. Conclusions

Our approach in this paper is three-fold: to bring together in a common framework existing work on the exact successive refinement of the IB and related topics; to develop further this common framework, particularly through a geometric approach to the problem; and to then open up a line of research on the soft successive refinement of the IB.

The formal unity that we make explicit in this paper is mainly that between these three scientific questions: (i) that of informationally optimal incorporation of information—relevant in particular to developmental and skill learning; (ii) that of informationally optimal feed-forward processing—relevant in particular to describing and designing deep neural networks (DNNs); and (iii) that of channel order in statistical decision theory—which provides clear interpretations of distinct bottlenecks' comparison in terms of universal informativeness of an agent's sensor. Indeed, while we focused for most of the paper on the information incorporation interpretation, we saw in Section 4 that the two other ones are as legitimate as the first one.

Once the formal problem is motivated and set, we turn to the mathematical analysis of it. We first note that, for jointly Gaussian vectors (X, Y) or for deterministic $p(Y|X)$, successive refinability can be easily drawn from existing IB literature [33,34]. Then, we propose a new geometric characterisation of SR, which builds on the intuition that what is "known" by a bottleneck is the convex hull of its decoder conditional probabilities. This new point of view, associated with an active approach that reformulates the IB problem as that of finding the lower convex envelope of a well-chosen function [35–39], provides a new tool for theoretical research on this topic. We exemplify this potential fertility by proving, thanks to the combination of our convex hull characterisation with the convexity approach to the IB, the successive refinability of binary source X and binary relevancy Y (Proposition 5). This convex hull characterisation also allows one to numerically investigate SR with a linear program, which can be helpful for computational studies on this topic.

Our own minimal numerical experiments suggest that (i) successive refinement does not always hold for the IB, (ii) the successive refinement patterns are shaped by IB bifurcations, and (iii) even when successive refinement seems to break, sometimes it is "close" to being satisfied, in the sense of the convex hull condition being only "slightly" violated.

To formalise this latter intuition, we propose to soften the traditional notion of SR into a *quantification* of the loss of information optimality incurred by several-stage processing. For that purpose, we call on the measure of unique information (UI) used in [40]. Intuitively, this quantity measures the information that only the coarser bottleneck T_1, and not the finer one T_2, holds about the source X, and it can be generalised to an arbitrary number of processing stages. Our minimal experiments, in the case of two processing stages, unveil a rich structure of soft SR that was partially hidden by exact SR, which only makes the distinction between vanishing UI (if there is SR) and positive UI (if there is no SR). Even though the UI landscapes depend strongly on the distribution $p(X, Y)$ that defines the IB problem, some qualitative features seem to emerge: (i) the "more" the convex hull condition is broken, the higher the unique information; (ii) the IB bifurcations crucially shape the UI landscape, with sharp decreases in unique information in particular when the finer trade-off parameter λ_2 crosses a bifurcation critical value; and (iii) in any case, this violation of successive refinement seems to always be mild compared to the system's globally processed information.

The features exhibited by these numerical experiments offer a "first outlook" of potentially general properties of exact and soft successive refinement for the IB problem, thus providing a guide for future theoretical research. These potential properties might provide interesting perspectives on the scientific questions that motivate the formalism, particularly in terms of the incorporation of fresh information into already learned models, and deep learning. For instance, the apparently important role of bifurcations in exact and soft successive refinement suggests that informationally optimal several-stage learning or processing should ideally be organised along well-chosen "checkpoints" on the information plane. Moreover, if the loss of information optimality induced by this sequential processing is indeed typically low (even though not entirely negligible) for the IB framework, this could be taken as an indication that incremental learning might be made highly efficient. These potential features thus provide a strong incentive to bring the formal framework presented here closer to maturity—for instance, along the lines of research proposed in Section 5.

Author Contributions: Conceptualisation, H.C., N.C.V. and D.P.; methodology, H.C., N.C.V. and D.P.; software, H.C.; validation, H.C., N.C.V. and D.P.; formal analysis, H.C., N.C.V. and D.P.; investigation, H.C., N.C.V. and D.P.; resources, N.C.V. and D.P.; writing—original draft preparation, H.C.; writing—review and editing, H.C., N.C.V. and D.P.; visualisation, H.C.; supervision, N.C.V. and D.P.; project administration, D.P.; funding acquisition, D.P. All authors have read and agreed to the published version of the manuscript.

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

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The code that we used for this work can be found at https://gitlab.com/uh-adapsys/successive-refinement-ib/, along with the explicit values of the example distributions $p(X, Y)$ that we used to generate our figures.

Acknowledgments: Thanks to Johannes Rauh and Pradeep Banerjee for insightful comments on unique information [40] and the iterative algorithm proposed to compute it in [74].

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

IB	Information Bottleneck
SR	Successive Refinement
UI	Unique Information
DNN	Deep Neural Network

Appendix A. Section 1 Details

Appendix A.1. Effective Cardinality

In [43], the effective cardinality of a Lagrangian bottleneck T is defined as the number of distinct $q(Y|t)$, whereas it is defined as that of distinct $q(X|t)$ in our Definition 1. However, as mentioned in Section 1.3, both choices happen to be equivalent:

Proposition A1. *Let $q(T|X)$ be a fixed solution to the Lagragian IB problem (3) for some β, where $p(X,Y)$ is discrete. The number of distinct $q(X|t)$ and that of distinct $q(Y|t)$ are equal.*

Proof. It is proven in [1] that a solution to the Lagrangian IB must satisfy for all $t \in \mathcal{T}$, $x \in \mathcal{X}$ the self-consistent equation

$$q(t|x) \;=\; \frac{q(t)}{Z(x)} \exp\big(-\beta\, \mathrm{D_{KL}}(p(Y|x)\,||\,q(Y|t))\big), \tag{A1}$$

where $Z(x)$ is the normalisation factor, and

$$q(t) := \sum_x q(t|x)p(x), \tag{A2}$$

$$q(y|t) := \sum_x p(y|x)q(x|t), \tag{A3}$$

with

$$q(x|t) := \frac{q(t|x)p(x)}{q(t)}. \tag{A4}$$

If $q(Y|t_1) = q(Y|t_2)$, then (A1) implies that $q(t_1|X) = q(t_2|X)$, which, combined with (A2), implies that we also have $q(t_1) = q(t_2)$. These two new equalities, combined with (A4), then prove that $q(X|t_1) = q(X|t_2)$. Conversely, if $q(X|t_1) = q(X|t_2)$, then Equation (A3) proves that $q(Y|t_1) = q(Y|t_2)$. □

Crucially, it is proven in [43] that a given Lagrangian bottleneck T can be reduced to effective cardinality while still being a bottleneck for the same trade-off parameter β by merging all bottleneck symbols $t_1 \ldots, t_r$ with equal decoder distributions $q(X|t_1) = \cdots = q(X|t_r)$ into a new symbol $[t]$ defined by

$$q([t]|x) := \sum_{i=1}^{r} q(t_i|x).$$

Moreover, this merging can also be carried out for primal bottlenecks (this is not a direct consequence of Proposition A1, as it is known that when the information curve is not strictly concave, the primal and Lagrangian problems might not be exactly equivalent [39,67]).

Proposition A2. *Let T be a primal bottleneck of parameter λ, i.e., a solution to (1), where $p(X,Y)$ is discrete. The bottleneck obtained from T by merging the symbols t with identical $q(X|t)$ is still a solution to (A38) for the same parameter λ.*

Proof. We will use the following reparametrisation of the IB problem (1) (see Section 2.2):

$$\underset{\substack{(q(T),q(X|T)):\\ \sum_t q(t)q(X|t)=p(X)\\ T-X-Y,\ I(X;T)\leq\lambda}}{\arg\max} \quad I(Y;T). \tag{A5}$$

We thus consider the bottleneck T from Proposition A2's statement as defined by a pair $(q(T), q(X|T))$ satisfying $\sum_t q(t)q(X|t) = p(X)$. Now, assume that there exist $t_1, t_2 \in \mathcal{T}$ such that $q(X|t_1) = q(X|t_2)$. Then,

$$\sum_x q(x|t_1)\log\left(\frac{q(x|t_1)}{p(x)}\right) = \sum_x q(x|t_2)\log\left(\frac{q(x|t_2)}{p(x)}\right),$$

so that

$$\begin{aligned} I(X;T) &= \sum_{t,x} q(t)q(x|t)\log\left(\frac{q(x|t)}{p(x)}\right) \\ &= \alpha_{t_1,t_2}\sum_x q(x|t_1)\log\left(\frac{q(x|t_1)}{p(x)}\right) + \sum_{t\notin\{t_1,t_2\},x} q(t)q(x|t)\log\left(\frac{q(x|t)}{p(x)}\right), \end{aligned} \tag{A6}$$

where $\alpha_{t_1,t_2} := q(t_1) + q(t_2)$. Moreover,

$$q(Y|t_1) = \sum_x q(x|t_1)p(y|x) = \sum_x q(x|t_2)p(y|x) = q(Y|t_2),$$

so that, similarly,

$$I(Y;T) = \alpha_{t_1,t_2}\sum_y q(y|t_1)\log\left(\frac{q(y|t_1)}{p(y)}\right) + \sum_{t\notin\{t_1,t_2\},y} q(t)q(y|t)\log\left(\frac{q(y|t)}{p(y)}\right). \tag{A7}$$

Eventually,

$$p(X) = \sum_t q(t)q(X|t) = \alpha_{t_1,t_2}q(X|t_1) + \sum_{t\notin\{t_1,t_2\}} q(t)q(X|t), \tag{A8}$$

where the first equality comes from the fact that $(q(T), q(X|T))$ is a solution to (A5) (so, in particular, it must satisfy the hard constraints required in the optimisation problem). Let us define the bottleneck $\tilde{T}$ on $\tilde{\mathcal{T}} := \mathcal{T} \setminus t_2$ by

$$\tilde{q}(t) := \begin{cases} \alpha_{t_1,t_2} & \text{if } t = t_1 \\ q(t) & \text{if } t \in \mathcal{T}\setminus\{t_1,t_2\}, \end{cases}$$

and, for all $t \in \mathcal{T} \setminus \{t_2\}$,

$$\tilde{q}(X|t) := q(X|t).$$

The last line of (A6) can then be rewritten as

$$\begin{aligned} I(X;T) &= \tilde{q}(t_1)\sum_x \tilde{q}(x|t_1)\log\left(\frac{\tilde{q}(x|t_1)}{p(x)}\right) + \sum_{t\notin\{t_1,t_2\},x}\tilde{q}(t)\tilde{q}(x|t)\log\left(\frac{\tilde{q}(x|t)}{p(x)}\right) \\ &= \sum_{t\in\tilde{\mathcal{T}}}\sum_{x\in\mathcal{X}}\tilde{q}(t)\tilde{q}(x|t)\log\left(\frac{\tilde{q}(x|t)}{p(x)}\right) \\ &= I(X,\tilde{T}). \end{aligned}$$

Similarly, from (A7), we obtain $I(Y;\tilde{T}) = I(Y;T)$, while from (A8), we have $\sum_t \tilde{q}(t)\tilde{q}(X|t) = p(X)$. In other words, $\tilde{T}$ is also a solution to the reparametrised primal bottleneck problem

(A5), for the same parameter λ. Moreover, it is clear that our definition of $(\tilde{q}(T), \tilde{q}(X|T))$ implies that $\tilde{q}(t_1|x) = q(t_1|x) + q(t_2|x)$ and $\tilde{q}(t|x) = q(t|x)$ for $t \in \mathcal{T} \setminus \{t_1, t_2\}$, so $\tilde{T}$ is the variable obtained from T by merging the symbols t_1 and t_2. The result follows by iterating this argument until all the $\tilde{q}(X|t)$ are distinct. □

Appendix B. Section 2 Details

Appendix B.1. Proof of Proposition 1

$(i) \Rightarrow (ii)$: Suppose that there are variables T_1 and $T_i := (T_{i-1}, S_i)$ for $2 \leq i \leq n$ such that each T_i is a bottleneck with parameter λ_i. Unrolling the iterative definitions of the T_i, we obtain

$$T_i = (T_1, S_2, \ldots, S_i),$$

which implies that, if $j < i$, then T_j is a deterministic function of T_i; in other words, given T_i, the variable T_j is independent of any other variable. So, first, we have $X - T_n - T_{n-1}$. Now, assume that for a given i, we have

$$X - T_n - \cdots - T_i. \tag{A9}$$

Given T_i, the variable T_{i-1} is independent of any other variable, so, in particular,

$$(X, T_n, \ldots, T_{i+1}) - T_i - T_{i-1}. \tag{A10}$$

The Markov chains (A9) and (A10) together imply that

$$X - T_n - \cdots - T_{i-1}.$$

Thus, a recurrence from $i = n$ to $i = 1$ proves that we do have $X - T_n - \cdots - T_1$, where, by assumption, each T_i is indeed a bottleneck of parameter λ_i.

$(iii) \Rightarrow (i)$: For all i, the Markov chain (5) implies that

$$I(X; T_i) = I(X; T_i'),$$
$$I(Y; T_i) = I(Y; T_i'),$$

where $T_i' := (T_i, \ldots, T_1)$. The Markov chain (5) also implies that these T_i' satisfy $Y - X - T_i'$. Thus, the T_i' are also bottlenecks with respective trade-off parameters $\lambda_1, \ldots, \lambda_n$. But, by construction, they satisfy $T_i' = (T_{i-1}', S_i)$, where, here, $S_i := T_i$.

$(ii) \Rightarrow (iii)$. We merely define $q(X, T_1, \ldots, T_n, Y)$ through the density

$$q(x, t_1, \ldots, t_n, y) := q(x, t_1, \ldots, t_n) q(y|x).$$

From this construction and the fact that each individual bottleneck must by definition satisfy $Y - X - T_i$, it is clear that $q(X, T_1, \ldots, T_n, Y)$ is indeed an extension of the individual bottleneck probabilities $q(X, Y, T_i)$. Moreover, by construction, we have

$$Y - X - (T_n, \ldots, T_1).$$

This latter Markov chain, combined with the assumed Markov chain (4), together imply that the Markov chain (5) holds.

Appendix B.2. Operational Interpretation of Successive Refinement

This section describes the operational interpretation—for the case of discrete variables X,Y—of successive refinement, which was already proposed in [30,31], as well as, in a slightly more general fashion, in [32]. We will here rely on the content from the latter work (even though our notations will be different). We will denote, for a variable Z, by Z^l, the concatenation of l i.i.d. variables with the same law as Z.

Definition A1. *For $l \in \mathbb{N}$, an* n-stage $(l, M_1, \ldots, M_n)$-code *consists of n encoder functions*

$$\phi_i^l \ : \ \mathcal{X}^l \to \{1, \ldots, M_i\}$$

and n decoder functions

$$\psi_i^l \ : \ \{1, \ldots, M_1\} \times \cdots \times \{1, \ldots, M_i\} \to \mathcal{Y}^l.$$

For a given source X, the i-th output of the $(l, M_1, \ldots, M_n)$-code will be written

$$\hat{Y}_i^l := \psi_i^l(\phi_1^l(X^l), \ldots, \phi_i^l(X^l)).$$

Intuitively, each new encoder extracts additional information from the same source, and, crucially, each new decoder is allowed to rely on *all* the information encoded until the i-th stage. Note that the output space of the decoder is modelled on that of the relevancy variable because this is the one about which one wants to extract information.

Definition A2. *The* relevance-complexity region *is the set of tuples $(R_1, \ldots, R_n, \mu_1, \ldots, \mu_n)$ such that there exists a sequence of n-stage $(l, M_1, \ldots, M_n)$-codes for all $1 \leq i \leq n$,*

$$\forall l \in \mathbb{N}, \quad \frac{1}{l} \log M_i \ \leq \ R_i$$

and

$$\forall l \in \mathbb{N}, \quad \frac{1}{l} I(Y^l; \hat{Y}_i^l) \ \geq \ \mu_i.$$

Intuitively, for a tuple to be in the relevance-complexity region, there must be an n-stage code such that the i-th encoder adds information at a rate no larger than R_i, and the i-th decoder yields information about the target variable Y no lower than μ_i. In other words, the relevance-complexity region is made of all the tuples that are achievable by n-stage codes.

Now, let us give the operational definition of successive refinement. We will denote, for a parameter λ, by $I_Y(\lambda)$, the maximum value of $I(Y;T)$ in the primal IB problem (1).

Definition A3. *Let $0 \leq \lambda_1 < \cdots < \lambda_n$. An IB problem defined by $p(X,Y)$ is said to be* operationally successively refinable, *or* O-SR, *for rates $(\lambda_1, \ldots, \lambda_n)$, if the tuple*

$$(\lambda_1 \ , \ \lambda_2 - \lambda_1 \ , \ \ldots \ , \ \lambda_n - \lambda_{n-1}, \ I_Y(\lambda_1), \ \ldots, \ I_Y(\lambda_n))$$

is in the relevance-complexity region.

Intuitively, in the case $n = 2$, assume one is given a total rate λ_2 to "spend" on encoding a source X. One can choose to encode the source in a single processing stage, yielding at best, after decoding, asymptotic relevant information $I_Y(\lambda_2)$ (see [2]). Alternatively, one can choose to break up the total rate λ_2 into two rates $R_1 := \lambda_1 < \lambda_2$ and $R_2 := \lambda_2 - \lambda_1$, and successively encode potentially different aspects of the source at these rates. Operational SR means that even though this second alternative "spends" the total rate λ_2 along two distinct stages, it can still, after decoding, also yield asymptotic relevant information of $I_Y(\lambda_2)$. Naturally, in this case, the relevant information decodable from only the first stage must also be the optimal one, i.e., $I_Y(\lambda_1)$—otherwise, the "waste" in spending the rate λ_1 would prevent the second-stage decoder, which partially relies on the information encoded at the first stage, from ever achieving the optimal relevant information $I_Y(\lambda_2)$.

We then have the following single-letter characterisation:

Proposition A3. *The IB problem defined by $p(X,Y)$ is O-SR for rates $(\lambda_1,\dots,\lambda_n)$ if and only if there exist variables $T_1,\dots,T_n$ such that*

(i) We have the Markov chain $Y - X - T_n - \cdots - T_1$;
(ii) The variables $T_1,\dots,T_n$ are each bottlenecks with respective parameters $\lambda_1,\dots,\lambda_n$.

Proof. This single-letter characterisation is a consequence of Remark 1 in [32], which states the following: a tuple $(R_1,\dots,R_n,\mu_1,\dots,\mu_n)$ is in the relevance-complexity region if and only if there exist variables $T_1,\dots,T_n$ such that the Markov chain $Y - X - T_n - \cdots - T_1$ holds, and such that, for all $i = 1,\dots,n$,

$$\sum_{j=1}^{i} I(X;T_j|T_1,\dots,T_{j-1}) \leq \sum_{j=1}^{i} R_j, \tag{A11}$$

$$I(Y;T_i) \geq \mu_i. \tag{A12}$$

By simplifying the left-hand side in (A11) through the chain rule for mutual information, defining $\lambda_i := \sum_{j=1}^{i} R_j$, and applying the statement with $\mu_i := I_Y(\lambda_i)$, we obtain that the IB problem is O-SR for rates $(\lambda_1,\dots,\lambda_n)$ if and only if there exist variables $T_1,\dots,T_n$ such that

1. We have the Markov chain $Y - X - T_n - \cdots - T_1$; and
2. We have, for all $i = 1,\dots,n$,

$$I(X;T_i) \leq \lambda_i, \tag{A13}$$

$$I(Y;T_i) \geq I_Y(\lambda_i). \tag{A14}$$

However, if point 1 above holds, then, particularly for all $i = 1,\dots,n$, we have the Markov chain $Y - X - T_i$. As a consequence, by definition of the primal IB problem (1), the inequality in (A14) can be replaced by an equality, and thus point 2 as a whole can be replaced by the condition that T_i is a bottleneck of parameter λ_i for the IB problem defined by $p(X,Y)$. Hence, we are left with points (i) and (ii) of Theorem A3's statement. □

It is worth mentioning that our Proposition A3 is also essentially Theorem 7 in [30], which proves the same single-letter characterisation for the same operational problem—up to the difference that the result is limited to $n = 2$, and that the latter work does not consider any decoder functions ψ_i^l. Moreover, Proposition A3 is a consequence of Lemma 4 in [31].

It is clear that the conditions of Theorem A3 are exactly those of Proposition 4-(iii), so the operational Definition A3 and the single-letter Definition 5 are equivalent; in other words, the notion studied in our work does have an operational interpretation. Crucially, the operational construction of Definitions A1–A3 also goes clearly along the interpretation in terms of the successive incorporation of information.

Appendix B.3. Proof of Proposition 2

First of all, note that even though in the Definition 5 of successive refinement, the term "bottleneck" refers to a solution to the primal problem (1), the definition makes as much sense if now by "bottleneck" we mean a solution to the Lagrangian problem (3). This is, therefore, what we will be speaking about in this section. With this Lagrangian version, the Markov chain characterisation given by Proposition 1 still holds. More precisely:

Proposition A4. *Let (X,Y) be jointly Gaussian, and $1 \leq \beta_1 < \cdots < \beta_n$. The following are equivalent:*

(i) There is successive refinement for Lagrangian parameters $(\beta_1,\dots,\beta_n)$.

(*ii*) *There exist Lagrangian bottlenecks* $T_1, \ldots, T_n$*, of common source* X *and relevancy* Y*, with respective parameters* $\beta_1, \ldots, \beta_n$*, and an extension* $q(Y, X, T_1, \ldots, T_n)$ *of the* $q_i := q_i(Y, X, T_i)$*, such that, under* q*, we have the Markov chain*

$$Y - X - T_n - \cdots - T_1. \tag{A15}$$

Proof. One can directly verify that the proof given for Proposition 1 (see Appendix B.1) does not involve the explicit form of the IB problem, so the very same proof can be used for the Lagrangian formulation. □

The statement of Proposition 2 is now fully explicit.

Proof of Proposition 2. For the case of the Lagrangian IB problem with jointly Gaussian source X and relevancy Y, an analytic solution was given in [75], which proves among other things that the functions $(\beta \mapsto I_\beta(X;T))$ and $(\beta \mapsto I_\beta(Y;T))$ are continuous and increasing, where $I_\beta(X;T)$ and $I_\beta(Y;T)$ are defined by bottlenecks T of Lagrangian trade-off parameter β. Let us define

$$\beta_{IB}(X,Y) := \sup\{\beta \in \mathbb{R} \,:\, I_\beta(X;T) = 0\},$$

where we must have $\beta_{IB}(X,Y) \geq 1$ (see Section 1.3). Moreover, from the continuity of the function $(\beta \mapsto I_\beta(X;T))$, this supremum is a maximum, and from the monotonicity of the latter function, $I_\beta(X;T) = 0$ for all $\beta \leq \beta_{IB}(X,Y)$, whereas, by definition of $\beta_{IB}(X,Y)$, we have $I_\beta(X;T) > 0$ for all $\beta > \beta_{IB}(X,Y)$. Thus, $\beta_{IB}(X,Y)$ delimits trivial from non-trivial solutions, and we can, without loss of generality, choose $\beta \geq \beta_{IB}(X,Y)$.

Let us now turn to the *semigroup structure* of the Gaussian IB problem, which was both defined and proved in [33]. In short, this structure means that one can *compose* two Gaussian bottlenecks, while still obtaining a Gaussian bottleneck for the original problem. More precisely, let $\beta_2 > \beta_{IB}(X,Y)$, and define T_2 as the analytical solution to the Lagrangian IB from [75]. This provides one with a joint distribution $q_2(Y,X,T_2)$, which, importantly for us here, happens to define a Gaussian vector as well. Then, we consider a new IB problem with still the same relevancy variable Y, but now with T_2 as the source, i.e.,

$$\underset{q(T_1|T_2)\,:\,T_1-T_2-Y,}{\arg\min} \; I(T_2;T_1) \;-\; \beta_1' \, I(Y;T_1), \tag{A16}$$

where $\beta_1' \geq \beta_{IB}(T_2,Y)$. As T_2 and Y are jointly Gaussian, the problem above is again a Gaussian IB problem, so we can again analytically define a solution T_1 with the formulas from [75], yielding a distribution $q_1(Y,T_2,T_1)$. The semigroup structure proven in [33] refers to the following feature:

Proposition A5. *Assume that* T_1 *and* T_2 *are built as above, and define the extension* $q(Y,X,T_1,T_2)$ *of* $q_1(Y,X,T_1)$ *and* $q_2(Y,X,T_2)$ *through*

$$q(y,x,t_1,t_2) := q_2(y,x,t_2)q_1(t_1|t_2). \tag{A17}$$

Then, the marginal $q(Y,X,T_1)$ *defines a Lagrangian bottleneck of source* X *and relevancy* Y *for some parameter* β_1 *uniquely defined, with* $\beta_{IB}(X,Y) \leq \beta_1 < \beta_2$*.*

Thus, we can define a binary operator "$\circ$", which, for every $\beta_2 > \beta_{IB}(X,Y) \geq 1$ and $\beta_1' \geq \beta_{IB}(T_1,Y)$, provides the parameter $\beta_1 := \beta_2' \circ \beta_1$ defined by Proposition A5. Ref. [33] gives an explicit formula for this binary operator :

$$\beta_1' \circ \beta_2 = \frac{\beta_1'\beta_2}{\beta_1' + \beta_2 - 1}, \tag{A18}$$

which is well-defined for $\beta_2 > \beta_{IB}(X,Y)$ and $\beta_1' \geq \beta_{IB}(T_2,Y)$, because $\beta_{IB}(X,Y) \geq 1$ and $\beta_{IB}(T_2,Y) \geq 1 \geq 0$ imply that $\beta_1' + \beta_2 - 1 > 0$. This formula implies the following:

Proposition A6. *Let $\beta_2 > \beta_{IB}(X,Y)$. For any β_1 such that $\beta_{IB}(X,Y) \leq \beta_1 < \beta_2$, there exists a β_1' such that $\beta_1 = \beta_1' \circ \beta_2$.*

Proof. Let f denote the function $\beta_1' \mapsto \beta_1' \circ \beta_2$, which is well-defined and continuous on the interval $[\beta_{IB}(T_1,Y), +\infty[$. It is clear from formula (A18) that

$$\lim_{\beta_1' \to \infty} f(\beta_1') = \beta_2. \tag{A19}$$

On the other hand, note first that as $\beta_{IB}(T_2,Y)$ delimits trivial from non-trivial solutions, we have $I_{\beta_{IB}(T_2,Y)}(T_2;T_1) = 0$. But, by construction, under q given by Equation (A17), we have the Markov chain $Y - X - T_2 - T_1$. Thus, $I_{\beta_{IB}(T_2,Y)\circ\beta_2}(X;T_1) \leq I_{\beta_{IB}(T_2,Y)}(T_2,T_1)$, i.e., $I_{\beta_{IB}(T_2,Y)\circ\beta_2}(X;T_1) = 0$. So, by definition of $\beta_{IB}(X,Y)$, we have

$$\beta_{IB}(T_2,Y) \circ \beta_2 \leq \beta_{IB}(X,Y), \tag{A20}$$

i.e.,

$$f(\beta_{IB}(T_2,Y)) \leq \beta_{IB}(X,Y). \tag{A21}$$

Now, Equations (A19) and (A21), combined with the continuity of f, imply that

$$[\beta_{IB}(X,Y), \beta_2[\subseteq f\big([\beta_{IB}(T_2,Y), \infty[\big),$$

which yields the result. □

Now let us consider a family of parameters $\beta_{IB}(X,Y) \leq \beta_1 < \cdots < \beta_n$. By iterating Propositions A5 and A6 used together, we obtain that there exist bottlenecks $T_1, \ldots, T_n$ of common source X and relevancy Y, with respective parameters $\beta_1, \ldots, \beta_n$, and an extension $q(Y, X, T_1, \ldots, T_n)$ of these bottlenecks defined by

$$q(y, x, t_1, \ldots, t_n) := q(y, x, t_n) q(t_{n-1}|t_n) \ldots q(t_1|t_2).$$

By construction, under q, the Markov chain $Y - X - T_n - \cdots - T_1$ holds. In other words, condition (ii) from Proposition A4 is satisfied, which proves the successive refinability of jointly Gaussian vectors for the Lagrangian IB problem. □

Appendix B.4. Proof of Proposition 3

Here, for $\alpha \in [0,1]$, we denote by T_α the variable defined by

$$\begin{aligned} q(T_\alpha = Y|X) &= \alpha \\ q(T_\alpha = e|X) &= 1 - \alpha, \end{aligned} \tag{A22}$$

where e denotes a dummy symbol not pertaining to either $\mathcal{X}$ or $\mathcal{Y}$. It was proven in [67] that, for every primal parameter $\lambda \in [0, I(X;Y)]$, there exists an α such that T_α is a bottleneck of parameter λ. Note that we must have

$$\lambda = I(X;T_\alpha) = \alpha I(X;Y), \tag{A23}$$

where the first equality comes the general fact that a bottleneck must saturate the information constraint in (1) (see Section 1.3), and the second equality is a direct computation from (A22). Thus, α is a bijective and increasing function of λ, and it is sufficient, for

proving successive refinement, to prove that, for $0 \leq \alpha_1 < \cdots < \alpha_n \leq 1$, we can design a joint distribution $q(X, T_{\alpha_1}, \ldots, T_{\alpha_n})$ such that we have the Markov chain

$$X - T_{\alpha_n} - \cdots - T_{\alpha_1}.$$

Let us first focus on the case $n = 2$. We define a bottleneck $T_2 := T_{\alpha_2}$, i.e, we set $q(X, T_2) := q(X, T_{\alpha_2})$ and then a distribution $q(T_1, T_2)$ through

$$\begin{aligned} q(T_1 = Y|T_2 = Y) &:= \frac{\alpha_1}{\alpha_2} \\ q(T_1 = e|T_2 = Y) &:= \frac{\alpha_2 - \alpha_1}{\alpha_2} \\ q(T_1 = Y|T_2 = e) &:= 0 \\ q(T_1 = e|T_2 = e) &:= 1. \end{aligned}$$

We then define an extension $q(X, T_1, T_2)$ of $q(X, T_2)$ and $q(T_1, T_2)$ through

$$q(x, t_1, t_2) := q(x, t_2)q(t_1|t_2),$$

which implies by construction the Markov chain $X - T_2 - T_1$. But it also implies that

$$\begin{aligned} q(T_1 = Y|x) &= q(T_1 = Y|T_2 = Y)q(T_2 = Y|x) + q(T_1 = Y|T_2 = e)q(T_2 = e|x) \\ &= \frac{\alpha_1}{\alpha_2}\alpha_2 + 0 \times (1 - \alpha_2) \\ &= \alpha_1, \end{aligned}$$

and thus, necessarily, $q(T_1 = e|X) = 1 - \alpha_1$. So, $q(X, T_1) = q(X, T_{\alpha_1})$. Thus, we built a joint law $q(X, T_{\alpha_1}, T_{\alpha_2})$ such that $X - T_{\alpha_2} - T_{\alpha_1}$, which proves successive refinement for the case $n = 2$. The case of arbitrary n follows by direct iteration of the previous reasoning, where one starts from defining $q(X, T_n)$ through $T_n := T_{\alpha_n}$, and then iteratively defines $q(X, T_i, T_{i+1}, \ldots, T_n)$ through a well-chosen $q(T_i|T_{i+1})$ and the Markov chain condition $X - T_n - \cdots - T_{i+1} - T_i$.

Appendix B.5. Proof of Proposition 4

The result is a consequence of the following general fact, where we will eventually set $U := T_1$, $V := T_2$, and $W := X$.

Proposition A7. *Let $q(U, W)$ and $q(V, W)$ be full-support consistent distributions, defined on discrete alphabets $\mathcal{U} \times \mathcal{W}$ and $\mathcal{V} \times \mathcal{W}$, respectively. Consider the following properties:*

(i) There exists an extension $\tilde{q}(U, V, W)$ of $q(U, W)$ and $q(V, W)$ under which the Markov chain $U - V - W$ holds.

(ii) For each $u \in \mathcal{U}$, there exists a family of convex combination coefficients $\{\alpha_{v,u}\ ,\ v \in \mathcal{V}\}$ such that

$$q(W|u) = \sum_v \alpha_{v,u}\, q(W|v).$$

Then, we always have $(i) \Rightarrow (ii)$ and, if, moreover, the channel $q(W|V)$ is injective, then we also have $(ii) \Rightarrow (i)$, and the extension $\tilde{q}$ is uniquely defined.

Note the abuse of notations in the statement of Proposition A7: we write q for both $q(U, V)$ and $q(V, W)$, which are distinct distributions on partially distinct alphabets, even though they are consistent; in addition, along the proof, context, if not explicit statements, will make clear which distribution we are referring to.

Proof. Along the proof, we will be using the fact that a probability distribution is equivalent to a family of convex combination coefficients several times; indeed, both notions define a family of non-negative numbers such that their sum equals one.

$(i) \Rightarrow (ii)$. For all u, w, assumption (i) provides a $\tilde{q}(U, V, W)$ such that

$$\begin{aligned} q(w|u) &= \tilde{q}(w|u) \\ &= \sum_v \tilde{q}(w, v|u) \\ &= \sum_v \tilde{q}(v|u)\, \tilde{q}(w|v) \\ &= \sum_v \tilde{q}(v|u)\, q(w|v), \end{aligned}$$

where the first and fourth equalities use the fact that $\tilde{q}(U, V, W)$ is an extension of $q(U, W)$ and $q(V, W)$, and the third equality uses the fact that, under $\tilde{q}(U, V, W)$, the Markov chain $U - V - W$ holds. Let us define $\alpha_{v,u} := q(v|u)$. For each $u \in \mathcal{U}$, the family $\{\alpha_{v,u},\ v \in \mathcal{V}\}$ is a probability distribution, and thus a family of convex combination coefficients.

$(ii) \Rightarrow (i)$. We want to design a distribution $\tilde{q}$ that is both consistent with $q(U, W)$ and $q(V, W)$, and satisfies $U - V - W$. Thus, such a distribution is wholly defined by $\tilde{q}(V|U)$, because it must satisfy

$$\begin{aligned} \tilde{q}(u, v, w) &= \tilde{q}(u)\tilde{q}(v|u)\tilde{q}(w|v) \\ &= q(u)\tilde{q}(v|u)q(w|v), \end{aligned} \tag{A24}$$

where $q(U)$ is obtained by marginalising $q(U, W)$, whereas $q(W|V)$ is obtained from $q(V, W)$. Assumption (ii) provides a candidate: let us define $\tilde{q}(v|u) := \alpha_{v,u}$, which makes sense because, for each u, the family $(\alpha_{v,u})_v$ is made of convex combination coefficients. From assumption (ii), for all u, w,

$$q(w|u) = \sum_v \tilde{q}(v|u)\, q(w|v), \tag{A25}$$

and the corresponding $\tilde{q}(U, V, W)$ defined through Equation (A24) satisfies the Markov chain $U - V - W$.

To prove that $\tilde{q}$ is an extension of $q(U, W)$ and $q(V, W)$, let us prove first that $\tilde{q}$ is consistent with $q(U, W)$. We have

$$\begin{aligned} \tilde{q}(u, w) &= \sum_v \tilde{q}(u, v, w) \\ &= \sum_v q(u)\tilde{q}(v|u)q(w|v) \\ &= q(u) \sum_v \tilde{q}(v|u)q(w|v) \\ &= q(u)q(w|u) \\ &= q(u, w), \end{aligned}$$

where the first equality is the definition of the marginal $\tilde{q}(u, w)$; the second equality uses Equation (A24); and the fourth equality uses (A25). Thus, $\tilde{q}(U, V, W)$ is consistent with $q(U, W)$.

Now, let us prove that $\tilde{q}(V, W) = q(V, W)$. This is equivalent to the channel $\tilde{q}(V|U)$ sending the marginal $q(U)$ on the marginal $q(V)$:

Lemma A1. *We have $\tilde{q}(V, W) = q(V, W)$ if and only if*

$$\tilde{Q}_{vu} q_u = q_v, \tag{A26}$$

where q_u and q_v are the column vectors defined by $q(U)$ and $q(V)$, respectively, and $\tilde{Q}_{vu}$ is the column transition matrix defined by $\tilde{q}(V|U)$.

Proof. For all v, w,

$$\begin{aligned}\tilde{q}(v,w) &= \sum_u \tilde{q}(u,v,w)\\ &= \Big(\sum_u q(u)\tilde{q}(v|u)\Big)q(w|v),\end{aligned}$$

where the first equality is the definition of the marginal $\tilde{q}(v,w)$, and the second one uses Equation (A24). Thus, for all v, w,

$$\begin{aligned}\tilde{q}(v,w) = q(v,w) \quad &\Leftrightarrow \quad q(v,w) = \Big(\sum_u q(u)\tilde{q}(v|u)\Big)\tilde{q}(w|v)\\ &\Leftrightarrow \quad q(v)q(w|v) = \Big(\sum_u q(u)\tilde{q}(v|u)\Big)q(w|v),\end{aligned}$$

and, eventually, for all v, w,

$$\tilde{q}(v,w) = q(v,w) \quad \Leftrightarrow \quad q(w|v) = 0 \quad \text{or} \quad q(v) = \sum_u q(u)\tilde{q}(v|u). \tag{A27}$$

Let us momentarily fix $v \in \mathcal{V}$. Since $q(W|v)$ is a probability, there must be some w_0 such that $q(w_0|v) > 0$. Choosing that w_0, we find that, for the given v, the vector equality $\tilde{q}(v,W) = q(v,W)$ implies, through Equation (A27), that the scalar equality $q(v) = \sum_u q(u)\tilde{q}(v|u)$. By now applying this reasoning to each $v \in \mathcal{V}$, we obtain that $\tilde{q}(V,W) = q(V,W)$ implies that

$$\forall v \in \mathcal{V}, \quad \sum_u q(u)\tilde{q}(v|u) = q(v), \tag{A28}$$

whose matrix formulation is precisely (A26). Conversely, if (A28) holds, then Equation (A27) shows that $\tilde{q}(V,W) = q(V,W)$. □

We now prove that Equation (A26) indeed holds. Let us also write Q_{wv} and Q_{wu} for the column transition matrices defined by $q(W|V)$ and $q(W|U)$, respectively. Then, Equation (A25), which, here, is our assumption, can be rewritten as

$$Q_{wu} = Q_{wv}\tilde{Q}_{vu}. \tag{A29}$$

Thus,

$$Q_{wv}\tilde{Q}_{vu}q_u = Q_{wu}q_u = q_w = Q_{wv}q_v$$

where q_w is the column vector defined by $q(W)$, and the second and third equalities are the matrix versions of the decompositions $q(W) = \sum_u q(u)q(W|u)$ and $q(W) = \sum_v q(v)q(W|v)$, respectively. In other words,

$$Q_{wv}(\tilde{Q}_{vu}q_u - q_v) = 0. \tag{A30}$$

The injectivity of Q_{wv} implies that (A26) indeed holds, so, from Lemma A1, we have $\tilde{q}(V,W) = q(V,W)$. We have thus proven that $\tilde{q}$ extends both $q(U,W)$ and $q(V,W)$, so point (ii) holds.

Eventually, let us prove the uniqueness. Let $\tilde{q}' := \tilde{q}'(U,V,W)$ be another extension of $q(U,W)$ and $q(V,W)$ such that, under $\tilde{q}'$, the Markov chain $U - V - W$ holds. For the same reasons as above, $\tilde{q}'$ must satisfy Equation (A24) with $\tilde{q}$ replaced by $\tilde{q}'$, so $\tilde{q}'$ is wholly

specified by $\tilde{q}'(V|U)$, and is enough to prove that $\tilde{q}'(V|U) = \tilde{q}(V|U)$. Now, using the assumptions of consistency and the Markov chain for $\tilde{q}'$, we obtain

$$\begin{aligned} q(w|u) &= \tilde{q}'(w|u) \\ &= \sum_v \tilde{q}'(v,w|u) \\ &= \sum_v \tilde{q}'(v|u)\tilde{q}'(w|v) \\ &= \sum_v \tilde{q}'(v|u)q(w|v), \end{aligned} \tag{A31}$$

i.e., in matrix terms, if $\tilde{Q}'_{uv}$ is the column transition matrix representing $\tilde{q}'(V|U)$,

$$Q_{wu} = Q_{wv}\tilde{Q}'_{vu}.$$

Combining this with Equation (A29), we have $Q_{wv}(\tilde{Q}'_{vu} - \tilde{Q}_{vu}) = 0$. In other words, if c_i is the i-th column of $\tilde{Q}'_{vu} - \tilde{Q}_{vu}$, then $Q_{wv}c_i = 0$, which, by injectivity of Q_{wv}, means that $c_i = 0$. Thus, $\tilde{Q}'_{vu} - \tilde{Q}_{vu} = 0$, i.e., $\tilde{q}(U|V) = \tilde{q}'(U|V)$.

This ends the proof of Proposition A7. □

Now, first of all, note that if we set $U := T_1$, $V := T_2$ and $W := X$, then point (ii) in Proposition A7 is equivalent to the convex hull condition (7).

If there is successive refinement for parameters (λ_1, λ_2), then, from Proposition 1, there are bottlenecks T_1, T_2 of parameters λ_1, λ_2, respectively, such that $X - T_2 - T_1$; and the direction $(i) \Rightarrow (ii)$ of Proposition A7 implies that the convex hull condition (7) is satisfied.

Conversely, assume that the convex hull condition is satisfied for some bottlenecks T_1, T_2 of parameters λ_1, λ_2, respectively, such that $q_2(X|T_2)$ is injective. Then, the sense $(ii) \Rightarrow (i)$ of Proposition A7 shows that there exists a unique extension $\tilde{q}(X, T_1, T_2)$ of $q_1(X, T_1)$ and $q_2(X, T_2)$ such that we have $X - T_2 - T_1$. We then conclude with the Markov chain characterisation of successive refinement (Proposition 1).

Appendix B.6. Linear Program Used to Compute the Convex Hull Condition (7)

Consider, for points $u, v_1, \ldots, v_k \in \mathbb{R}^m$, the condition

$$u \in Hull\{v_i,\ i = 1, \ldots, k\}. \tag{A32}$$

A linear program can be used to check whether this condition holds or not; in short, it consists of the first step of the simplex method (see, e.g., [68], Section 5.6), which asserts the existence or not of an initial feasible basis, and computes this basis if it exists. More precisely, let us first note V the $m \times k$ matrix whose columns are the points v_i, and define

$$M := \begin{pmatrix} & V & \\ 1 & \ldots & 1 \end{pmatrix}, \qquad \tilde{u} := \begin{pmatrix} u \\ 1 \end{pmatrix}.$$

Then, condition (A32) can be reformulated as

$$\exists \alpha := (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k : \quad \begin{cases} M\alpha = \tilde{u}, \\ \alpha_i \geq 0 \ \text{ for } \ i = 1, \ldots, k. \end{cases} \tag{A33}$$

We now consider the linear program defined for the augmented variable

$$\tilde{\alpha} := (\alpha_1, \ldots, \alpha_k, \alpha_{k+1}, \ldots, \alpha_{k+m+1}) \in \mathbb{R}^{k+m+1}$$

as

$$\min_{\substack{\tilde{M}\tilde{\alpha}=\tilde{u} \\ \forall i=1,\dots,k+m+1,\ \alpha_i\geq 0}} \alpha_{k+1}+\cdots+\alpha_{k+m+1}, \tag{A34}$$

where $\tilde{M} := (M|I_{m+1})$ is obtained by appending the $(m+1)\times(m+1)$ identity matrix to M to the right. It can be directly verified that (A33), and thus, equivalently, (A32), holds if and only if the minimum is 0 in the linear program (A34), and that if this is the case, then the first k coordinates $\alpha_1,\dots,\alpha_k$ of any of the program's solutions provide coefficients for obtaining u as a convex combination of the v_i.

Now, consider two bottleneck distributions $q_1 := q_1(X,T_1)$ and $q_2 := q_2(X,T_2)$ such that $q(X|T_2)$ is injective. We want to check the convex hull condition (7), which holds if and only if for every $t_1 \in \mathcal{T}_1$, we have

$$q(X|t_1) \in Hull\{q(X|t_2),\ t_2 \in \mathcal{T}_2\}. \tag{A35}$$

This condition can be checked, for every fixed t_1, with the linear program described above, where if the condition holds, the algorithm also outputs a family of coefficients $(\alpha_{t_2,t_1})_{t_2}$ such that

$$q(X|t_1) = \sum_{t_2} \alpha_{t_2,t_1} q(X|t_2). \tag{A36}$$

Let us define $q(t_2|t_1) := \alpha_{t_2,t_1}$ and a joint distribution $q(X,T_1,T_2)$ through

$$q(x,t_1,t_2) := q_1(t_1)q(t_2|t_1)q_2(x|t_2). \tag{A37}$$

By construction, under q, we have the Markov chain $X - T_2 - T_1$. Moreover thanks to Equation (A36) and the injectivity of $q(X|T_2)$, Proposition A7 shows that q is indeed an extension of $q_1(X,T_1)$ and $q_2(X,T_2)$. Thus, the linear program above allows one both to check whether or not the convex hull condition holds and, when it does, to obtain Theorem 4's unique extension $q(X,T_1,T_2)$ such that $X - T_2 - T_1$.

Let us turn to considering the algorithm's complexity. For each $t_1 \in \mathcal{T}_1$, we want to know if the point $q(X|t_1)$, which is made of $m = |\mathcal{X}|$ coordinates, is in the convex hull of $k = |\mathcal{T}_2|$ points, where we can always choose $|\mathcal{T}_2| \leq |\mathcal{X}|+1$ (see Section 1.3). One can directly verify that the linear program (A34) thus consists of at most $2|\mathcal{X}|+2$ variables and $3|\mathcal{X}|+2$ equality and inequality constraints. Moreover, we want to check condition (A35) for every $t_1 \in \mathcal{T}_1$, where we can always choose $|\mathcal{T}_1| \leq |\mathcal{X}|+1$. As a consequence, the time complexity of checking the convex hull condition (7) can be bounded as $O((|\mathcal{X}|+1)\,K)$, where K is the complexity bound of a linear program with $2|\mathcal{X}|+2$ variables and $3|\mathcal{X}|+2$ constraints. By changing the multiplicative constant in the definition of the $O(\cdot)$ notation, the bound $O((|\mathcal{X}|+1)\,K)$ clearly simplifies to $O(|\mathcal{X}|\,K)$. Eventually, Ref. [69] shows that

$$K = \tilde{O}\left(|\mathcal{X}|^{\omega}\log\left(\frac{|\mathcal{X}|}{\delta}\right)\right)$$

where $\omega \approx 2.38$ corresponds to the complexity of matrix multiplication and δ is the relative accuracy. Here the notation $\tilde{O}(\cdot)$ hides polylogarithmic factors: i.e., for two functions f and g defined over positive integers, $f(n) = \tilde{O}(g(n))$ means that there exists some $r \in \mathbb{N}$ such that $f(n) = O(g(n)\log^r(g(n)))$. Overall, the convex hull condition (7) can thus be checked with an algorithm of time complexity no worse than $\tilde{O}(|\mathcal{X}|^{\omega+1}\log(\frac{|\mathcal{X}|}{\delta}))$.

Note that as the convex hull condition holds if and only if the linear program's output is 0 for all $t_1 \in \mathcal{T}_1$, in numerical computations, the threshold for rounding the program's output impacts the answer. In our numerical experiments, we chose the threshold 10^{-6}.

Appendix B.7. Proof of Proposition 5

We will first present the framework developed in [35,37] and the original content of this proof, which starts with Lemma A4 below. A full plan of this proof is presented in the main text.

We already noticed (in Section 2.2) that the primal IB problem (1) can be reformulated as an optimisation over the pairs $(q(T), q(X|T))$, i.e., Equation (6). Using the identity $I(U;V) = H(U) - H(U|V)$, and recalling that a bottleneck T must satisfy $I(X;T) = \lambda$ [37], we can further reformulate the problem (6) as

$$\underset{\substack{(q(T),\, q(X|T)) \\ \sum_t q(t)q(X|t)=p(X) \\ H(X|T)=\nu}}{\arg\min} \quad H(Y|T), \tag{A38}$$

where $\nu := H(X) - \lambda$. In particular, we can assume, without loss of generality, that $0 \leq \nu \leq H(X)$ (see Section 1.3), where $\nu = H(X)$ corresponds to $I(X;T) = 0$. Similarly as we denoted before by $I_Y(\lambda)$ the maximum in the classic IB problem (1), here, we denote by $H_Y(\nu)$ the minimum in (A38). Rather than considering the information curve, i.e., the graph of I_Y, and following [35] upon which we rely, here, we consider the graph of H_Y, which we will refer to as the conditional entropy (CE) curve. This curve is convex [35], and it is just an affine translation of the information curve. Let us now define, for $\beta \geq 1$, the function

$$\begin{aligned} F_\beta \; : \; & \Delta_{\mathcal{X}} \to \mathbb{R} \\ & p \mapsto H(\kappa p) - \beta^{-1} H(p), \end{aligned}$$

where κ is the column transition matrix defined by the conditional probability $p(Y|X)$. Note that, for $p = p(X)$, we have $\kappa p = p(Y)$. (In this section, we choose notations close to those from [37], as long as they do not clash with the ones we already established; most notably, what we denote here by β would correspond to β^{-1} in [37].)

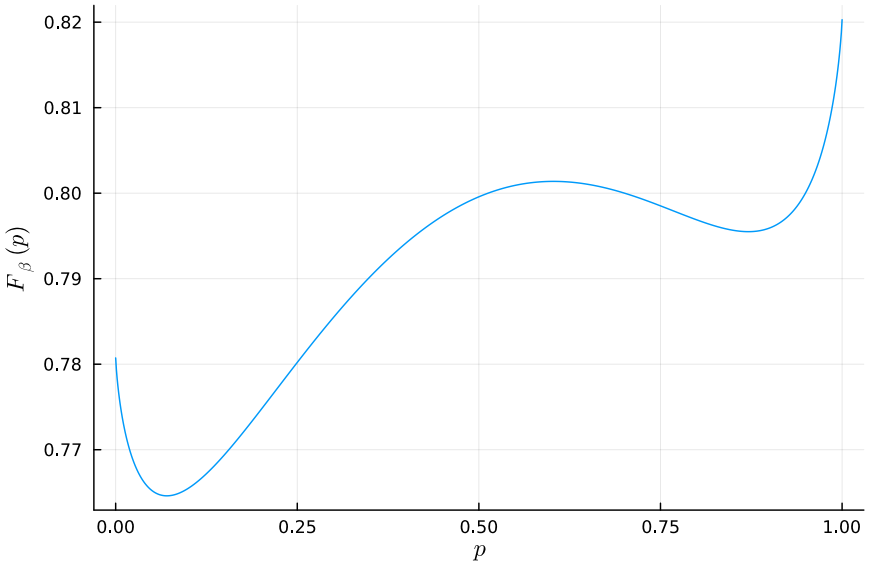

Figure A1. The function F_β for example values of β and $p(X,Y)$, where the source and relevancy are binary. Here, on the x-axis, p parameterises the binary distribution $[p, 1-p]$.

The function F_β is plotted in Figure A1 for example values of β and $p(X,Y)$, where the source and relevancy are binary. As a difference in concave functions, the function is a priori neither concave nor convex, but we can define its *lower convex envelope*, i.e., the largest convex function, which is still inferior or equal to F_β everywhere: we will denote it by

$\mathcal{K}_\cup(F_\beta)$. In Section IV in [35], through convex duality arguments, the following relationship between bottlenecks and F_β was proven:

Proposition A8. *If a pair $(q(T), q(X|T))$ solves the reformulated primal IB problem (A38), then*

$$\sum_t q(t) F_\beta(q(X|t)) = \mathcal{K}_\cup(F_\beta)(p(X)), \tag{A39}$$

for some $\beta \geq 1$ such that β^{-1} is the slope of a tangent to the CE curve at the point $(\nu, H_Y(\nu))$.

Let us also define the set of points where F_β differs from its lower convex envelope:

$$\mathcal{P}(\beta) := \{p \in \Delta_\mathcal{X} \ : \ F_\beta(p) \neq \mathcal{K}_\cup(F_\beta)(p)\}, \tag{A40}$$

which will happen to be crucial for our considerations on successive refinement. As already noted (see [37], Section II.B), this set grows when β increases:

Lemma A2. *If $\beta_1 \leq \beta_2$, then $\mathcal{P}(\beta_1) \subseteq \mathcal{P}(\beta_2)$.*

Proof. For the sake of self-containedness, we reproduce the computation from [37]. Let $p \notin \mathcal{P}(\beta_2)$, which means that $\mathcal{K}_\cup\Big(F_{\beta_2}\Big)(p) = F_{\beta_2}(p)$. For all $\beta_1 \leq \beta_2$,

$$\begin{aligned} F_{\beta_1}(p) &= H(\kappa p) - \beta_1^{-1} H(p) \\ &= F_{\beta_2}(p) - (\beta_1^{-1} - \beta_2^{-1}) H(p), \end{aligned}$$

so

$$\begin{aligned} \mathcal{K}_\cup\Big(F_{\beta_1}\Big)(p) &= \mathcal{K}_\cup\Big(F_{\beta_2} - (\beta_1^{-1} - \beta_2^{-1})H\Big)(p) \\ &\geq \mathcal{K}_\cup\Big(F_{\beta_2}\Big)(p) + \mathcal{K}_\cup\Big(- (\beta_1^{-1} - \beta_2^{-1})H\Big)(p) \\ &= \mathcal{K}_\cup\Big(F_{\beta_2}\Big)(p) - (\beta_1^{-1} - \beta_2^{-1}) H(p), \end{aligned}$$

where the last equality comes from the convexity of the function $p \mapsto -(\beta_1^{-1} - \beta_2^{-1})H(p)$. Thus,

$$\begin{aligned} \mathcal{K}_\cup\Big(F_{\beta_1}\Big)(p) &\geq \mathcal{K}_\cup\Big(F_{\beta_1}\Big)(p) - (\beta_1^{-1} - \beta_2^{-1}) H(p) \\ &= F_{\beta_2}(p) - (\beta_1^{-1} - \beta_2^{-1}) H(p) \\ &= F_{\beta_1}(p). \end{aligned}$$

But, by definition, we have $\mathcal{K}_\cup\Big(F_{\beta_1}\Big)(p) \leq F_{\beta_1}(p)$, so $\mathcal{K}_\cup\Big(F_{\beta_1}\Big)(p) = F_{\beta_1}(p)$; in other words, $p \notin \mathcal{P}(\beta_1)$. Thus, we have proved that $\mathcal{P}(\beta_2)^c \subseteq \mathcal{P}(\beta_1)^c$, which is equivalent to $\mathcal{P}(\beta_1) \subseteq \mathcal{P}(\beta_2)$. □

Let us now assume that $|\mathcal{X}| = |\mathcal{Y}| = 2$. As we already proved successive refinability for deterministic $p(Y|X)$ in Proposition 3, we can assume that $p(Y|X)$ is not deterministic. But, the case of $|\mathcal{X}| = |\mathcal{Y}| = 2$ and non-deterministic $p(Y|X)$ is exhaustively studied in [35] (Section IV.A, IV.B and IV.D). The latter work implies that, in this case:

Lemma A3. *Let $0 \leq \nu < H(X)$, let $(q(T), q(X|T))$ be a a solution to (A38) with parameter ν, and let β be given by Proposition A8. Then, the set $\mathcal{P}(\beta)$ is a non-empty open interval and, for a pair $(q(T), q(X|T))$ to satisfy (A39), the set of points*

$$\{q(X|t),\ t \in \mathcal{T}\}$$

must coincide with the extreme points of the interval $\mathcal{P}(\beta)$.

Equipped with these previously established facts, we can leverage them to prove successive refinement when $|\mathcal{X}| = |\mathcal{Y}| = 2$ and $p(Y|X)$ is not deterministic. Note that the computations from [35] that yield Lemma A3 extract crucial information from the fact that the sign of F''_β is given here by a quadratic polynomial. These computations are not straightforwardly generalisable to larger source and relevancy cardinalities—even though they might serve as inspiration for potential generalisations. Let us start with the following lemma.

Lemma A4. *Let* $0 \leq \nu < H(X)$. *Then, we can assume, without loss of generality, that* $|\mathcal{T}| = 2$. *Moreover, in this case, a solution* $(q(T), q(X|T))$ *to the reformulated IB problem* (A38) *is such that* $q(X|T)$, *seen as a probability transition matrix, is injective.*

Proof. Let $(q(T), q(X|T))$ be a solution to (A38) for parameter ν, and let β be given by Proposition A8. From Lemma A3, each $q(X|t)$ must correspond to one of the two extreme points of the interval $\mathcal{P}(\beta)$. Moreover, Proposition A2 ensures that, for any primal bottleneck (or equivalently, any solution to (A38)), we still obtain a bottleneck for the same parameter if we merge symbols t with identical $q(X|t)$. Thus, we can assume, without loss of generality, that $|\mathcal{T}| = 2$, and, in this case, the decoder $q(X|T)$ is, up to permutation of bottleneck symbols, uniquely defined by β.

Moreover, as $\mathcal{P}(\beta)$ is open and non-empty, these extreme points are distinct; in other words, the column transition matrix Q defined by $q(X|T)$ has its columns made of two distinct points on the simplex $\Delta_{\mathcal{X}}$. These points must thus be linearly independent as vectors in $\mathbb{R}^2$, so the rank of Q is 2. By the null rank theorem and as $|\mathcal{T}| = 2$, this implies that Q is injective. □

Let us now first consider SR for the case of $n = 2$ processing stages. Let $0 < \lambda_1 < \lambda_2 \leq H(X)$, and let T_1, T_2 be solutions to the primal IB problem (1) of respective parameters λ_1, λ_2. Equivalently, T_1 and T_2 are solutions to the reformulated IB problem (A38) with resp. parameters ν_1, ν_2, where $0 \leq \nu_2 < \nu_1 < H(X)$. From Lemma A4, we can assume that $q(X|T_2)$ is injective. Moreover, from Proposition A8, the bottleneck pairs $(q(T_1), q(X|T_1))$ and $(q(T_2), q(X|T_2))$ are solutions to (A39) for parameters β_1, β_2, respectively, which correspond to inverse slopes of the CE curve at $(\nu_1, H_Y(\nu_1))$ and $(\nu_2, H_Y(\nu_2))$, respectively. By convexity of the CE curve [35], we have $\beta_1 \leq \beta_2$. Thus, from Lemma A2,

$$\mathcal{P}(\beta_1) \subseteq \mathcal{P}(\beta_2).$$

This is equivalent to

$$Hull\big(\overline{\mathcal{P}(\beta_1)}\big) = \overline{\mathcal{P}(\beta_1)} \subseteq \overline{\mathcal{P}(\beta_2)} = Hull\big(\overline{\mathcal{P}(\beta_2)}\big),$$

where $\overline{E}$ denotes the closure of a set E, so, here, $\mathcal{P}(\beta_i)$ and $\overline{\mathcal{P}(\beta_i)}$ only differ by taking or not taking the segment's extreme points, and the equalities come from the convexity of this segment. From Lemma A3, this can be rewritten as

$$Hull\big\{q(X|t_1),\, t_1 \in \mathcal{T}_1\big\} \subseteq Hull\big\{q(X|t_2),\, t_2 \in \mathcal{T}_2\big\}.$$

But this is exactly the convex hull condition (7). As we chose an injective $q(X|T_2)$, we can use the convex hull characterisation (Theorem 4) to conclude that T_1 and T_2 achieve successive refinement. Thus, we have proved SR for $n = 2$ stages.

Appendix B.8. Computation of Bifurcations Values

In this work, we compute the bottlenecks' bifurcation parameters as the values where the effective cardinality changes [43]: i.e., a bifurcation is a trade-off parameter value λ for

which the number of distinct $q_\lambda(X|t)$ changes in a neighborhood of λ (see Section 1.3). With this naive method, the threshold chosen to numerically equate points $q(X|t)$ impacts the computed critical values, which could be avoided by using more sophisticated methods for computing these bifurcation values [42,43,71]. However, the bifurcation values computed by our naive method did correspond, on our minimal examples, to parameters where the smoothness of the functions $I_X(\beta) := I_\beta(X;T)$ and $I_Y(\beta) := I_\beta(Y;T)$ breaks. Thus, our method seemingly identifies discontinuities of the first-order derivative of I_X and I_Y, which are those of second-order derivatives of the Lagrangian in (3) (see Corollary 1 in [43]). In this sense, our naive method still identifies the IB bifurcations, if defined as second-order bifurcations of the IB Lagragian as in, e.g., [42,43].

Appendix C. Section 3 Details

Appendix C.1. Proof of Proposition 6

We recall that Δ_{q_1,q_2} is the space of extensions $q(X,T_1,T_2)$ of $q_1(X,T_1)$ and $q_2(X,T_2)$, and that $\Delta_{SR,2}$ is the space of all distributions $r(X,T_1,T_2)$ (not necessarily consistent with q_1 and q_2) under which the Markov chain $X - T_2 - T_1$ holds. We write the proof for discrete variables for ease of presentation, but the very same proof works for continuous variables if we replace sums by integrals. For $q(X,T_1,T_2) \in \Delta_{q_1,q_2}$ and $r(X,T_1,T_2) \in \Delta_{SR,2}$,

$$\begin{aligned}
D_{KL}(q||r) &= \sum q(x,t_1,t_2)\log\left(\frac{q(x,t_1,t_2)}{r(x,t_1,t_2)}\right)\\
&= \sum q(x,t_1,t_2)\log\left(\frac{q(x,t_2)q(t_1|x,t_2)}{r(x,t_2)r(t_1|t_2)}\right)\\
&= \sum q(x,t_1,t_2)\log\left(\frac{q(t_1|x,t_2)}{r(t_1|t_2)}\right) + D_{KL}(q(X,T_2)||r(X,T_2))\\
&\geq \sum q(x,t_1,t_2)\log\left(\frac{q(t_1|x,t_2)}{r(t_1|t_2)}\right)\\
&= \sum q(x,t_1,t_2)\log\left(\frac{q(t_1|x,t_2)}{q(t_1|t_2)}\right) + \sum q(t_2)D_{KL}(q(T_1|t_2)||r(T_1|t_2))\\
&\geq \sum q(x,t_1,t_2)\log\left(\frac{q(t_1|x,t_2)}{q(t_1|t_2)}\right)
\end{aligned} \tag{A41}$$

The last term is $D_{KL}(q||r_0)$, with

$$r_0(X,T_1,T_2) := q(X)q(T_2|X)q(T_1|T_2) \in \Delta_{SR,2},$$

because, under r_0, the Markov chain $X - T_2 - T_1$ holds. So, from the last inequality in (A41),

$$\inf_{r\in\Delta_{SR,2}} D_{KL}(q||r) = D_{KL}(q||r_0).$$

But, the last term of (A41) is also $I_q(X;T_1|T_2)$. Thus,

$$\begin{aligned}
D_{KL}(\Delta_{q_1,q_2}||\Delta_{SR}) &= \inf_{q\in\Delta_{q_1,q_2}}\inf_{r\in\Delta_{SR}} D_{KL}(q,r)\\
&= \inf_{q\in\Delta_{q_1,q_2}} D_{KL}(q||r_0)\\
&= \inf_{q\in\Delta_{q_1,q_2}} I_q(X;T_1|T_2)\\
&= UI(X:T_1\setminus T_2).
\end{aligned}$$

Appendix D. The Unicity and Injectivity Conjecture, and Technical Subtleties It Would Solve

In this section, we describe in more details some technical subtleties encountered in the main text, and present a conjecture that, if true, would make them fade away in cases where the information curve is strictly concave. Let us start by stating the conjecture, which is also interesting in itself. We recall that a bottleneck T is in canonical form when all the pointwise conditional probabilities $q(X|t)$ are distinct (see Section 1.3), and that every primal bottleneck can be reduced to canonical form (see Proposition A2).

Conjecture 1. *Let $p(X,Y)$ be such that the information curve is strictly concave. Then, the set of solutions $(q(T), q(X|T))$ to the primal IB problem (6) that are expressed in canonical form is such that*

(i) *The pair $(q(T), q(X|T))$ is, up to permuting bottleneck symbols, uniquely determined.*
(ii) *The channel $q(X|T)$, seen as a linear operator on probability distributions, is injective.*

Note that point (ii) in the conjecture was always numerically satisfied in our minimal numerical experiments, where we also always observed a strictly concave information curve. The strict concavity assumption is necessary for this conjecture to be possibly true, because it has been shown that for a non-strictly concave information curve, the channel $q(X|T)$ can be non-injective [39].

The convex hull characterisation of exact SR, i.e., Theorem 4, would, with Conjecture 1, be made more complete for the strictly concave case. Indeed, one can prove that the conjecture would imply the following one (Conjecture 2 can be obtained by combining Theorem 4 and Conjecture 1; as the latter is in any case not a statement for now, we omit the details):

Conjecture 2. *Let X and Y be discrete variables, let $\lambda_1 < \lambda_2$, and assume that the information curve is strictly concave. Then, there is successive refinement for parameters (λ_1, λ_2) if and only if, equivalently:*

(i) *There exist bottlenecks T_1, T_2 of parameters λ_1, λ_2, respectively, such that the convex hull condition (7) holds;*
(ii) *For any bottlenecks T_1, T_2 of parameters λ_1, λ_2, respectively, the convex hull condition (7) holds.*

In particular, assuming that the information curve is strictly concave and that Conjecture 1 is true, then if the convex hull condition breaks for some bottlenecks T_1 and T_2 of parameters λ_1 and λ_2, respectively, this is enough to conclude that there is not SR for parameters (λ_1, λ_2). Recalling that, in our numerical experiments, we observed strictly concave information curves, this would make the exact SR patterns in Figures 4–6 (right), Figures 10–12 (left), and Figure A3 (middle) exact characterisations of successive refinement. On the contrary, with Theorem 4 in its current state, in the latter figures, we are indeed guaranteed that SR holds in the blue areas where the convex hull condition is satisfied, but we are not formally guaranteed that SR does not hold in the red areas where the convex hull condition breaks. This is the reason for why, in the main text, we refer to these figures as mere numerical proxies for successive refinement.

Conjecture 1 being true would also, in the strictly concave case, solve a potential ambiguity in the definition of soft SR. Indeed, we do not provide, in Section 3.1, any formal guarantee that the quantity $UI(X : T_1 \setminus T_2)$ does not depend on the choice of the bottlenecks T_1 and T_2, among all those that solve the IB problems with respective trade-off parameters λ_1 and λ_2. To make sure that there is no such dependency, we should rather consider

$$\delta(\lambda_1, \lambda_2) := \inf_{q(T_1|X) \in IB(\lambda_1),\, q(T_2|X) \in IB(\lambda_2)} UI(X : T_1 \setminus T_2),$$

where, here, $IB(\lambda)$ denotes the set of distributions $q(T|X)$ that solve the IB problem (1) with trade-off parameter λ. In practice, there currently exists, to the best of our knowledge, no

algorithm to compute, for a given bottleneck problem, *all* the solutions in $IB(\lambda)$. This is the reason for why, in this paper, we stick to computing $UI(X : T_1 \setminus T_2)$ for fixed bottlenecks T_1 and T_2. Careful readers should take this number to be, a priori, only an upper bound on the true measure of soft successive refinement $\delta(\lambda_1, \lambda_2)$.

However, one can directly verify that either permuting bottleneck symbols t or merging those with identical $q(X|t)$—so as to obtain a canonical bottleneck—leaves the unique information invariant. Thus, if Conjecture 1-(i) is true, it proves that, for a strictly concave information curve, $UI(X : T_1 \setminus T_2)$ is actually uniquely defined by the trade-off parameters λ_1, λ_2, because any pair of corresponding bottlenecks T_1 and T_2 results in the same unique information.

Eventually, Conjecture 1 seems interesting in itself. Indeed, it would provide crucial information on the trajectory of the bottlenecks' pointwise decoders $q_\lambda(X|t)$ over λ, which could then help for theoretical advances on the successive refinement of the IB.

Appendix E. Sample $p(Y|X)$ Used in Sections 2.3 and 3.2

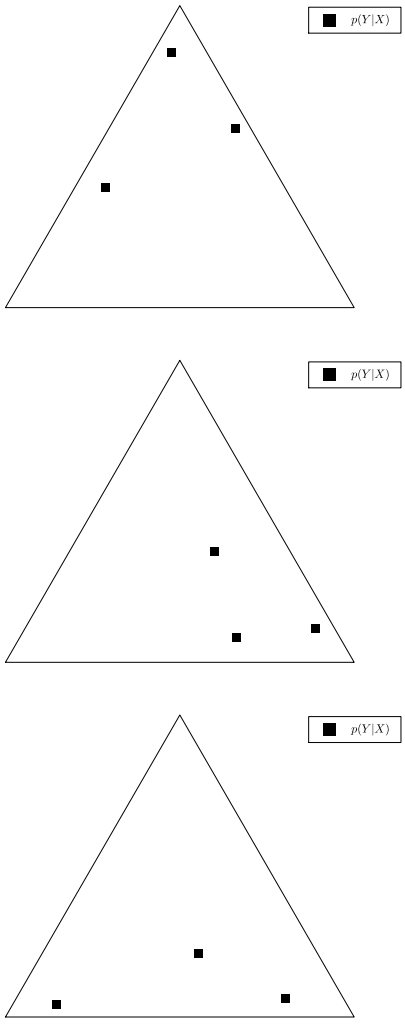

Figure A2. Plot of the sample distributions $p(Y|X)$ used in, respectively, from top to bottom: (i) Figures 4 and 7; (ii) Figures 5 and 8; (iii) Figures 6 and 9. The simplex depicted here is $\Delta_\mathcal{Y}$, where $|\mathcal{Y}| = 3$, and each black square corresponds to a symbol-wise conditional probability $p(Y|x) \in \Delta_\mathcal{Y}$. Note that the corresponding $p(X) \in \Delta_\mathcal{X}$ is shown in the left parts of Figures 4–9, which depict the simplex $\Delta_\mathcal{X}$, where, here, we also have $|\mathcal{X}| = 3$. The explicit values of the corresponding $p(X, Y)$ can be found at: https://gitlab.com/uh-adapsys/successive-refinement-ib/.

Appendix F. Additional Plots for Exact and Soft Successive Refinement

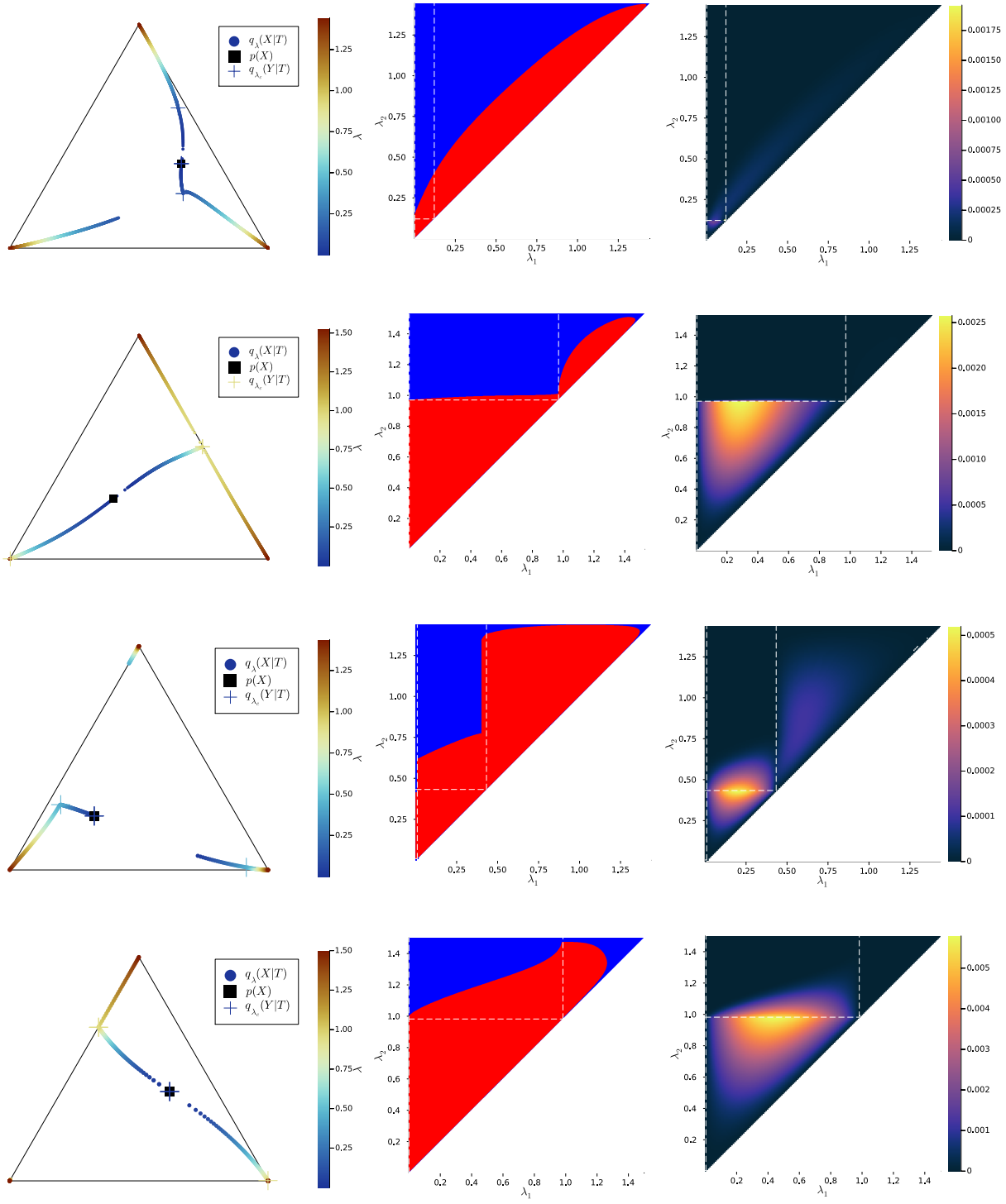

Figure A3. Additional examples for $|\mathcal{X}| = |\mathcal{Y}| = 3$: comparison of bottleneck trajectories (left) with exact SR patterns (center) and unique information landscapes (right). See Figures 4 and 7 for more details on the legends. The conditional distributions $p(Y|X)$ corresponding to each row in this figure are plotted in Figure A4. The explicit values of the corresponding $p(X, Y)$ can be found at: https://gitlab.com/uh-adapsys/successive-refinement-ib/.

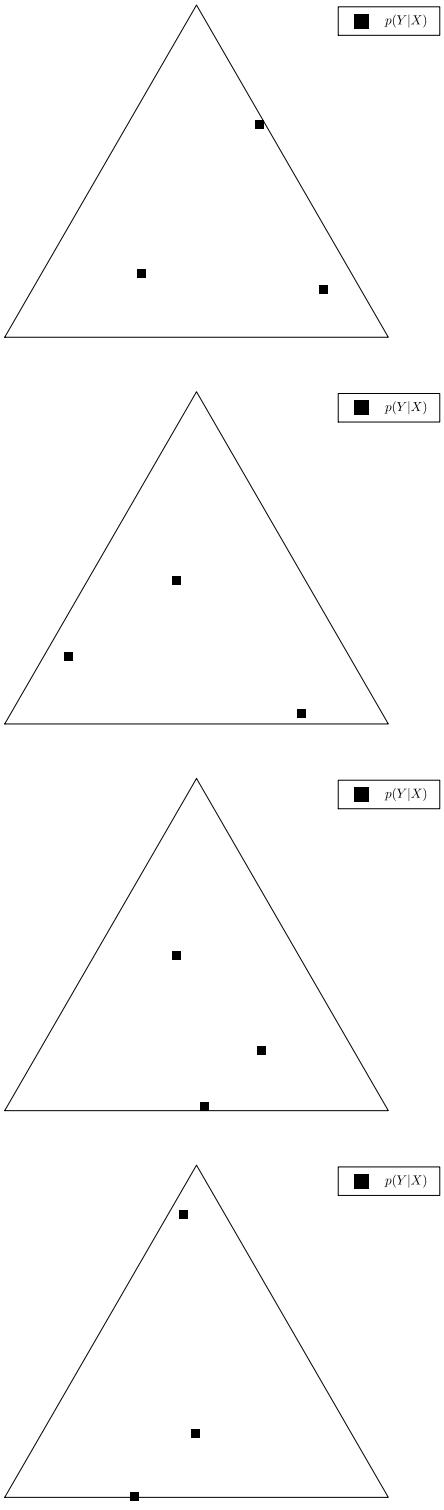

Figure A4. Sample distributions $p(Y|X)$ used in Figure A3, where the vertical order here corresponds to that of Figure A3. The simplex depicted here is $\Delta_{\mathcal{Y}}$, where $|\mathcal{Y}| = 3$, and each black square corresponds to a symbol-wise conditional probability $p(Y|x) \in \Delta_{\mathcal{Y}}$. Note that the corresponding $p(X) \in \Delta_{\mathcal{X}}$ is shown in the left parts of each row in Figure A3, which depict the simplex $\Delta_{\mathcal{X}}$, where, here, we also have $|\mathcal{X}| = 3$. The explicit values of the corresponding $p(X, Y)$ can be found at: https://gitlab.com/uh-adapsys/successive-refinement-ib/.

References

1. Tishby, N.; Pereira, F.; Bialek, W. The Information Bottleneck Method. In Proceedings of the 37th Allerton Conference on Communication, Control and Computation, 22–24 September 1999; Volume 49.
2. Gilad-Bachrach, R.; Navot, A.; Tishby, N. An Information Theoretic Tradeoff between Complexity and Accuracy. In *Learning Theory and Kernel Machines;* Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2003. [CrossRef]
3. Bialek, W.; De Ruyter Van Steveninck, R.R.; Tishby, N. Efficient representation as a design principle for neural coding and computation. In Proceedings of the 2006 IEEE International Symposium on Information Theory, Seattle, WA, USA, 9–14 July 2006; pp. 659–663. [CrossRef]
4. Creutzig, F.; Globerson, A.; Tishby, N. Past-future information bottleneck in dynamical systems. *Phys. Rev. E* **2009**, *79*, 041925. [CrossRef]
5. Amir, N.; Tiomkin, S.; Tishby, N. Past-future Information Bottleneck for linear feedback systems. In Proceedings of the 2015 54th IEEE Conference on Decision and Control (CDC), Osaka, Japan, 15–18 December 2015; pp. 5737–5742. [CrossRef]
6. Sachdeva, V.; Mora, T.; Walczak, A.M.; Palmer, S.E. Optimal prediction with resource constraints using the information bottleneck. *PLoS Comput. Biol.* **2021**, *17*, e1008743. [CrossRef] [PubMed]
7. Klampfl, S.; Legenstein, R.; Maass, W. Spiking Neurons Can Learn to Solve Information Bottleneck Problems and Extract Independent Components. *Neural Comput.* **2009**, *21*, 911–959. [CrossRef] [PubMed]
8. Buesing, L.; Maass, W. A Spiking Neuron as Information Bottleneck. *Neural Comput.* **2010**, *22*, 1961–1992. [CrossRef]
9. Chalk, M.; Marre, O.; Tkačik, G. Toward a unified theory of efficient, predictive, and sparse coding. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 186–191. [CrossRef]
10. Palmer, S.E.; Marre, O.; Berry, M.J.; Bialek, W. Predictive information in a sensory population. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 6908–6913. [CrossRef]
11. Wang, S.; Segev, I.; Borst, A.; Palmer, S. Maximally efficient prediction in the early fly visual system may support evasive flight maneuvers. *PLoS Comput. Biol.* **2021**, *17*, e1008965. [CrossRef] [PubMed]
12. Buddha, S.K.; So, K.; Carmena, J.M.; Gastpar, M.C. Function Identification in Neuron Populations via Information Bottleneck. *Entropy* **2013**, *15*, 1587–1608. [CrossRef]
13. Kleinman, M.; Wang, T.; Xiao, D.; Feghhi, E.; Lee, K.; Carr, N.; Li, Y.; Hadidi, N.; Chandrasekaran, C.; Kao, J.C. A cortical information bottleneck during decision-making. *bioRxiv* **2023**. [CrossRef]
14. Nehaniv, C.L.; Polani, D.; Dautenhahn, K.; te Beokhorst, R.; Cañamero, L. Meaningful Information, Sensor Evolution, and the Temporal Horizon of Embodied Organisms. In *Artificial life VIII;* ICAL 2003; MIT Press: Cambridge, MA, USA, 2002; pp. 345–349.
15. Klyubin, A.; Polani, D.; Nehaniv, C. Organization of the information flow in the perception-action loop of evolved agents. In Proceedings of the 2004 NASA/DoD Conference on Evolvable Hardware, Seattle, WA, USA, 24–26 June 2004; pp. 177–180. [CrossRef]
16. van Dijk, S.G.; Polani, D. Informational Drives for Sensor Evolution. Vol. ALIFE 2012: The Thirteenth International Conference on the Synthesis and Simulation of Living Systems, ALIFE 2022: The 2022 Conference on Artificial Life. 2012. Available online: https://direct.mit.edu/isal/proceedings-pdf/alife2012/24/333/1901044/978-0-262-31050-5-ch044.pdf (accessed on 12 September 2023).
17. Möller, M.; Polani, D. Emergence of common concepts, symmetries and conformity in agent groups—An information-theoretic model. *Interface Focus* **2023**, *13*, 20230006. [CrossRef]
18. Catenacci Volpi, N.; Polani, D. Space Emerges from What We Know-Spatial Categorisations Induced by Information Constraints. *Entropy* **2020**, *20*, 1179. [CrossRef]
19. Zaslavsky, N.; Kemp, C.; Regier, T.; Tishby, N. Efficient compression in color naming and its evolution. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 201800521. [CrossRef]
20. Zaslavsky, N.; Garvin, K.; Kemp, C.; Tishby, N.; Regier, T. The evolution of color naming reflects pressure for efficiency: Evidence from the recent past. *bioRxiv* **2022**. [CrossRef]
21. Tucker, M.; Levy, R.P.; Shah, J.; Zaslavsky, N. Trading off Utility, Informativeness, and Complexity in Emergent Communication. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 22214–22228.
22. Pacelli, V.; Majumdar, A. Task-Driven Estimation and Control via Information Bottlenecks. *arXiv* **2018**, arXiv:1809.07874. [CrossRef]
23. Lamb, A.; Islam, R.; Efroni, Y.; Didolkar, A.; Misra, D.; Foster, D.; Molu, L.; Chari, R.; Krishnamurthy, A.; Langford, J. Guaranteed Discovery of Control-Endogenous Latent States with Multi-Step Inverse Models. *arXiv* **2022**, arXiv:2207.08229. [CrossRef]
24. Goyal, A.; Islam, R.; Strouse, D.; Ahmed, Z.; Larochelle, H.; Botvinick, M.; Levine, S.; Bengio, Y. Transfer and Exploration via the Information Bottleneck. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
25. Koshelev, V. Hierarchical Coding of Discrete Sources. *Probl. Peredachi Inf.* **1980**, *16*, 31–49.
26. Equitz, W.; Cover, T. Successive refinement of information. *IEEE Trans. Inf. Theory* **1991**, *37*, 269–275. [CrossRef]
27. Rimoldi, B. Successive refinement of information: Characterization of the achievable rates. *IEEE Trans. Inf. Theory* **1994**, *40*, 253–259. [CrossRef]
28. Tuncel, E.; Rose, K. Computation and analysis of the N-Layer scalable rate-distortion function. *IEEE Trans. Inf. Theory* **2003**, *49*, 1218–1230. [CrossRef]

29. Kostina, V.; Tuncel, E. Successive Refinement of Abstract Sources. *IEEE Trans. Inf. Theory* **2019**, *65*, 6385–6398. [CrossRef]
30. Tian, C.; Chen, J. Successive Refinement for Hypothesis Testing and Lossless One-Helper Problem. *IEEE Trans. Inf. Theory* **2008**, *54*, 4666–4681. [CrossRef]
31. Tuncel, E. Capacity/Storage Tradeoff in High-Dimensional Identification Systems. In Proceedings of the 2006 IEEE International Symposium on Information Theory, Seattle, WA, USA, 9–14 July 2006; pp. 1929–1933. [CrossRef]
32. Mahvari, M.M.; Kobayashi, M.; Zaidi, A. On the Relevance-Complexity Region of Scalable Information Bottleneck. *arXiv* **2020**, arXiv:2011.01352. [CrossRef]
33. Kline, A.G.; Palmer, S.E. Gaussian information bottleneck and the non-perturbative renormalization group. *New J. Phys.* **2022**, *24*, 033007. [CrossRef]
34. Kolchinsky, A.; Tracey, B.D.; Van Kuyk, S. Caveats for information bottleneck in deterministic scenarios. *arXiv* **2018**, arXiv:1808.07593. [CrossRef]
35. Witsenhausen, H.; Wyner, A. A conditional entropy bound for a pair of discrete random variables. *IEEE Trans. Inf. Theory* **1975**, *21*, 493–501. [CrossRef]
36. Hsu, H.; Asoodeh, S.; Salamatian, S.; Calmon, F.P. Generalizing Bottleneck Problems. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 531–535. [CrossRef]
37. Asoodeh, S.; Calmon, F. Bottleneck Problems: An Information and Estimation-Theoretic View. *Entropy* **2020**, *22*, 1325. [CrossRef]
38. Dikshtein, M.; Shamai, S. A Class of Nonbinary Symmetric Information Bottleneck Problems. *arXiv* **2021**, arXiv:cs.IT/2110.00985.
39. Benger, E.; Asoodeh, S.; Chen, J. The Cardinality Bound on the Information Bottleneck Representations is Tight. *arXiv* **2023**, arXiv:cs.IT/2305.07000.
40. Bertschinger, N.; Rauh, J.; Olbrich, E.; Ay, N. Quantifying Unique Information. *Entropy* **2013**, *16*, 2161–2183. [CrossRef]
41. Parker, A.E.; Gedeon, T.; Dimitrov, A. The Lack of Convexity of the Relevance-Compression Function. *arXiv* **2022**, arXiv:2204.10957. [CrossRef]
42. Wu, T.; Fischer, I. Phase Transitions for the Information Bottleneck in Representation Learning. *arXiv* **2020**, arXiv:2001.01878.
43. Zaslavsky, N.; Tishby, N. Deterministic Annealing and the Evolution of Information Bottleneck Representations. 2019. Available online: https://www.nogsky.com/publication/2019-evo-ib/2019-evo-IB.pdf (accessed on 12 September 2023).
44. Ngampruetikorn, V.; Schwab, D.J. Perturbation Theory for the Information Bottleneck. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 21008–21018.
45. Bertschinger, N.; Rauh, J. The Blackwell relation defines no lattice. In Proceedings of the 2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, 29 June–4 July 2014; pp. 2479–2483. [CrossRef]
46. Yang, Q.; Piantanida, P.; Gündüz, D. The Multi-layer Information Bottleneck Problem. *arXiv* **2017**, arXiv:1711.05102. [CrossRef]
47. Cover, T.; Thomas, J. *Elements of Information Theory*; Wiley-Interscience: Hoboken, NJ, USA, 2006.
48. Zaidi, A.; Estella-Aguerri, I.; Shamai (Shitz), S. On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views. *Entropy* **2020**, *22*, 151. [CrossRef] [PubMed]
49. Tishby, N.; Zaslavsky, N. Deep Learning and the Information Bottleneck Principle. In Proceedings of the 2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, 26 April–1 May 2015. [CrossRef]
50. Shwartz-Ziv, R.; Tishby, N. Opening the Black Box of Deep Neural Networks via Information. 2017. Available online: http://xxx.lanl.gov/abs/1703.00810 (accessed on 12 September 2023).
51. Shwartz-Ziv, R.; Painsky, A.; Tishby, N. Representation Compression and Generalization in Deep Neural Networks, 2019. Available online: https://openreview.net/pdf?id=SkeL6sCqK7 (accessed on 12 September 2023).
52. Saxe, A.M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B.D.; Cox, D.D. On the information bottleneck theory of deep learning. *J. Stat. Mech. Theory Exp.* **2019**, *2019*, 124020. [CrossRef]
53. Achille, A.; Soatto, S. Emergence of Invariance and Disentanglement in Deep Representations. In Proceedings of the 2018 Information Theory and Applications Workshop (ITA), San Diego, CA, USA, 11–16 February 2018; pp. 1–9. [CrossRef]
54. Elad, A.; Haviv, D.; Blau, Y.; Michaeli, T. Direct Validation of the Information Bottleneck Principle for Deep Nets. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27–28 October 2019; pp. 758–762. [CrossRef]
55. Lorenzen, S.S.; Igel, C.; Nielsen, M. Information Bottleneck: Exact Analysis of (Quantized) Neural Networks. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
56. Kawaguchi, K.; Deng, Z.; Ji, X.; Huang, J. How Does Information Bottleneck Help Deep Learning? 2023. Available online: https://proceedings.mlr.press/v202/kawaguchi23a/kawaguchi23a.pdf (accessed on 12 September 2023).
57. Yousfi, Y.; Akyol, E. Successive Information Bottleneck and Applications in Deep Learning. In Proceedings of the 2020 54th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 1–4 November 2020; pp. 1210–1213. [CrossRef]
58. No, A. Universality of Logarithmic Loss in Successive Refinement. *Entropy* **2019**, *21*, 158. [CrossRef]
59. Nasser, R. On the input-degradedness and input-equivalence between channels. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 2453–2457. [CrossRef]
60. Lastras, L.; Berger, T. All sources are nearly successively refinable. *IEEE Trans. Inf. Theory* **2001**, *47*, 918–926. [CrossRef]
61. Williams, P.L.; Beer, R.D. Nonnegative Decomposition of Multivariate Information. 2010. Available online: https://arxiv.org/pdf/1004.2515 (accessed on 12 September 2023).

62. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J. Shared Information—New Insights and Problems in Decomposing Information in Complex Systems. In Proceedings of the European Conference on Complex Systems, 2012; Springer International Publishing: Berlin/Heidelberg, Germany, 2013; pp. 251–269. [CrossRef]
63. Griffith, V.; Koch, C. Quantifying Synergistic Mutual Information. In *Guided Self-Organization: Inception*; Prokopenko, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 159–190. [CrossRef]
64. Harder, M.; Salge, C.; Polani, D. Bivariate measure of redundant information. *Phys. Rev. E* **2013**, *87*, 012130. [CrossRef] [PubMed]
65. Blackwell, D. Equivalent Comparisons of Experiments. *Ann. Math. Stat.* **1953**, *24*, 265–272. [CrossRef]
66. Lemaréchal, C. Lagrangian Relaxation. In *Computational Combinatorial Optimization: Optimal or Provably Near-Optimal Solutions*; Jünger, M.; Naddef, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; pp. 112–156. [CrossRef]
67. Kolchinsky, A.; Tracey, B.; Wolpert, D. Nonlinear Information Bottleneck. *Entropy* **2017**, *21*, 1181. [CrossRef]
68. Matousek, J.; Gärtner, B. *Understanding and Using Linear Programming*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 2007.
69. van den Brand, J. A Deterministic Linear Program Solver in Current Matrix Multiplication Time. In Proceedings of the Thirty-First Annual ACM-SIAM Symposium on Discrete Algorithms; Society for Industrial and Applied Mathematics (SODA'20), Salt Lake City, UT, USA, 5–8 January 2020; pp. 259–278.
70. Rose, K. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE* **1998**, *86*, 2210–2239. [CrossRef]
71. Gedeon, T.; Parker, A.E.; Dimitrov, A.G. The Mathematical Structure of Information Bottleneck Methods. *Entropy* **2012**, *14*, 456–479. [CrossRef]
72. Shamir, O.; Sabato, S.; Tishby, N. Learning and generalization with the information bottleneck. *Theor. Comput. Sci.* **2010**, *411*, 2696–2711. [CrossRef]
73. Rauh, J.; Banerjee, P.K.; Olbrich, E.; Jost, J. Unique Information and Secret Key Decompositions. In Proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT), Paris, France, 7–12 July 2019. [CrossRef]
74. Banerjee, P.; Rauh, J.; Montufar, G. Computing the Unique Information. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 141–145. [CrossRef]
75. Chechik, G.; Globerson, A.; Tishby, N.; Weiss, Y. Information bottleneck for Gaussian variables. *J. Mach. Learn. Res.* **2005**, *6*, 165–188.
76. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; The MIT Press: Cambridge, MA, USA, 2018.
77. Goldfeld, Z.; Polyanskiy, Y. The Information Bottleneck Problem and its Applications in Machine Learning. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 19–38. [CrossRef]
78. Mahvari, M.M.; Kobayashi, M.; Zaidi, A. Scalable Vector Gaussian Information Bottleneck. In Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT), Melbourne, Australia, 12–20 July 2021; pp. 37–42. [CrossRef]