Information Bottleneck on General Alphabets

Georg Pichler, Günther Koliander

***Abstract*—We prove rigorously a source coding theorem that can probably be considered folklore, a generalization to arbitrary alphabets of a problem motivated by the Information Bottleneck method. For general random variables (Y, X), we show essentially that for some $n \in \mathbb{N}$, a function f with rate limit $\log|f| \leq nR$ and $\mathrm{I}(\mathsf{Y}^n; f(\mathsf{X}^n)) \geq nS$ exists if and only if there is a random variable U such that the Markov chain $\mathsf{Y} \multimap \mathsf{X} \multimap \mathsf{U}$ holds, $\mathrm{I}(\mathsf{U}; \mathsf{X}) \leq R$ and $\mathrm{I}(\mathsf{U}; \mathsf{Y}) \geq S$. The proof relies on the well established discrete case and showcases a technique for lifting discrete coding theorems to arbitrary alphabets.**

I. Introduction

Since its inception [1], the *Information Bottleneck* (IB) method became a widely applied tool, especially in the context of machine learning problems. It has been successfully applied to various problems in machine learning [2], computer vision [3], and communications [5], [6], [7]. Furthermore, it is a valuable tool for channel output compression in a communication system [8], [9].

In the underlying information-theoretic problem, we define a pair $(S, R) \in \mathbb{R}^2$ to be *achievable* for the two arbitrary random sources (Y, X), if there exists a function f with rate limited range $\frac{1}{n}\log|f| \leq R$ and $\mathrm{I}(\mathbf{Y}; f(\mathbf{X})) \geq nS$, where $(\mathbf{Y}, \mathbf{X})$ are n independent and identically distributed (i.i.d.) copies of (Y, X).

While this Shannon-theoretic problem and variants thereof were also considered (e. g., [10], [11]), a large part of the literature is aimed at studying the IB function

$$S_{\mathrm{IB}}(R) = \sup_{\substack{\mathsf{U}\,:\,\mathrm{I}(\mathsf{U};\mathsf{X}) \leq R \\ \mathsf{Y} \multimap \mathsf{X} \multimap \mathsf{U}}} \mathrm{I}(\mathsf{U}; \mathsf{Y}) \tag{1}$$

in different contexts. In particular, several works (e. g., [1], [2], [12], [13], [14]) intend to compute a probability distribution that achieves the supremum in (1). The resulting distribution is then used as a building block in numerical algorithms, e. g., for document clustering [2] or dimensionality reduction [12].

In the discrete case, $S_{\mathrm{IB}}(R)$ is equal to the maximum of all S such that (S, R) is in the *achievable region* (closure of the set of all achievable pairs). This statement has been re-proven many times in different contexts [15], [11], [16], [17]. In this note, we prove a theorem, which can probably be considered folklore, extending this result from discrete to arbitrary random variables. Formally speaking, using the definitions in [18], we prove that a pair (S, R) is in the achievable region of an arbitrary source (Y, X) if and only if, for every $\varepsilon > 0$, there exists a random variable U with $\mathsf{Y} \multimap \mathsf{X} \multimap \mathsf{U}$, $\mathrm{I}(\mathsf{X}; \mathsf{U}) \leq R + \varepsilon$, and $\mathrm{I}(\mathsf{Y}; \mathsf{U}) \geq S - \varepsilon$. This provides a single-letter solution to the information-theoretic problem behind the information bottleneck method for arbitrary random sources and in particular it shows, that the information bottleneck for Gaussian random variables [12] is indeed the solution to a Shannon-theoretic problem.

G. Pichler is with the Institute of Telecommunications, Technische Universität Wien, Vienna, Austria

G. Koliander is with the Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria

Funding by WWTF Grants MA16-053, ICT15-119, and NXT17-013.

The proof relies on the discrete case. Thus, the techniques employed could be useful for lifting other discrete coding theorems to the case of arbitrary alphabets.

II. Main Result

Let Y and X be random variables with arbitrary alphabets $\mathcal{S}_{\mathsf{Y}}$ and $\mathcal{S}_{\mathsf{X}}$, respectively. The bold-faced random vectors $\mathbf{Y}$ and $\mathbf{X}$ are n i.i.d. copies of Y and X, respectively. We then have the following definitions.

Definition 1. *A pair $(S, R) \in \mathbb{R}^2$ is* achievable *if for some $n \in \mathbb{N}$ there exists a measurable function $f\colon \mathcal{S}_{\mathsf{X}}^n \to \mathcal{M}$ for some finite set $\mathcal{M}$ with bounded cardinality $\frac{1}{n}\log|\mathcal{M}| \leq R$ and*

$$\frac{1}{n}\mathrm{I}\big(\mathbf{Y}; f(\mathbf{X})\big) \geq S. \tag{2}$$

The set of all achievable pairs is denoted $\mathcal{R} \subseteq \mathbb{R}^2$.

Definition 2. *A pair $(S, R) \in \mathbb{R}^2$ is* IB-achievable *if there exists an additional random variable U with arbitrary alphabet $\mathcal{S}_{\mathsf{U}}$, satisfying $\mathsf{Y} \multimap \mathsf{X} \multimap \mathsf{U}$ and*

$$R \geq \mathrm{I}(\mathsf{X}; \mathsf{U}), \tag{3}$$
$$S \leq \mathrm{I}(\mathsf{Y}; \mathsf{U}). \tag{4}$$

The set of all IB-achievable pairs is denoted $\mathcal{R}_{\mathrm{IB}} \subseteq \mathbb{R}^2$.

In what follows, we will prove the following theorem.

Theorem 3. *The equality $\overline{\mathcal{R}_{\mathrm{IB}}} = \overline{\mathcal{R}}$ holds.*

III. Preliminaries

When introducing a function, we implicitly assume it to be measurable w.r.t. the appropriate σ-algebras. The σ-algebra associated with a finite set is its power set and the σ-algebra associated with $\mathbb{R}$ is the Borel σ-algebra. The symbol $\varnothing$ is used for the empty set and for a constant random variable. When there is no possibility for confusion, we will not distinguish between a single-element set and its element, e. g., we write x instead of $\{x\}$ and $\mathbb{1}_x$ for the indicator function of $\{x\}$. We use $A \triangle B := (A \setminus B) \cup (B \setminus A)$ to denote the symmetric set difference.

Let (Ω, Σ, μ) be a probability space. A random variable $\mathsf{X}\colon \Omega \to \mathcal{S}_{\mathsf{X}}$ takes values in the measurable space $(\mathcal{S}_{\mathsf{X}}, \mathcal{A}_{\mathsf{X}})$. The push-forward probability measure $\mu_{\mathsf{X}}\colon \mathcal{A}_{\mathsf{X}} \to [0, 1]$ is defined by $\mu_{\mathsf{X}}(A) = \mu\big(\mathsf{X}^{-1}(A)\big)$ for all $A \in \mathcal{A}_{\mathsf{X}}$. We will state most results in terms of push-forward measures and usually ignore the background probability space. When multiple random variables are defined, we implicitly assume the push-forward measures to be consistent in the sense that, e. g., $\mu_{\mathsf{X}}(A) = \mu_{\mathsf{XY}}(A \times \mathcal{S}_{\mathsf{Y}})$ for all $A \in \mathcal{A}_{\mathsf{X}}$.

For $n \in \mathbb{N}$ let Ω^n denote the n-fold Cartesian product of (Ω, Σ, μ). A bold-faced random vector, e. g., $\mathbf{X}$, defined on Ω^n, is an n-fold copy of X, i. e., $\mathbf{X} = \mathsf{X}^n$. Accordingly, the corresponding push-forward measure, e. g., $\mu_{\mathbf{X}}$ is the n-fold product measure.

For a random variable X let a_{X}, b_{X}, and c_{X} denote arbitrary functions on $\mathcal{S}_{\mathsf{X}}$, each with finite range. We will use the symbol $\mathcal{M}_{\mathsf{X}}$ to denote the range of a_{X}, i. e., $a_{\mathsf{X}}\colon \mathcal{S}_{\mathsf{X}} \to \mathcal{M}_{\mathsf{X}}$.

Definition 4 ([19, Def. 8.11]). *The* conditional expectation *of a random variable* X *with* $\mathcal{S}_\mathsf{X} = \mathbb{R}$*, given a random variable* Y*, is a random variable* $\mathbb{E}[\mathsf{X}|\mathsf{Y}]$ *such that*

1) $\mathbb{E}[\mathsf{X}|\mathsf{Y}]$ *is* $\sigma(\mathsf{Y})$*-measurable, and*
2) *for all* $A \in \sigma(\mathsf{Y})$*, we have* $\mathbb{E}\big[\mathbb{1}_A \mathbb{E}[\mathsf{X}|\mathsf{Y}]\big] = \mathbb{E}[\mathbb{1}_A \mathsf{X}]$.

The conditional probability *of an event* $B \in \Sigma$ *given* Y *is defined as* $\mathrm{P}\{B|\mathsf{Y}\} := \mathbb{E}[\mathbb{1}_B|\mathsf{Y}]$.

The conditional expectation and therefore also the conditional probability exists and is unique up to equality almost surely by [19, Thm. 8.12]. Furthermore, if $(\mathcal{S}_\mathsf{X}, \mathcal{A}_\mathsf{X})$ is a standard space [18, Sec. 1.5], there even exists a *regular conditional distribution* of X given Y [19, Thm. 8.37].

Definition 5. *For two random variables* X *and* Y *a* regular conditional distribution *of* X *given* Y *is a function* $\kappa_{\mathsf{X}|\mathsf{Y}} \colon \Omega \times \mathcal{A}_\mathsf{X} \to [0,1]$ *such that*

1) *for every* $\omega \in \Omega$*, the set function* $\kappa_{\mathsf{X}|\mathsf{Y}}(\omega) := \kappa_{\mathsf{X}|\mathsf{Y}}(\omega; \,\cdot\,)$ *is a probability measure on* $(\mathcal{S}_\mathsf{X}, \mathcal{A}_\mathsf{X})$.
2) *for every set* $A \in \mathcal{A}_\mathsf{X}$*, the function* $\kappa_{\mathsf{X}|\mathsf{Y}}(\,\cdot\,; A)$ *is* $\sigma(\mathsf{Y})$*-measurable.*
3) *for* μ*-a. e.* $\omega \in \Omega$ *and all* $A \in \mathcal{A}_\mathsf{X}$*, we have* $\kappa_{\mathsf{X}|\mathsf{Y}}(\omega; A) = \mathrm{P}\big\{\mathsf{X}^{-1}(A)\big|\mathsf{Y}\big\}(\omega)$ *(cf. Def. 4).*

Note, in particular, that finite spaces are standard spaces.

Remark 1. If the random variable Y is discrete, then $\kappa_{\mathsf{X}|\mathsf{Y}}$ reduces to conditioning given events $\mathsf{Y} = y$ for $y \in \mathcal{S}_\mathsf{Y}$, i. e., $\kappa_{\mathsf{X}|\mathsf{Y}}(\omega; A) = \frac{\mu_{\mathsf{XY}}(A \times \mathsf{Y}(\omega))}{\mu_\mathsf{Y}(\mathsf{Y}(\omega))}$ (cf. [19, Lem. 8.10]).

We use the following definitions and results from [18], [19].

Definition 6. *For random variables* X *and* Y *with* $|\mathcal{S}_\mathsf{X}| < \infty$ *the* conditional entropy *is defined as [18, Sec. 5.5]*

$$\mathrm{H}(\mathsf{X}|\mathsf{Y}) := \int \mathrm{H}\big(\kappa_{\mathsf{X}|\mathsf{Y}}\big)\, d\mu, \tag{5}$$

where $\mathrm{H}(\,\cdot\,)$ *denotes discrete entropy on* $\mathcal{S}_\mathsf{X}$*. For arbitrary random variables* X*,* Y*, and* Z *the* conditional mutual information *is defined as [18, Lem. 5.5.7]*

$$\begin{aligned}\mathrm{I}(\mathsf{X};\mathsf{Y}|\mathsf{Z}) &:= \sup_{a_\mathsf{X}, a_\mathsf{Y}} \int \mathrm{D}\big(\kappa_{a_\mathsf{X}(\mathsf{X})a_\mathsf{Y}(\mathsf{Y})|\mathsf{Z}} \big\| \kappa_{a_\mathsf{X}(\mathsf{X})|\mathsf{Z}} \times \kappa_{a_\mathsf{Y}(\mathsf{Y})|\mathsf{Z}}\big)\, d\mu & (6)\\ &= \sup_{a_\mathsf{X}, a_\mathsf{Y}} \big[\mathrm{H}(a_\mathsf{X}(\mathsf{X})|\mathsf{Z}) + \mathrm{H}(a_\mathsf{Y}(\mathsf{Y})|\mathsf{Z}) - \mathrm{H}(a_\mathsf{X}(\mathsf{X})a_\mathsf{Y}(\mathsf{Y})|\mathsf{Z})\big], & (7)\end{aligned}$$

where $\mathrm{D}(\,\cdot\,\|\,\cdot\,)$ *denotes Kullback-Leibler divergence [18, Sec. 2.3] and the supremum is taken over all* a_X *and* a_Y *with finite range. The* mutual information *is given by [18, Lem. 5.5.1]* $\mathrm{I}(\mathsf{X};\mathsf{Y}) := \mathrm{I}(\mathsf{X};\mathsf{Y}|\varnothing)$.

Definition 7 ([19, Def. 12.20]). *For arbitrary random variables* X*,* Y*, and* Z*, the Markov chain* $\mathsf{X} \multimap \mathsf{Y} \multimap \mathsf{Z}$ *holds if, for any* $A \in \mathcal{A}_\mathsf{X}$*,* $B \in \mathcal{A}_\mathsf{Z}$*, the following holds* μ*-a. e.:*

$$\mathrm{P}\big\{\mathsf{X}^{-1}(A) \cap \mathsf{Z}^{-1}(B)\big|\mathsf{Y}\big\} = \mathrm{P}\big\{\mathsf{X}^{-1}(A)\big|\mathsf{Y}\big\}\mathrm{P}\big\{\mathsf{Z}^{-1}(B)\big|\mathsf{Y}\big\}. \tag{8}$$

In the following, we collect some properties of these definitions.

Lemma 8. *For random variables* X*,* Y*, and* Z *the following properties hold:*

(i) $\mathrm{I}(\mathsf{X};\mathsf{Y}|\mathsf{Z}) \geq 0$ *with equality if and only if* $\mathsf{X} \multimap \mathsf{Z} \multimap \mathsf{Y}$.
(ii) *For discrete* X*, i. e.,* $|\mathcal{S}_\mathsf{X}| < \infty$*, we have* $\mathrm{I}(\mathsf{X};\mathsf{Y}) = \mathrm{H}(\mathsf{X}) - \mathrm{H}(\mathsf{X}|\mathsf{Y})$.
(iii) $\mathrm{I}(\mathsf{X};\mathsf{YZ}) = \mathrm{I}(\mathsf{X};\mathsf{Z}) + \mathrm{I}(\mathsf{X};\mathsf{Y}|\mathsf{Z})$.
(iv) *If* $\mathsf{X} \multimap \mathsf{Y} \multimap \mathsf{Z}$*, then* $\mathrm{I}(\mathsf{X};\mathsf{Y}) \geq \mathrm{I}(\mathsf{X};\mathsf{Z})$.

Proof. (i): The claim $\mathrm{I}(\mathsf{X};\mathsf{Y}|\mathsf{Z}) \geq 0$ follows directly from (6) and the non-negativity of divergence.

Assume that $\mathsf{X} \multimap \mathsf{Z} \multimap \mathsf{Y}$, i. e., $\mathrm{P}\big\{\mathsf{X}^{-1}(A) \cap \mathsf{Y}^{-1}(B)\big|\mathsf{Z}\big\} = \mathrm{P}\big\{\mathsf{X}^{-1}(A)\big|\mathsf{Z}\big\}\mathrm{P}\big\{\mathsf{Y}^{-1}(B)\big|\mathsf{Z}\big\}$ almost everywhere. Let $a_\mathsf{X} \colon \mathcal{S}_\mathsf{X} \to \mathcal{M}_\mathsf{X}$ and $a_\mathsf{Y} \colon \mathcal{S}_\mathsf{Y} \to \mathcal{M}_\mathsf{Y}$ be functions with finite range. Pick two arbitrary sets $A \subseteq \mathcal{M}_\mathsf{X}$, $B \subseteq \mathcal{M}_\mathsf{Y}$ and we obtain μ-a. e.

$$\begin{aligned}&\kappa_{a_\mathsf{X}(\mathsf{X})a_\mathsf{Y}(\mathsf{Y})|\mathsf{Z}}(\,\cdot\,; A \times B)\\ &\quad= \mathrm{P}\big\{\mathsf{X}^{-1}(a_\mathsf{X}^{-1}(A)) \cap \mathsf{Y}^{-1}(a_\mathsf{Y}^{-1}(B))\big|\mathsf{Z}\big\} & (9)\\ &\quad= \mathrm{P}\big\{\mathsf{X}^{-1}(a_\mathsf{X}^{-1}(A))\big|\mathsf{Z}\big\}\mathrm{P}\big\{\mathsf{Y}^{-1}(a_\mathsf{Y}^{-1}(B))\big|\mathsf{Z}\big\} & (10)\\ &\quad= \kappa_{a_\mathsf{X}(\mathsf{X})|\mathsf{Z}}(\,\cdot\,; A)\kappa_{a_\mathsf{Y}(\mathsf{Y})|\mathsf{Z}}(\,\cdot\,; B), & (11)\end{aligned}$$

where (9) and (11) follow from part 3 of Def. 5. This proves that μ-a. e. the equality of measures $\kappa_{a_\mathsf{X}(\mathsf{X})a_\mathsf{Y}(\mathsf{Y})|\mathsf{Z}} = \kappa_{a_\mathsf{X}(\mathsf{X})|\mathsf{Z}} \times \kappa_{a_\mathsf{Y}(\mathsf{Y})|\mathsf{Z}}$ holds. By the properties of Kullback-Leibler divergence [18, Thm. 2.3.1] we have $\mathrm{I}(\mathsf{X};\mathsf{Y}|\mathsf{Z}) = 0$ due to (6).

On the other hand, assume $\mathrm{I}(\mathsf{X};\mathsf{Y}|\mathsf{Z}) = 0$ and choose arbitrary sets $A \in \mathcal{A}_\mathsf{X}$ and $B \in \mathcal{A}_\mathsf{Y}$. We define $a_\mathsf{X} := \mathbb{1}_A$, $a_\mathsf{Y} := \mathbb{1}_B$, $\hat{\mathsf{X}} := a_\mathsf{X}(\mathsf{X})$, and $\hat{\mathsf{Y}} := a_\mathsf{Y}(\mathsf{Y})$. By (6) we have $\mathrm{D}\big(\kappa_{\hat{\mathsf{X}}\hat{\mathsf{Y}}|\mathsf{Z}}(\omega)\big\|\kappa_{\hat{\mathsf{X}}|\mathsf{Z}}(\omega) \times \kappa_{\hat{\mathsf{Y}}|\mathsf{Z}}(\omega)\big) = 0$ for μ-a. e. $\omega \in \Omega$, which is equivalent to the equality μ-a. e. of the measures $\kappa_{\hat{\mathsf{X}}\hat{\mathsf{Y}}|\mathsf{Z}} = \kappa_{\hat{\mathsf{X}}|\mathsf{Z}} \times \kappa_{\hat{\mathsf{Y}}|\mathsf{Z}}$. We obtain μ-a. e.,

$$\begin{aligned}\mathrm{P}\big\{\mathsf{X}^{-1}(A) \cap \mathsf{Y}^{-1}(B)\big|\mathsf{Z}\big\} &= \kappa_{\hat{\mathsf{X}}\hat{\mathsf{Y}}|\mathsf{Z}}(\,\cdot\,; 1 \times 1) & (12)\\ &= \kappa_{\hat{\mathsf{X}}|\mathsf{Z}}(\,\cdot\,; 1)\kappa_{\hat{\mathsf{Y}}|\mathsf{Z}}(\,\cdot\,; 1) & (13)\\ &= \mathrm{P}\big\{\mathsf{X}^{-1}(A)\big|\mathsf{Z}\big\}\mathrm{P}\big\{\mathsf{Y}^{-1}(B)\big|\mathsf{Z}\big\}. & (14)\end{aligned}$$

(ii): See [18, Lem. 5.5.6].
(iii): See [18, Lem. 5.5.7].
(iv): Using Prop. (i) we have $\mathrm{I}(\mathsf{X};\mathsf{Z}|\mathsf{Y}) = 0$ and by Prop. (iii) it follows that

$$\begin{aligned}\mathrm{I}(\mathsf{X};\mathsf{Z}) &\leq \mathrm{I}(\mathsf{X};\mathsf{YZ}) & (15)\\ &= \mathrm{I}(\mathsf{X};\mathsf{Y}) + \mathrm{I}(\mathsf{X};\mathsf{Z}|\mathsf{Y}) = \mathrm{I}(\mathsf{X};\mathsf{Y}). & \blacksquare\end{aligned}$$

Occasionally we will interpret a probability measure on a finite space $\mathcal{M}$ as a vector in $[0,1]^\mathcal{M}$, equipped with the Borel σ-algebra. We will use the L_∞-distance on this space.

Definition 9. *For two probability measures* μ *and* ν *on a finite space* $\mathcal{M}$*, their distance is defined as the* L_∞*-distance* $\mathrm{d}(\mu,\nu) := \max_{m\in\mathcal{M}} |\mu(m) - \nu(m)|$*. The diameter of* $A \subseteq [0,1]^\mathcal{M}$ *is defined as* $\operatorname{diam}(A) = \sup_{\mu,\nu\in A} \mathrm{d}(\mu,\nu)$.

Lemma 10 ([20, Lem. 2.7]). *For two probability measures* μ *and* ν *on a finite space* $\mathcal{M}$ *with* $\mathrm{d}(\mu,\nu) \leq \varepsilon \leq \frac{1}{2}$ *the inequality* $|\mathrm{H}(\mu) - \mathrm{H}(\nu)| \leq -\varepsilon|\mathcal{M}|\log\varepsilon$ *holds.*

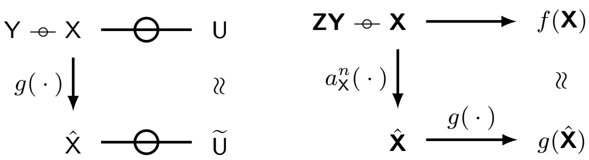
IV. Proof of $\mathcal{R}_{\mathrm{IB}} \subseteq \overline{\mathcal{R}}$

For finite spaces $\mathcal{S}_\mathsf{Y}$, $\mathcal{S}_\mathsf{X}$, and $\mathcal{S}_\mathsf{U}$, the statement $\mathcal{R}_{\mathrm{IB}} \subseteq \overline{\mathcal{R}}$ is well known, cf., [10, Sec. IV], [11, Sec. III.F]. We restate it in the form of the following lemma.

Lemma 11. *For random variables* Y*,* X*, and* U *with finite* $\mathcal{S}_\mathsf{Y}$*,* $\mathcal{S}_\mathsf{X}$*, and* $\mathcal{S}_\mathsf{U}$*, assume that* $\mathsf{Y} \multimap \mathsf{X} \multimap \mathsf{U}$ *holds. Then, for any* $\varepsilon > 0$*, there exists* $n \in \mathbb{N}$ *and a function* $f \colon \mathcal{S}_\mathsf{X}^n \to \mathcal{M}$ *with* $\frac{1}{n}\log|\mathcal{M}| \leq \mathrm{I}(\mathsf{X};\mathsf{U}) + \varepsilon$ *such that* $\frac{1}{n}\mathrm{I}\big(\mathbf{Y}; f(\mathbf{X})\big) \geq \mathrm{I}(\mathsf{Y};\mathsf{U}) - \varepsilon$.

In a first step, we will utilize Lem. 11 to show $\mathcal{R}_{\mathrm{IB}} \subseteq \overline{\mathcal{R}}$ for an arbitrary alphabet $\mathcal{S}_\mathsf{X}$, i. e., we wish to prove the following Proposition 12, lifting the restriction $|\mathcal{S}_\mathsf{X}| < \infty$.

Proposition 12. *For random variables* Y*,* X*, and* U *with finite* $\mathcal{S}_\mathsf{Y}$ *and* $\mathcal{S}_\mathsf{U}$*, assume that* $\mathsf{Y} \multimap \mathsf{X} \multimap \mathsf{U}$ *holds. Then, for any*



(a) $\mathcal{R}_{\mathrm{IB}} \subseteq \overline{\mathcal{R}}$. (b) $\mathcal{R} \subseteq \overline{\mathcal{R}_{\mathrm{IB}}}$.

Fig. 1: Illustrations.

$\varepsilon > 0$, there exists $n \in \mathbb{N}$ and a function $f\colon \mathcal{S}_{\mathsf{X}}^n \to \mathcal{M}$ with $\frac{1}{n}\log|\mathcal{M}| \le \mathrm{I}(\mathsf{X};\mathsf{U}) + \varepsilon$ such that

$$\frac{1}{n}\mathrm{I}\big(\mathbf{Y}; f(\mathbf{X})\big) \ge \mathrm{I}(\mathsf{Y};\mathsf{U}) - \varepsilon. \tag{16}$$

Remark 2. Considering that both definitions of achievability (Defs. 1 and 2) only rely on the notion of mutual information, one may assume that Def. 6 can be used to directly infer Proposition 12 from Lem. 11. However, this is not the case. For an arbitrary discretization $a_{\mathsf{X}}(\mathsf{X})$ of X, we do have $\mathrm{I}(a_{\mathsf{X}}(\mathsf{X});\mathsf{U}) \le \mathrm{I}(\mathsf{X};\mathsf{U})$. However, the Markov chain $\mathsf{Y} -\!\circ\!- a_{\mathsf{X}}(\mathsf{X}) -\!\circ\!- \mathsf{U}$ does not hold in general. To circumvent this problem, we will use a discrete random variable $\hat{\mathsf{X}} = g(\mathsf{X})$ with an appropriate quantizer g and construct a new random variable $\widetilde{\mathsf{U}}$, satisfying the Markov chain $\mathsf{Y} -\!\circ\!- \hat{\mathsf{X}} -\!\circ\!- \widetilde{\mathsf{U}}$ such that $\mathrm{I}(\mathsf{Y};\widetilde{\mathsf{U}})$ is close to $\mathrm{I}(\mathsf{Y};\mathsf{U})$. Fig. 1a illustrates this strategy. We choose the quantizer g based on the conditional probability distribution of U given X, i. e., quantization based on $\kappa_{\mathsf{U}|\mathsf{X}}$ using L_∞-distance (cf. Def. 9). Subsequently, we will use that, by Lem. 10, a small L_∞-distance guarantees a small gap in terms of information measures.

Proof of Proposition 12. Let μ_{YXU} be a probability measure on $\Omega := \mathcal{S}_{\mathsf{Y}} \times \mathcal{S}_{\mathsf{X}} \times \mathcal{S}_{\mathsf{U}}$, such that $\mathsf{Y} -\!\circ\!- \mathsf{X} -\!\circ\!- \mathsf{U}$ holds. Fix $0 < \delta \le \frac{1}{2}$ and find a finite, measurable partition $(P_i)_{i\in\mathcal{I}}$ of the space of probability measures on $\mathcal{S}_{\mathsf{U}}$ such that for every $i \in \mathcal{I}$ we have $\operatorname{diam}(P_i) \le \delta$ and fix some $\nu_i \in P_i$ for every $i \in \mathcal{I}$. Define the random variable $\hat{\mathsf{X}}\colon \Omega \to \mathcal{I}$ as $\hat{\mathsf{X}} = i$ if $\kappa_{\mathsf{U}|\mathsf{X}} \in P_i$. The random variable $\hat{\mathsf{X}}$ is $\sigma(\mathsf{X})$-measurable (see Appendix A). We can therefore find a measurable function g such that $\hat{\mathsf{X}} = g(\mathsf{X})$ by the factorization lemma [19, Corollary 1.97]. Define the new probability space $\Omega \times \bigtimes_{i\in\mathcal{I}} \mathcal{S}_{\mathsf{U}}$, equipped with the probability measure $\mu_{\mathsf{YXU}\widetilde{\mathsf{U}}_{\mathcal{I}}} := \mu_{\mathsf{YXU}} \times \bigtimes_{i\in\mathcal{I}} \nu_i$. Slightly abusing notation, we define the random variables $\mathsf{Y}, \mathsf{X}, \mathsf{U}$, and $\widetilde{\mathsf{U}}_i$ (for every $i \in \mathcal{I}$) as the according projections. We also use $\hat{\mathsf{X}} = g(\mathsf{X})$ and define the random variable $\widetilde{\mathsf{U}} = \widetilde{\mathsf{U}}_{\hat{\mathsf{X}}}$. From this construction we have $\mu_{\mathsf{YXU}\widetilde{\mathsf{U}}_{\mathcal{I}}}$-a. e. the equality of measures $\kappa_{\widetilde{\mathsf{U}}|\hat{\mathsf{X}}} = \kappa_{\widetilde{\mathsf{U}}|\mathsf{X}} = \nu_{\hat{\mathsf{X}}}$, as well as $\mathsf{Y} -\!\circ\!- \hat{\mathsf{X}} -\!\circ\!- \widetilde{\mathsf{U}}$ and $\mathsf{Y} -\!\circ\!- \mathsf{X} -\!\circ\!- \widetilde{\mathsf{U}}$ (see Appendix B). Therefore, we have $\mu_{\mathsf{YXU}\widetilde{\mathsf{U}}_{\mathcal{I}}}$-a. e.

$$\mathrm{d}(\kappa_{\widetilde{\mathsf{U}}|\hat{\mathsf{X}}}, \kappa_{\mathsf{U}|\mathsf{X}}) \le \delta, \text{ and } \qquad \mathrm{d}(\kappa_{\widetilde{\mathsf{U}}|\mathsf{X}}, \kappa_{\mathsf{U}|\mathsf{X}}) \le \delta, \tag{17}$$

by $\kappa_{\widetilde{\mathsf{U}}|\hat{\mathsf{X}}} = \kappa_{\widetilde{\mathsf{U}}|\mathsf{X}} = \nu_{\hat{\mathsf{X}}}$ and $\kappa_{\mathsf{U}|\mathsf{X}}, \nu_{\hat{\mathsf{X}}} \in P_{\hat{\mathsf{X}}}$. Thus, for any $u \in \mathcal{S}_{\mathsf{U}}$,

$$\begin{aligned}\mu_{\mathsf{U}}(u) &= \int \kappa_{\mathsf{U}|\mathsf{X}}(\,\cdot\,; u)\, d\mu_{\mathsf{YXU}} &(18)\\ &\le \int (\kappa_{\widetilde{\mathsf{U}}|\mathsf{X}}(\,\cdot\,; u) + \delta)\, d\mu_{\mathsf{YXU}\widetilde{\mathsf{U}}_{\mathcal{I}}} = \mu_{\widetilde{\mathsf{U}}}(u) + \delta &(19)\end{aligned}$$

and, by the same argument, $\mu_{\mathsf{U}}(u) \ge \mu_{\widetilde{\mathsf{U}}}(u) - \delta$, i. e., in total,

$$\mathrm{d}(\mu_{\mathsf{U}}, \mu_{\widetilde{\mathsf{U}}}) \le \delta. \tag{20}$$

Thus, we obtain

$$\begin{aligned}\mathrm{I}(\mathsf{X};\mathsf{U}) &= \mathrm{H}(\mu_{\mathsf{U}}) - \mathrm{H}(\mathsf{U}|\mathsf{X}) &(21)\\ &\overset{(20)}{\ge} \mathrm{H}\big(\mu_{\widetilde{\mathsf{U}}}\big) + \delta|\mathcal{S}_{\mathsf{U}}|\log\delta - \int \mathrm{H}\big(\kappa_{\mathsf{U}|\mathsf{X}}\big)\, d\mu_{\mathsf{YXU}} &(22)\\ &\overset{(17)}{\ge} \mathrm{H}\big(\mu_{\widetilde{\mathsf{U}}}\big) + 2\delta|\mathcal{S}_{\mathsf{U}}|\log\delta - \int \mathrm{H}\big(\kappa_{\widetilde{\mathsf{U}}|\hat{\mathsf{X}}}\big)\, d\mu_{\mathsf{YXU}\widetilde{\mathsf{U}}_{\mathcal{I}}} &(23)\\ &= \mathrm{I}(\hat{\mathsf{X}};\widetilde{\mathsf{U}}) + 2\delta|\mathcal{S}_{\mathsf{U}}|\log\delta, &(24)\end{aligned}$$

where (21) and (24) follow from Prop. (ii) of Lem. 8, and in both (22) and (23) we used Lem. 10. From $\mathsf{Y} -\!\circ\!- \mathsf{X} -\!\circ\!- \mathsf{U}$ and Prop. (i) of Lem. 8, we know that μ_{YXU}-a. e., we have the equality of measures $\kappa_{\mathsf{YU}|\mathsf{X}} = \kappa_{\mathsf{Y}|\mathsf{X}} \times \kappa_{\mathsf{U}|\mathsf{X}}$. Using this equality in (26) we obtain

$$\begin{aligned}\mu_{\mathsf{YU}}(y \times u) &= \int \kappa_{\mathsf{YU}|\mathsf{X}}(\,\cdot\,; y \times u)\, d\mu_{\mathsf{YXU}} &(25)\\ &= \int \kappa_{\mathsf{Y}|\mathsf{X}}(\,\cdot\,; y)\kappa_{\mathsf{U}|\mathsf{X}}(\,\cdot\,; u)\, d\mu_{\mathsf{YXU}} &(26)\\ &\overset{(17)}{\le} \int \kappa_{\mathsf{Y}|\mathsf{X}}(\,\cdot\,; y)(\kappa_{\widetilde{\mathsf{U}}|\mathsf{X}}(\,\cdot\,; u) + \delta)\, d\mu_{\mathsf{YXU}\widetilde{\mathsf{U}}_{\mathcal{I}}} &(27)\\ &\le \int \kappa_{\mathsf{Y}\widetilde{\mathsf{U}}|\mathsf{X}}(\,\cdot\,; y \times u)\, d\mu_{\mathsf{YXU}\widetilde{\mathsf{U}}_{\mathcal{I}}} + \delta &(28)\\ &= \mu_{\mathsf{Y}\widetilde{\mathsf{U}}}(y \times u) + \delta, &(29)\end{aligned}$$

where (25) and (29) follow from the defining property of conditional probability, part 2 of Def. 4, and (28) follows from $\mathsf{Y} -\!\circ\!- \mathsf{X} -\!\circ\!- \widetilde{\mathsf{U}}$ and Prop. (i) of Lem. 8. By the same argument, one can show that $\mu_{\mathsf{YU}}(y \times u) \ge \mu_{\mathsf{Y}\widetilde{\mathsf{U}}}(y \times u) - \delta$. Therefore, in total, $\mathrm{d}(\mu_{\mathsf{YU}}, \mu_{\mathsf{Y}\widetilde{\mathsf{U}}}) \le \delta$ and, by Lem. 10,

$$|\mathrm{H}(\mathsf{YU}) - \mathrm{H}(\mathsf{Y}\widetilde{\mathsf{U}})| \le -\delta|\mathcal{S}_{\mathsf{Y}}||\mathcal{S}_{\mathsf{U}}|\log\delta. \tag{30}$$

Thus, the mutual information can be bounded by

$$\begin{aligned}\mathrm{I}(\mathsf{Y};\mathsf{U}) &= \mathrm{H}(\mathsf{Y}) + \mathrm{H}(\mathsf{U}) - \mathrm{H}(\mathsf{YU}) &(31)\\ &\overset{(20)}{\le} \mathrm{H}(\mathsf{Y}) + \mathrm{H}(\widetilde{\mathsf{U}}) - \delta|\mathcal{S}_{\mathsf{U}}|\log\delta - \mathrm{H}(\mathsf{YU}) &(32)\\ &\overset{(30)}{\le} \mathrm{I}(\mathsf{Y};\widetilde{\mathsf{U}}) - \delta(|\mathcal{S}_{\mathsf{Y}}| + 1)|\mathcal{S}_{\mathsf{U}}|\log\delta &(33)\\ &\le \mathrm{I}(\mathsf{Y};\widetilde{\mathsf{U}}) - 2\delta|\mathcal{S}_{\mathsf{Y}}||\mathcal{S}_{\mathsf{U}}|\log\delta, &(34)\end{aligned}$$

where we applied Lem. 10 in (32) and (33). We apply Lem. 11 to the three random variables $\mathsf{Y}, \hat{\mathsf{X}}$, and $\widetilde{\mathsf{U}}$ and obtain a function $\hat{f}\colon \mathcal{I}^n \to \mathcal{M}$ with $\frac{1}{n}\mathrm{I}\big(\mathbf{Y}; \hat{f}(\hat{\mathbf{X}})\big) \ge \mathrm{I}(\mathsf{Y};\widetilde{\mathsf{U}}) - \delta$ and

$$\frac{1}{n}\log|\mathcal{M}| \le \mathrm{I}(\hat{\mathsf{X}};\widetilde{\mathsf{U}}) + \delta \overset{(24)}{\le} \mathrm{I}(\mathsf{X};\mathsf{U}) + \delta - 2\delta|\mathcal{S}_{\mathsf{U}}|\log\delta. \tag{35}$$

We have $\hat{\mathbf{X}} = g^n \circ \mathbf{X}$ and defining $f := \hat{f} \circ g^n$, we obtain

$$\begin{aligned}\frac{1}{n}\mathrm{I}(\mathbf{Y}; f(\mathbf{X})) = \frac{1}{n}\mathrm{I}(\mathbf{Y}; \hat{f}(\hat{\mathbf{X}})) &\ge \mathrm{I}(\mathsf{Y};\widetilde{\mathsf{U}}) - \delta &(36)\\ &\overset{(34)}{\ge} \mathrm{I}(\mathsf{Y};\mathsf{U}) + 2\delta|\mathcal{S}_{\mathsf{Y}}||\mathcal{S}_{\mathsf{U}}|\log\delta - \delta. &(37)\end{aligned}$$

Choosing δ such that $\varepsilon \ge -2\delta|\mathcal{S}_{\mathsf{Y}}||\mathcal{S}_{\mathsf{U}}|\log\delta + \delta$ completes the proof. ∎

We can now complete the proof by showing the following lemma.

Lemma 13. *$\mathcal{R}_{\mathrm{IB}} \subseteq \overline{\mathcal{R}}$.*

Proof. Assuming $(S,R) \in \mathcal{R}_{\mathrm{IB}}$, choose μ_{YXU} according to Def. 2. Clearly $\mathrm{I}(\mathsf{X};\mathsf{U}) < \infty$ to satisfy (3) and thus also $\mathrm{I}(\mathsf{Y};\mathsf{U}) < \infty$ by Prop. (iv) of Lem. 8 as $\mathsf{Y} -\!\circ\!- \mathsf{X} -\!\circ\!- \mathsf{U}$ holds. Pick $\varepsilon > 0$, select functions a_{X}, a_{U} such that $\mathrm{I}\big(a_{\mathsf{X}}(\mathsf{X}); a_{\mathsf{U}}(\mathsf{U})\big) \geq \mathrm{I}(\mathsf{X};\mathsf{U}) - \varepsilon$, and select functions b_{Y}, b_{U} such that $\mathrm{I}\big(b_{\mathsf{Y}}(\mathsf{Y}); b_{\mathsf{U}}(\mathsf{U})\big) \geq \mathrm{I}(\mathsf{Y};\mathsf{U}) - \varepsilon$ (cf. (7)). Using $\hat{\mathsf{U}} := \big(a_{\mathsf{U}}(\mathsf{U}), b_{\mathsf{U}}(\mathsf{U})\big)$ and $\hat{\mathsf{Y}} := b_{\mathsf{Y}}(\mathsf{Y})$, we have

$$0 = \mathrm{I}(\mathsf{Y};\mathsf{U}|\mathsf{X}) = \sup_{c_{\mathsf{Y}}, c_{\mathsf{U}}} \mathrm{I}(c_{\mathsf{Y}}(\mathsf{Y}); c_{\mathsf{U}}(\mathsf{U})|\mathsf{X}) \geq \mathrm{I}(\hat{\mathsf{Y}};\hat{\mathsf{U}}|\mathsf{X}) \geq 0 \quad (38)$$

as well as

$$\mathrm{I}(\mathsf{X};\mathsf{U}) = \sup_{c_{\mathsf{X}}, c_{\mathsf{U}}} \mathrm{I}\big(c_{\mathsf{X}}(\mathsf{X}); c_{\mathsf{U}}(\mathsf{U})\big) \quad (39)$$

$$\geq \sup_{c_{\mathsf{X}}} \mathrm{I}(c_{\mathsf{X}}(\mathsf{X}); \hat{\mathsf{U}}) = \mathrm{I}(\mathsf{X};\hat{\mathsf{U}}), \text{ and} \quad (40)$$

$$\mathrm{I}(\mathsf{Y};\mathsf{U}) - \varepsilon \leq \mathrm{I}\big(b_{\mathsf{Y}}(\mathsf{Y}); b_{\mathsf{U}}(\mathsf{U})\big) \leq \mathrm{I}(\hat{\mathsf{Y}};\hat{\mathsf{U}}). \quad (41)$$

We apply Proposition 12, substituting $\hat{\mathsf{U}} \to \mathsf{U}$ and $\hat{\mathsf{Y}} \to \mathsf{Y}$. Proposition 12 guarantees the existence of a function $f\colon \mathcal{S}_{\mathsf{X}}^n \to \mathcal{M}$ with $\frac{1}{n}\log|\mathcal{M}| \leq \mathrm{I}(\mathsf{X};\hat{\mathsf{U}}) + \varepsilon \overset{(40)}{\leq} \mathrm{I}(\mathsf{X};\mathsf{U}) + \varepsilon \overset{(3)}{\leq} R + \varepsilon$ and

$$\frac{1}{n}\mathrm{I}(\mathbf{Y}; f(\mathbf{X})) = \frac{1}{n}\sup_{c_{\mathbf{Y}}} \mathrm{I}(c_{\mathbf{Y}} \circ \mathbf{Y}; f(\mathbf{X})) \quad (42)$$

$$\geq \frac{1}{n}\mathrm{I}(b_{\mathsf{Y}}^n \circ \mathbf{Y}; f(\mathbf{X})) = \frac{1}{n}\mathrm{I}(\hat{\mathbf{Y}}; f(\mathbf{X})) \quad (43)$$

$$\overset{(16)}{\geq} \mathrm{I}(\hat{\mathsf{Y}};\hat{\mathsf{U}}) - \varepsilon \overset{(41)}{\geq} \mathrm{I}(\mathsf{Y};\mathsf{U}) - 2\varepsilon \overset{(4)}{\geq} S - 2\varepsilon. \quad (44)$$

Thus, $(S - 2\varepsilon, R - \varepsilon) \in \mathcal{R}$ and therefore $(S, R) \in \overline{\mathcal{R}}$. ∎

V. Proof of $\mathcal{R} \subseteq \overline{\mathcal{R}_{\mathrm{IB}}}$

We start with the well-known result $\mathcal{R}_{\mathrm{IB}} \subseteq \overline{\mathcal{R}}$ for finite spaces $\mathcal{S}_{\mathsf{Y}}$, $\mathcal{S}_{\mathsf{X}}$, and $\mathcal{S}_{\mathsf{U}}$, cf., [10, Sec. IV], [11, Sec. III.F]. The statement is rephrased in the following lemma.

Lemma 14. *Assume that the spaces $\mathcal{S}_{\mathsf{Y}}$ and $\mathcal{S}_{\mathsf{X}}$ are both finite and μ_{YX} is fixed. For some $n \in \mathbb{N}$, let $f\colon \mathcal{S}_{\mathsf{X}}^n \to \mathcal{M}$ be a function with $|\mathcal{M}| < \infty$. Then there exists a probability measure μ_{YXU}, extending μ_{YX}, such that $\mathcal{S}_{\mathsf{U}}$ is finite, $\mathsf{Y} \mathrel{-\!\!\circ\!\!-} \mathsf{X} \mathrel{-\!\!\circ\!\!-} \mathsf{U}$, and*

$$\mathrm{I}(\mathsf{X};\mathsf{U}) \leq \frac{1}{n}\log|\mathcal{M}|, \quad (45)$$

$$\mathrm{I}(\mathsf{Y};\mathsf{U}) \geq \frac{1}{n}\mathrm{I}(\mathbf{Y}; f(\mathbf{X})). \quad (46)$$

We can slightly strengthen Lem. 14.

Corollary 15. *Assume that, in the setting of Lem. 14, we are given μ_{ZYX} on $\mathcal{S}_{\mathsf{Z}} \times \mathcal{S}_{\mathsf{Y}} \times \mathcal{S}_{\mathsf{X}}$, extending μ_{YX}, where $\mathcal{S}_{\mathsf{Z}}$ is arbitrary, not necessarily finite. Then there exists a probability measure μ_{ZYXU}, extending μ_{ZYX}, such that $\mathcal{S}_{\mathsf{U}}$ is finite and $\mathsf{ZY} \mathrel{-\!\!\circ\!\!-} \mathsf{X} \mathrel{-\!\!\circ\!\!-} \mathsf{U}$, (45), and (46) hold.*

Proof. Apply Lem. 14 to obtain μ_{YXU} on $\mathcal{S}_{\mathsf{Y}} \times \mathcal{S}_{\mathsf{X}} \times \mathcal{S}_{\mathsf{U}}$ satisfying (45), (46), and $\mathsf{Y} \mathrel{-\!\!\circ\!\!-} \mathsf{X} \mathrel{-\!\!\circ\!\!-} \mathsf{U}$. We define μ_{ZYXU} by

$$\mu_{\mathsf{ZYXU}}(A \times y \times x \times u) = \frac{\mu_{\mathsf{ZYX}}(A \times y \times x)}{\mu_{\mathsf{YX}}(y \times x)}\mu_{\mathsf{YXU}}(y \times x \times u) \quad (47)$$

for any $(y, x, u) \in \mathcal{S}_{\mathsf{Y}} \times \mathcal{S}_{\mathsf{X}} \times \mathcal{S}_{\mathsf{U}}$ and $A \in \mathcal{A}_{\mathsf{Z}}$. Pick arbitrary $A \in \mathcal{A}_{\mathsf{Z}}$, $y \in \mathcal{S}_{\mathsf{Y}}$, and $u \in \mathcal{S}_{\mathsf{U}}$. The Markov chain $\mathsf{ZY} \mathrel{-\!\!\circ\!\!-} \mathsf{X} \mathrel{-\!\!\circ\!\!-} \mathsf{U}$ now follows as the events $\mathsf{Z}^{-1}(A) \cap \mathsf{Y}^{-1}(y)$ and $\mathsf{U}^{-1}(u)$ are independent given $\mathsf{X}^{-1}(x)$ for any $x \in \mathcal{S}_{\mathsf{X}}$ (cf. Rmk. 1). ∎

Again, we proceed by extending Cor. 15, lifting the restriction that $\mathcal{S}_{\mathsf{X}}$ is finite and obtain the following proposition.

Proposition 16. *Given a probability measure μ_{ZYX} as in Cor. 15, assume that $|\mathcal{S}_{\mathsf{Y}}| < \infty$. For some $n \in \mathbb{N}$, let $f\colon \mathcal{S}_{\mathsf{X}}^n \to \mathcal{M}$ be a function with $|\mathcal{M}| < \infty$. Then, for any $\varepsilon > 0$, there exists a probability measure μ_{ZYXU}, extending μ_{ZYX} with $\mathsf{ZY} \mathrel{-\!\!\circ\!\!-} \mathsf{X} \mathrel{-\!\!\circ\!\!-} \mathsf{U}$ and*

$$\mathrm{I}(\mathsf{X};\mathsf{U}) \leq \frac{1}{n}\log|\mathcal{M}| \quad (48)$$

$$\mathrm{I}(\mathsf{Y};\mathsf{U}) \geq \frac{1}{n}\mathrm{I}\big(\mathbf{Y}; f(\mathbf{X})\big) - \varepsilon. \quad (49)$$

Remark 3. In contrast to Proposition 12, Proposition 16 could be proved by the usual single-letterization + time-sharing strategy, by showing that the necessary Markov chains hold. However, we will rely on the discrete case (Lem. 14) and showcase a technique to lift it to general alphabets.

Remark 4. In the proof of Proposition 16, we face a similar problem as outlined in Rmk. 2. We need to construct a function $g(\hat{\mathbf{X}})$ of a "per-letter" quantization $\hat{\mathbf{X}} := a_{\mathsf{X}}^n(\mathbf{X})$, that is close to $f(\mathbf{X})$ in distribution. Fig. 1b provides a sketch.

Proof of Proposition 16. We can partition $\mathcal{S}_{\mathsf{X}}^n = \bigcup_{m \in \mathcal{M}} \mathcal{Q}_m$ into finitely many measurable, mutually disjoint sets $\mathcal{Q}_m := f^{-1}(m)$, $m \in \mathcal{M}$. We want to approximate the sets $\mathcal{Q}_m$ by a finite union of rectangles in the semiring [19, Def. 1.9] $\Xi := \big\{\mathcal{B} : \mathcal{B} = \bigtimes_{i=1}^n B_i \text{ with } B_i \in \mathcal{A}_{\mathsf{X}}\big\}$. We choose $\delta > 0$, which will be specified later. According to [19, Thm. 1.65(ii)], we obtain $\mathcal{B}^{(m)} := \bigcup_{k=1}^K \mathcal{B}_k^{(m)}$ for each $m \in \mathcal{M}$, where $\mathcal{B}_k^{(m)} \in \Xi$ are mutually disjoint sets, satisfying $\mu_{\mathbf{X}}(\mathcal{B}^{(m)} \triangle \mathcal{Q}_m) \leq \delta$. Since $\mathcal{B}_k^{(m)} \in \Xi$, we have $\mathcal{B}_k^{(m)} = \bigtimes_{i=1}^n B_{k,i}^{(m)}$ for some $B_{k,i}^{(m)} \in \mathcal{A}_{\mathsf{X}}$. We can construct functions a_{X} and g such that $g \circ a_{\mathsf{X}}^n(\boldsymbol{x}) = m$ whenever $\boldsymbol{x} \in \mathcal{B}^{(m)}$ and $\boldsymbol{x} \notin \mathcal{B}^{(\not m)}$ with $\mathcal{B}^{(\not m)} := \bigcup_{m' \neq m} \mathcal{B}^{(m')}$. Indeed, we obtain a_{X} by finding a measurable partition of $\mathcal{S}_{\mathsf{X}}$ that is finer than $(B_{k,i}^{(m)}, (B_{k,i}^{(m)})^{\mathrm{c}})$ for all i, k, m. For fixed $m \in \mathcal{M}$,

$$\mathcal{Q}_m \subseteq \mathcal{Q}_m \cup \big(\mathcal{B}^{(m)} \setminus \mathcal{B}^{(\not m)}\big) \quad (50)$$

$$\subseteq \big(\mathcal{B}^{(m)} \setminus \mathcal{B}^{(\not m)}\big) \cup \big(\mathcal{Q}_m \setminus \mathcal{B}^{(m)}\big) \cup \bigcup_{m' \neq m} \mathcal{Q}_m \cap \mathcal{B}^{(m')} \quad (51)$$

$$\subseteq \big(\mathcal{B}^{(m)} \setminus \mathcal{B}^{(\not m)}\big) \cup \big(\mathcal{Q}_m \triangle \mathcal{B}^{(m)}\big) \cup \bigcup_{m' \neq m} \mathcal{B}^{(m')} \setminus \mathcal{Q}_{m'} \quad (52)$$

$$\subseteq \big(\mathcal{B}^{(m)} \setminus \mathcal{B}^{(\not m)}\big) \cup \bigcup_{m'} \mathcal{B}^{(m')} \triangle \mathcal{Q}_{m'}, \quad (53)$$

where we used the fact that $\mathcal{Q}_m \cap \mathcal{Q}_{m'} = \varnothing$ for $m \neq m'$ in (52). Using $\hat{\mathsf{X}} := a_{\mathsf{X}}(\mathsf{X})$, we obtain for any $\boldsymbol{y} \in \mathcal{S}_{\mathsf{Y}}^n$

$$\mu_{\mathbf{Y}f(\mathbf{X})}(\boldsymbol{y} \times m) = \mu_{\mathbf{YX}}(\boldsymbol{y} \times \mathcal{Q}_m) \quad (54)$$

$$\overset{(53)}{\leq} \mu_{\mathbf{YX}}\big(\boldsymbol{y} \times (\mathcal{B}^{(m)} \setminus \mathcal{B}^{(\not m)})\big) + \sum_{m'} \mu_{\mathbf{X}}(\mathcal{B}^{(m')} \triangle \mathcal{Q}_{m'}) \quad (55)$$

$$\leq \mu_{\mathbf{Y}g(\hat{\mathbf{X}})}(\boldsymbol{y} \times m) + |\mathcal{M}|\delta. \quad (56)$$

On the other hand, we have

$$\mu_{\mathbf{Y}f(\mathbf{X})}(\boldsymbol{y} \times m) = \mu_{\mathbf{Y}}(\boldsymbol{y}) - \sum_{m' \neq m} \mu_{\mathbf{Y}f(\mathbf{X})}(\boldsymbol{y} \times m') \quad (57)$$

$$\overset{(56)}{\geq} \mu_{\mathbf{Y}}(\boldsymbol{y}) - \sum_{m' \neq m} \Big(\mu_{\mathbf{Y}g(\hat{\mathbf{X}})}(\boldsymbol{y} \times m') + |\mathcal{M}|\delta\Big) \quad (58)$$

$$\geq \mu_{\mathbf{Y}g(\hat{\mathbf{X}})}(\boldsymbol{y} \times m) - |\mathcal{M}|^2\delta. \quad (59)$$

We thus obtain $\mathrm{d}(\mu_{\mathbf{Y}f(\mathbf{X})}, \mu_{\mathbf{Y}g(\hat{\mathbf{X}})}) \leq |\mathcal{M}|^2\delta$. This also implies $\mathrm{d}(\mu_{f(\mathbf{X})}, \mu_{g(\hat{\mathbf{X}})}) \leq |\mathcal{S}_{\mathsf{Y}}|^n|\mathcal{M}|^2\delta$. Assume $|\mathcal{S}_{\mathsf{Y}}|^n|\mathcal{M}|^2\delta \leq \frac{1}{2}$ and apply Cor. 15 substituting $\hat{\mathsf{X}} \to \mathsf{X}$, $\mathsf{XZ} \to \mathsf{Z}$, and the function

$g \to f$. This yields a random variable U with $\mathsf{XZY} \mathrel{-\!\!\circ\!\!-} \hat{\mathsf{X}} \mathrel{-\!\!\circ\!\!-} \mathsf{U}$,

$$\mathrm{I}(\hat{\mathsf{X}};\mathsf{U}) \leq \frac{1}{n}\log|\mathcal{M}|, \text{ and } \quad \mathrm{I}(\mathsf{Y};\mathsf{U}) \geq \frac{1}{n}\mathrm{I}\big(\mathbf{Y}; g(\hat{\mathbf{X}})\big). \tag{60}$$

We also obtain $\mathsf{ZY} \mathrel{-\!\!\circ\!\!-} \mathsf{X} \mathrel{-\!\!\circ\!\!-} \mathsf{U}$ due to

$$\begin{aligned} 0 &= \mathrm{I}(\mathsf{XZY};\mathsf{U}|\hat{\mathsf{X}}) && (61)\\ &= \mathrm{I}(\mathsf{XZY};\mathsf{U}) - \mathrm{I}(\mathsf{U};\hat{\mathsf{X}}) && (62)\\ &\geq \mathrm{I}(\mathsf{XZY};\mathsf{U}) - \mathrm{I}(\mathsf{U};\mathsf{X}) && (63)\\ &= \mathrm{I}(\mathsf{ZY};\mathsf{U}|\mathsf{X}) && (64)\\ &\geq 0, && (65) \end{aligned}$$

where (61) follows from $\mathsf{XZY} \mathrel{-\!\!\circ\!\!-} \hat{\mathsf{X}} \mathrel{-\!\!\circ\!\!-} \mathsf{U}$ using Prop. (i) of Lem. 8, (62) and (64) follow from Prop. (iii) of Lem. 8, (63) is a consequence of Def. 6, and we used Prop. (i) of Lem. 8 in (65). This also immediately implies $0 = \mathrm{I}(\mathsf{X};\mathsf{U}|\hat{\mathsf{X}})$ and hence

$$\begin{aligned} \frac{1}{n}\log|\mathcal{M}| &\overset{(60)}{\geq} \mathrm{I}(\hat{\mathsf{X}};\mathsf{U}) = \mathrm{I}(\hat{\mathsf{X}};\mathsf{U}) + \mathrm{I}(\mathsf{X};\mathsf{U}|\hat{\mathsf{X}}) && (66)\\ &= \mathrm{I}(\mathsf{X}\hat{\mathsf{X}};\mathsf{U}) = \mathrm{I}(\mathsf{X};\mathsf{U}), && (67) \end{aligned}$$

where we used Prop. (iii) of Lem. 8 in (67). We also have

$$\begin{aligned} \mathrm{I}(\mathsf{Y};\mathsf{U}) &\overset{(60)}{\geq} \frac{1}{n}\mathrm{I}\big(\mathbf{Y}; g(\hat{\mathbf{X}})\big) && (68)\\ &= \frac{1}{n}\Big(\mathrm{H}(\mathbf{Y}) + \mathrm{H}(g(\hat{\mathbf{X}})) - \mathrm{H}(\mathbf{Y}g(\hat{\mathbf{X}}))\Big) && (69)\\ &\geq \frac{1}{n}\mathrm{I}\big(\mathbf{Y}; f(\mathbf{X})\big) + \frac{1}{n}|\mathcal{S}_{\mathsf{Y}}|^n|\mathcal{M}|^3\delta\log(|\mathcal{M}|^2\delta)\\ &\qquad + \frac{1}{n}|\mathcal{S}_{\mathsf{Y}}|^n|\mathcal{M}|^3\delta\log(|\mathcal{S}_{\mathsf{Y}}|^n|\mathcal{M}|^2\delta) && (70)\\ &\geq \frac{1}{n}\mathrm{I}\big(\mathbf{Y}; f(\mathbf{X})\big) + \frac{2}{n}|\mathcal{S}_{\mathsf{Y}}|^n|\mathcal{M}|^3\delta\log(|\mathcal{M}|^2\delta) && (71) \end{aligned}$$

where we used Lem. 10 in (70). Select δ such that $\varepsilon \geq -\frac{2}{n}|\mathcal{S}_{\mathsf{Y}}|^n|\mathcal{M}|^3\delta\log(|\mathcal{M}|^2\delta)$. ∎

We can now finish the proof by showing the following lemma.

Lemma 17. *$\mathcal{R} \subseteq \overline{\mathcal{R}_{\mathrm{IB}}}$.*

Proof. Assume $(S,R) \in \mathcal{R}$ and choose $n \in \mathbb{N}$ and f, satisfying $\frac{1}{n}\log|\mathcal{M}| \leq R$ and (2). Choose any $\varepsilon > 0$ and find a_{Y} such that

$$\mathrm{I}\big(a_{\mathsf{Y}}^n(\mathbf{Y}); f(\mathbf{X})\big) \geq \mathrm{I}\big(\mathbf{Y}; f(\mathbf{X})\big) - \varepsilon \overset{(2)}{\geq} nS - \varepsilon. \tag{72}$$

This is possible by applying [18, Lem. 5.2.2] with the algebra that is generated by the rectangles (cf. the paragraph above [18, Lem. 5.5.1]). We apply Proposition 16, substituting $a_{\mathsf{Y}}(\mathsf{Y}) \to \mathsf{Y}$ and $\mathsf{Y} \to \mathsf{Z}$. For arbitrary $\varepsilon > 0$, Proposition 16 provides U with $\mathsf{Y}a_{\mathsf{Y}}(\mathsf{Y}) \mathrel{-\!\!\circ\!\!-} \mathsf{X} \mathrel{-\!\!\circ\!\!-} \mathsf{U}$ (i. e., $\mathsf{Y} \mathrel{-\!\!\circ\!\!-} \mathsf{X} \mathrel{-\!\!\circ\!\!-} \mathsf{U}$) and

$$\begin{aligned} \mathrm{I}(\mathsf{X};\mathsf{U}) &\leq \frac{1}{n}\log|\mathcal{M}| \leq R && (73)\\ \mathrm{I}(\mathsf{Y};\mathsf{U}) &\geq \mathrm{I}(a_{\mathsf{Y}}(\mathsf{Y});\mathsf{U}) && (74)\\ &\overset{(49)}{\geq} \frac{1}{n}\mathrm{I}\big(a_{\mathsf{Y}}^n(\mathbf{Y}); f(\mathbf{X})\big) - \varepsilon \overset{(72)}{\geq} S - 2\varepsilon. && (75) \end{aligned}$$

Hence, $(S - 2\varepsilon, R) \in \mathcal{R}_{\mathrm{IB}}$ and consequently $(S,R) \in \overline{\mathcal{R}_{\mathrm{IB}}}$. ∎

Appendix

A. $\hat{\mathsf{X}}$ is $\sigma(\mathsf{X})$-measurable

For $u \in \mathcal{S}_{\mathsf{U}}$ consider the $\sigma(\mathsf{X})$-measurable function $h_u := \kappa_{\mathsf{U}|\mathsf{X}}(\,\cdot\,;u)$ on $[0,1]$. We obtain the vector valued function $h := (h_u)_{u\in\mathcal{S}_{\mathsf{U}}}$ on $[0,1]^{|\mathcal{S}_{\mathsf{U}}|}$. This function h is $\sigma(\mathsf{X})$-measurable as every component is $\sigma(\mathsf{X})$-measurable. Thus, we have $\hat{\mathsf{X}}^{-1}(i) = h^{-1}(P_i) \in \sigma(\mathsf{X})$.

B. Distribution of $\widetilde{\mathsf{U}}$ and Conditional Independence

We will first show that $\mu_{\mathsf{YXU}\widetilde{\mathsf{U}}_{\mathcal{I}}}$-a. e.

$$\kappa_{\widetilde{\mathsf{U}}|\hat{\mathsf{X}}} = \kappa_{\widetilde{\mathsf{U}}|\mathsf{X}} = \nu_{\hat{\mathsf{X}}}. \tag{76}$$

Clearly, $\nu_{\hat{\mathsf{X}}}$ is a probability measure everywhere. Fixing $u \in \mathcal{S}_{\mathsf{U}}$, we need that $\nu_{\hat{\mathsf{X}}}(u)$ is $\sigma(\hat{\mathsf{X}})$-measurable, which is shown by the factorization lemma [19, Corollary 1.97], when writing $\nu_{\hat{\mathsf{X}}}(u) = \nu_{(\cdot)}(u) \circ \hat{\mathsf{X}}$. Also, this proves $\sigma(\mathsf{X})$-measurability as $\hat{\mathsf{X}}$ is $\sigma(\mathsf{X})$-measurable, i. e., $\sigma(\hat{\mathsf{X}}) \subseteq \sigma(\mathsf{X})$. It remains to show the defining property of conditional probability, part 2 of Def. 4. Choosing $B \in \sigma(\mathsf{X})$ and $u \in \mathcal{S}_{\mathsf{U}}$, we need to show that

$$\mathbb{E}[\mathbb{1}_B \nu_{\hat{\mathsf{X}}}(u)] = \mathbb{E}\Big[\mathbb{1}_B \mathbb{1}_{\{u\}}(\widetilde{\mathsf{U}})\Big]. \tag{77}$$

The statement for $B \in \sigma(\hat{\mathsf{X}})$ then follows by $\sigma(\hat{\mathsf{X}}) \subseteq \sigma(\mathsf{X})$, i. e., the $\sigma(\mathsf{X})$-measurability of $\hat{\mathsf{X}}$. We prove (77) by

$$\begin{aligned} \mathbb{E}[\mathbb{1}_B \nu_{\hat{\mathsf{X}}}(u)] &= \sum_{i\in\mathcal{I}} \mathbb{E}\big[\mathbb{1}_i(\hat{\mathsf{X}})\mathbb{1}_B \nu_i(u)\big] && (78)\\ &= \sum_{i\in\mathcal{I}} \nu_i(u)\mathbb{E}\big[\mathbb{1}_i(\hat{\mathsf{X}})\mathbb{1}_B\big] && (79)\\ &= \sum_{i\in\mathcal{I}} \mathbb{E}\Big[\mathbb{1}_u(\widetilde{\mathsf{U}}_i)\Big]\mathbb{E}\big[\mathbb{1}_i(\hat{\mathsf{X}})\mathbb{1}_B\big] && (80)\\ &= \sum_{i\in\mathcal{I}} \mathbb{E}\Big[\mathbb{1}_i(\hat{\mathsf{X}})\mathbb{1}_B\mathbb{1}_u(\widetilde{\mathsf{U}}_i)\Big] && (81)\\ &= \sum_{i\in\mathcal{I}} \mathbb{E}\Big[\mathbb{1}_i(\hat{\mathsf{X}})\mathbb{1}_B\mathbb{1}_u(\widetilde{\mathsf{U}})\Big] && (82)\\ &= \mathbb{E}\Big[\mathbb{1}_B\mathbb{1}_u(\widetilde{\mathsf{U}})\Big], && (83) \end{aligned}$$

where we used Fubini's theorem [19, Thm. 14.16] in (81).

To prove $\mathrm{I}(\mathsf{Y};\widetilde{\mathsf{U}}|\mathsf{X}) = 0$, we need to show that for every $y \in \mathcal{S}_{\mathsf{Y}}$, $u \in \mathcal{S}_{\mathsf{U}}$, and $B \in \sigma(\mathsf{X})$, we have

$$\int \mathbb{1}_B \kappa_{\mathsf{Y}|\mathsf{X}}(\,\cdot\,;y)\nu_{\hat{\mathsf{X}}}(u)\, d\mu_{\mathsf{YXU}} = \int \mathbb{1}_B\mathbb{1}_u(\widetilde{\mathsf{U}})\mathbb{1}_y(\mathsf{Y})\, d\mu_{\mathsf{YXU}\widetilde{\mathsf{U}}_{\mathcal{I}}} \tag{84}$$

and by integrating, we indeed obtain

$$\begin{aligned} &\int \mathbb{1}_B \kappa_{\mathsf{Y}|\mathsf{X}}(\,\cdot\,;y)\nu_{\hat{\mathsf{X}}}(u)\, d\mu_{\mathsf{YXU}} && (85)\\ &= \sum_{i\in\mathcal{I}} \int \mathbb{1}_B\mathbb{1}_i(\hat{\mathsf{X}})\kappa_{\mathsf{Y}|\mathsf{X}}(\,\cdot\,;y)\nu_i(u)\, d\mu_{\mathsf{YXU}} && (86)\\ &= \sum_{i\in\mathcal{I}} \nu_i(u) \int \mathbb{1}_B\mathbb{1}_i(\hat{\mathsf{X}})\kappa_{\mathsf{Y}|\mathsf{X}}(\,\cdot\,;y)\, d\mu_{\mathsf{YXU}} && (87)\\ &= \sum_{i\in\mathcal{I}} \int \mathbb{1}_u(\widetilde{\mathsf{U}}_i)\, d\mu_{\widetilde{\mathsf{U}}_{\mathcal{I}}} \int \mathbb{1}_B\mathbb{1}_i(\hat{\mathsf{X}})\mathbb{1}_y(\mathsf{Y})\, d\mu_{\mathsf{YXU}} && (88)\\ &= \sum_{i\in\mathcal{I}} \int \mathbb{1}_B\mathbb{1}_u(\widetilde{\mathsf{U}}_i)\mathbb{1}_i(\hat{\mathsf{X}})\mathbb{1}_y(\mathsf{Y})\, d\mu_{\mathsf{YXU}\widetilde{\mathsf{U}}_{\mathcal{I}}} && (89)\\ &= \sum_{i\in\mathcal{I}} \int \mathbb{1}_B\mathbb{1}_u(\widetilde{\mathsf{U}})\mathbb{1}_i(\hat{\mathsf{X}})\mathbb{1}_y(\mathsf{Y})\, d\mu_{\mathsf{YXU}\widetilde{\mathsf{U}}_{\mathcal{I}}} && (90)\\ &= \int \mathbb{1}_B\mathbb{1}_u(\widetilde{\mathsf{U}})\mathbb{1}_y(\mathsf{Y})\, d\mu_{\mathsf{YXU}\widetilde{\mathsf{U}}_{\mathcal{I}}}, && (91) \end{aligned}$$

where we used part 2 of Def. 4 in (88) and Fubini's theorem [19, Thm. 14.16] in (89). By replacing $\kappa_{\mathsf{Y}|\mathsf{X}}$ with $\kappa_{\mathsf{Y}|\hat{\mathsf{X}}}$ and using $B \in \sigma(\hat{\mathsf{X}})$, the same argument can be used to show $\mathrm{I}(\mathsf{Y}; \widetilde{\mathsf{U}}|\hat{\mathsf{X}}) = 0$.

Acknowledgment

The authors would like to thank Michael Meidlinger for providing inspiration for this work.

References

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 1999, pp. 368–377.

[2] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '00, ACM. New York, NY, USA: ACM, 2000, pp. 208–215.

[3] S. Gordon, H. Greenspan, and J. Goldberger, "Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, Oct 2003, pp. 370–377 vol.1.

[4] E. Schneidman, N. Slonim, N. Tishby, R. deRuyter van Steveninck, and W. Bialek, "Analyzing neural codes using the information bottleneck method," *Proc. of Advances in Neural Information Processing System (NIPS-13)*, 2002.

[5] G. Zeitler, R. Kötter, G. Bauch, and J. Widmer, "On quantizer design for soft values in the multiple-access relay channel," in *Proc. IEEE ICC 2009*, Jun. 2009.

[6] G. Zeitler, A. Singer, and G. Kramer, "Low-precision A/D conversion for maximum information rate in channels with memory," *IEEE Trans. Communications*, vol. 60, no. 9, pp. 2511–2521, 9 2012.

[7] A. Winkelbauer and G. Matz, "Joint network-channel coding for the asymmetric multiple-access relay channel," in *Proc. IEEE ICC 2012*, Jun. 2012, pp. 2485–2489.

[8] A. Winkelbauer, S. Farthofer, and G. Matz, "The rate-information trade-off for Gaussian vector channels," in *IEEE Int. Symp. Information Theory*, Honolulu, HI, USA, 6 2014.

[9] A. Winkelbauer and G. Matz, "On quantization of log-likelihood ratios for maximum mutual information," in *Proc. IEEE SPAWC*, 6 2015, pp. 316–320.

[10] G. Pichler, P. Piantanida, and G. Matz, "Distributed information-theoretic biclustering of two memoryless sources," in *Proc. 53rd Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 2015, pp. 426–433.

[11] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, Jan. 2014.

[12] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for gaussian variables," *Journal of Machine Learning Research*, vol. 6, pp. 165–188, Jan. 2005.

[13] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Advances in Neural Information Processing Systems 12*, S. Solla, T. Leen, and K. Müller, Eds. MIT Press, 2000, pp. 617–623.

[14] B. M. Kurkoski, "On the relationship between the kl means algorithm and the information bottleneck method," in *Proc. 11th International ITG Conference on Systems, Communications and Coding (SCC)*, Hamburg, Germany, 2 2017, pp. 1–6.

[15] M. B. Westover and J. A. O'Sullivan, "Achievable rates for pattern recognition," *IEEE Trans. Inf. Theory*, vol. 54, no. 1, pp. 299–320, Jan. 2008.

[16] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 533–542, Jul. 1986.

[17] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 33, no. 6, pp. 759–772, Nov. 1987.

[18] R. M. Gray, *Entropy and Information Theory*, 1st ed. Springer-Verlag, 2013.

[19] A. Klenke, *Probability Theory*, ser. Universitext. Springer-Verlag GmbH, 2013.

[20] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, Aug. 2011.