An Alphabet-Size Bound for the Information Bottleneck Function

Christoph Hirche*
*QMATH
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen, Denmark
Email: christoph.hirche@gmail.com

Andreas Winter†‡
†Grup d'Informació Quàntica, Departament de Física
Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain
‡Institució Catalana de la Recerca i d'Estudis Avançats
Pg. Lluis Companys, 23, 08010 Barcelona, Spain
Email: andreas.winter@uab.cat

***Abstract*—The information bottleneck function gives a measure of optimal preservation of correlation between some random variable X and some side information Y while compressing X into a new random variable W with bounded remaining correlation to X. As such, the information bottleneck has found many natural applications in machine learning, coding and video compression. The main objective in order to calculate the information bottleneck is to find the optimal representation on W. This could in principle be arbitrarily complicated, but fortunately it is known that the cardinality of W can be restricted as $|\mathcal{W}| \leq |\mathcal{X}|+1$ which makes the calculation possible for finite $|\mathcal{X}|$. Now, for many practical applications, e.g. in machine learning, X represents a potentially very large data space, while Y is from a comparably small set of labels. This raises the question whether the known cardinality bound can be improved in such situations. We show that the information bottleneck function can always be approximated up to an error $\delta(\epsilon, |\mathcal{Y}|)$ with a cardinality $|\mathcal{W}| \leq f(\epsilon, |\mathcal{Y}|)$, for explicitly given functions δ and f of an approximation parameter $\epsilon > 0$ and the cardinality of $\mathcal{Y}$.**

Finally, we generalize the known cardinality bounds to the case were some of the random variables represent quantum information.

I. Introduction

Given a joint probability distribution P_{XY}, where X is often interpreted as some kind of data and Y as some side information (e.g. labels in supervised machine learning), the information bottleneck function [1] quantifies the optimal compression of X into W below a certain threshold while preserving the maximal amount of correlation with the side information. This is closely related to the question of finding approximate sufficient statistics for P_{XY} and has therefore found a multitude of applications such as e.g. in investigating deep neural networks [2], [3], video processing [4], clustering [5] and polar coding [6]. Formally, the information bottleneck function of a given joint distribution P_{XY} of two random variables X and Y is defined as [1]

$$I_{XY}(R) := \max I(Y:W) \text{ s.t. } Y - X - W \text{ Markov chain,} \\ I(X:W) \leq R. \tag{1}$$

Here, X is considered the "data" and Y the "classifier", which is assumed to capture the objective truth about each data point, or at least reflect it on a training set.

It is known, and not difficult to prove using Caratheodory's theorem, that the above maximum is attained with $|\mathcal{W}| \leq |\mathcal{X}| + 1$, which makes the information bottleneck function at least in principle computable. By Caratheodory's theorem we refer to the fundamental geometric fact that if a point in an n-dimensional real affine space, w.l.o.g. $\mathbb{R}^n$, is a convex combination of any set $\mathcal{E}$, then it can be written as a convex combination of a subset of cardinality $n+1$ [7]. (Some earlier work states the bound $|\mathcal{W}| \leq |\mathcal{X}| + 2$, however directly using the results in [8] gives the slightly better bound, see also [9].) However, in the purported machine learning applications, X takes values in an enormously large alphabet, e.g. digital pictures of a certain format, which would require optimisation over channels $P_{W|X}$ with potentially megabyte-sized input and output, which is clearly infeasible. On the other hand, Y is by definition from a small set, for example letters or numbers in the task of recognizing written text.

It was found that in some cases deep neural networks approximate the information bottleneck function well, despite their output alphabet seemingly being too small, or at least much smaller than the above guarantee [3]. This raises the question whether the information bottleneck can generally be well approximated with a smaller alphabet size $|\mathcal{W}|$, and in particular one with bounded cardinality depending on Y rather than on X. In the present paper we answer this question affirmatively, by

1) providing an explicit method that reduces the cardinality of X, by compressing it into a random variable X' whose range is determined solely by $|\mathcal{Y}|$ and an error parameter $\epsilon > 0$, that allows to approximately recover the statistics of X conditioned on Y from X';
2) showing how this implies that the information bottleneck function of P_{XY} is well-approximated by that of the compressed $P_{X'Y}$; and
3) showing that the same tools imply that evaluating the information bottleneck function on the original P_{XY}, but with bounded cardinality $|\mathcal{W}| \leq f(\epsilon, |\mathcal{Y}|) \equiv N$, also gives a good approximation of the unbounded case.

Besides the usual information bottleneck function, recently also a version based on quantum information was considered [10], [11], [12]. Dimension bounds on auxiliary quantum registers in quantum information theory are generally a much harder problem since no good tools for investigating them are known. In the final section of this work, we will generalize the previously known bound and the ones developed in this work to particular intermediate settings where some random variables are classical and others are quantum.

II. Approximation via recovery

The following lemma shows how to obtain an approximately sufficient statistics for X with respect to Y, with an alphabet size that depends only on $|\mathcal{Y}|$ and on the accuracy of the approximation. The main tool is the existence of a recovery map that can approximately revert the compression.

Lemma 1 *Given a joint distribution P_{XY} of two random variables X and Y, and assuming that there exist N probability distributions $Q_1, \dots, Q_N$ on $\mathcal{Y}$, and a function $f : \mathcal{X} \longrightarrow [N]$ with the property that*

$$\forall x \quad \frac{1}{2}\|P_{Y|X=x} - Q_{f(x)}\|_1 \leq \epsilon, \tag{2}$$

for some $\epsilon > 0$. Then there exists a recovery channel $S : [N] \longrightarrow \mathcal{X}$ such that the Markov chain $Y - X - X' - \widehat{X}$ defined by $X' = f(X)$ and $P_{\widehat{X}|X'} = S$ satisfies $P_X = P_{\widehat{X}}$ and $\frac{1}{2}\|P_{XY} - P_{\widehat{X}Y}\|_1 \leq \epsilon' = 2\epsilon$.

Proof: The function f defines a partition of $\mathcal{X} = \dot{\bigcup}_i \mathcal{X}_i$ into the pre-images $\mathcal{X}_i = f^{-1}(i)$. Define

$$S(x|i) = \begin{cases} \frac{1}{P_X(\mathcal{X}_i)} P_X(x) & \text{if } x \in \mathcal{X}_i, \\ 0 & \text{otherwise.} \end{cases}$$

It can easily be checked that this leads to the desired results by calculating $P_{\widehat{X}Y}$ and then bounding $\frac{1}{2}\|P_{XY} - P_{\widehat{X}Y}\|_1$ using triangle inequality and twice the assumption in Eq. 2. It actually furthermore has the property that it gives a Markov chain $Y - X - X' - \widehat{X} - X'$, since $X' = f(X) = f(\widehat{X})$ by construction. ■

The idea, in any case, is that f identifies different x that have almost the same correlation with Y, as expressed by the conditional distribution; S recovers X as far as its correlation with Y is concerned. Thus, X' is an approximately sufficient statistics for X.

Now the question is: How large does N need to be? Surely not larger than $|\mathcal{X}|$, but we assume that to be potentially unbounded. In the worst case, we need to choose an ϵ-net of the probability simplex $\mathcal{P}(\mathcal{Y})$ of all probability distributions on $\mathcal{Y}$ with respect to the total variational distance, which results in $N \leq \left(\frac{c}{\epsilon}\right)^{|\mathcal{Y}|}$, with some universal constant $c > 0$.

To be more precise, N is the minimum cardinality of an ϵ-net, and a standard upper bound is given by the corresponding covering number. A standard estimate for the covering number of $K \subset \mathbb{R}^n$ by a convex and symmetric $D \subset \mathbb{R}^n$ can be attained via the volume of the involved sets

$$\frac{vol(K)}{vol(D)} \leq N \leq \frac{vol(K + \frac{1}{2}D)}{vol(\frac{1}{2}D)}. \tag{3}$$

A probability simplex $\Delta_{\mathcal{Y}}$ can be understood as a subset of the unit sphere in $\mathbb{R}^{|\mathcal{Y}|}$ which again is the boundary of the corresponding unit ball. Using the above volume bound for ϵ-nets on the unit ball by ϵ-balls, we get that

$$N_\Delta \leq \left(\frac{2}{\epsilon} + 1\right)^{|\mathcal{Y}|} \leq \left(\frac{3}{\epsilon}\right)^{|\mathcal{Y}|}. \tag{4}$$

This is in fact independent of the metric used, as long as the unit ball and the ϵ-ball are defined via the same metric. However, using somewhat more involved ideas it is possible to determine a bound on the covering number of the unit sphere more directly (see [13, Lemma 5.3]), leading to

$$N_\Delta \leq \left(\frac{2}{\epsilon}\right)^{|\mathcal{Y}|}. \tag{5}$$

We will however get back to using the simpler estimates when discussing the classical-quantum case.

III. Application to approximating the information bottleneck function.

To apply the results from the last section to find the desired approximation results, we first prove another lemma, exploiting the Markov chain emerging from Lemma 1.

Lemma 2 *Let $Y - X - \widetilde{X}$ be a Markov chain. Then the IB function of P_{XY} dominates the IB function of $P_{\widetilde{X}Y}$ pointwise:*

$$I_{XY}(R) \geq I_{\widetilde{X}Y}(R) \quad \forall R.$$

Proof: This follows easily from the definition: recall that $I_{\widetilde{X}Y}(R) = \max I(Y : W)$ with Markov chain

$$Y - X - \widetilde{X} - W, \, I(\widetilde{X} : W) \leq R.$$

Hence, given any W eligible for $\widetilde{X}Y$, by data processing inequality we have $I(X : W) \leq R$, and so W is also eligible for XY. Since we are optimising the same objective function, the claim follows. ■

The lemma implies that for the Markov chain $Y - X - X' - \widehat{X} - X'$ resulting from Lemma 1, we have

$$I_{XY}(R) \geq I_{X'Y}(R) \geq I_{\widehat{X}Y}(R) \geq I_{X'Y}(R),$$

which implies

$$I_{XY}(R) \geq I_{X'Y}(R) = I_{\widehat{X}Y}(R). \tag{6}$$

Now we are ready for our main result, which bounds the gap in the latter inequality:

Corollary 3 *Under the assumptions of Lemma 1,*

$$I_{X'Y}(R) \leq I_{XY}(R) \leq I_{X'Y}(R) + \delta(\epsilon),$$

where $\delta(\epsilon,|\mathcal{Y}|) := \epsilon' \log|\mathcal{Y}| + (1+\epsilon')h\Big(\frac{\epsilon'}{1+\epsilon'}\Big)$.

Proof: The left hand side inequality has been shown in Lemma 2, concretely the inequality in Eq. (6). For the right hand side bound, we use the equality in Eq. (6), and consider a channel $P_{W|X} : \mathcal{X} \longrightarrow \mathcal{W}$. Define Markov chains $Y-X-W$ and $Y-\widehat{X}-\widehat{W}$ by letting

$$P_{YXW} = P_{XY}P_{W|X}, \quad P_{Y\widehat{X}\widehat{W}} = P_{\widehat{X}Y}P_{\widehat{W}|\widehat{X}},$$

where we set $P_{\widehat{W}|\widehat{X}} = P_{W|X}$. Then we have $I(X:W) = I(\widehat{X}:\widehat{W})$, since $P_X = P_{\widehat{X}}$ and hence $P_{WX} = P_{\widehat{W}\widehat{X}}$. That is, W is eligible for XY if and only if $\widehat{W}$ is eligible for $\widehat{X}Y$.

On the other hand,

$$\big|I(Y:W)-I(Y:\widehat{W})\big| = \big|H(Y|W)-H(Y|\widehat{W})\big| \leq \delta(\epsilon,|\mathcal{Y}|),$$

where we have used the Alicki-Fannes continuity bound for the conditional entropy in the form of [14]. ■

For Lemma 1 to be applicable, we can use an ϵ-net for the probability distributions on $\mathcal{Y}$, resulting in $N \leq \left(\frac{2}{\epsilon}\right)^{|\mathcal{Y}|}$. This means of course, that to approximate I_{XY} we might as well compute $I_{X'Y}$, and the latter can attain its optimal value with an alphabet size $|\mathcal{W}| \leq |\mathcal{X}'|+1 = N+1 \leq \left(\frac{2}{\epsilon}\right)^{|\mathcal{Y}|}+1$.

In practice, utilizing Corollary 3 will require some simple preprocessing on the used probability distribution associated with the considered dataset in order to calculate the simplified bottleneck function. In some cases, one might want to directly use the original distribution, but still restrict $|\mathcal{W}|$. This should in particular be useful when $\mathcal{X}$ is very large. In the following we denote by $I_{XY}(R,N)$ the bottleneck function $I_{XY}(R)$ with the additional constraint that $|\mathcal{W}| \leq N$.

Corollary 4 *Under the assumptions of Lemma 1,*

$$I_{XY}(R,N) \leq I_{XY}(R) \leq I_{XY}(R,N) + \delta(\epsilon,|\mathcal{Y}|),$$

where $\delta(\epsilon,|\mathcal{Y}|) := \epsilon' \log|\mathcal{Y}| + (1+\epsilon')h\Big(\frac{\epsilon'}{1+\epsilon'}\Big)$ *and* $N \leq \left(\frac{2}{\epsilon}\right)^{|\mathcal{Y}|}$.

Proof: Using the definitions in Lemma 1 and applying it to the joint distribution P_{YW} resulting from the optimal channel $P_{W|X}$, we get the following Markov chain, $Y-X-W-W'-\widehat{W}$. Here, W' is constructed via an ϵ-net on Y and therefore requires at most $N \leq \left(\frac{2}{\epsilon}\right)^{|\mathcal{Y}|}$. Furthermore, we have $\frac{1}{2}\|P_{YW}-P_{Y\widehat{W}}\|_1 \leq 2\epsilon$. It is immediately clear that restricting the maximization to a smaller cardinality can not increase the result and therefore the left hand side follows directly.

Given that $I(Y:W)$ is the optimal value for $I_{XY}(R)$, we can use data processing and a continuity bound to get

$$I(Y:W) \geq I(Y:W') \geq I(Y:\widehat{W}) \geq I(Y:W)-\delta(\epsilon,|\mathcal{Y}|). \quad (7)$$

From data processing we also immediately get that

$$I(X:W) \geq I(X:W'), \quad (8)$$

which means that for a given rate R, if $P_{W|X}$ is an eligible map, then also $P_{W'|X}$ is. Therefore, we observe that the restricted bottleneck function deviates from the unrestricted one by at most $\delta(\epsilon)$:

$$I_{XY}(R)-\delta(\epsilon,|\mathcal{Y}|) \leq I(Y:W') \leq I_{XY}(R,N). \qquad ■$$

While this corollary gives a slightly more direct approach to the problem, it has the downside that, a priori, we do not know anything about the possible distributions $P_{Y|w=w}$. Therefore, in general an ϵ-net on $\mathcal{Y}$ is needed.

A. *Examples*

For many applications, such as machine learning, the involved probability distribution P_{XY} often stems from a dataset of objects with assigned labels. These datasets are then used as training sets e.g. towards a classification task. Generally, if we have a multi-class classification problem every object is assigned exactly one label. The number of labels available determines $|\mathcal{Y}|$. Now, it might be the case that some objects appear multiple times in the dataset but with different labels; let us call such objects *ambiguous*. However, in the framework of Corollary 3, as long as the total number of such objects in the dataset is not too large, only a very restricted number of distributions $P_{Y|X=x}$ can appear and we can even achieve $\epsilon = 0$.

A particular case that is often encountered is that of deterministic datasets, i.e. $Y = f(X)$ for some single-valued function f, meaning that there are no ambiguous objects. This includes many commonly used datasets such as MNIST and CIFAR. It is known that in these cases the information bottleneck function takes a very simple form [15]. This is also the case when considering the conditions for our Lemma 1 since the possible probability distributions are limited to $P_{Y|X=x} = \delta_{y,f(x)}$, and therefore $N = |\mathcal{Y}|$ is sufficient to even achieve $\epsilon = 0$.

For a small amount of ambiguous objects, let's say a of them, one can still achieve $\epsilon = 0$ with $N = |\mathcal{Y}|+a$. However, for large datasets, a might quickly become unpractical and the approximating method via ϵ-nets becomes much more viable.

One could also consider a different class of datasets, namely those for multi-label datasets (every object can be assigned multiple labels). Let's assume that the dataset does not include any ambiguous objects. Now, with a little combinatorics one sees that we can achieve $\epsilon = 0$ with $N = 2^{|\mathcal{Y}|}$ if the number of labels per objects is only restricted by the total number of labels $|\mathcal{Y}|$. If every object is assigned exactly k labels we can achieve the same with $N = \binom{|\mathcal{Y}|}{k}$ and with a maximum of k labels per object, $N = \sum_{i=0}^{k} \binom{|\mathcal{Y}|}{i}$ suffices. While these quantities are not necessarily small, they are still significantly better than the ϵ-net approximation method.

IV. Classical-quantum setting

In quantum information theory the role of probability distributions is taken by quantum states, i.e. positive semi-definite hermitian matrices with trace one. Instead of maps between probabilities, these states are transformed by completely positive and trace preserving maps called quantum channels. For a complete introduction we refer the reader to the literature, e.g. [16].

In the traditional (classical) setting the main tool to prove alphabet size bounds is Caratheodory's theorem. One crucial property that allows us to apply Caratheodory's theorem is that every conditional entropy for a classical probability distribution p_{XY} can be written as

$$H(X|W) = \sum_{w} p(w) H(X|W = w). \tag{9}$$

This might seem like a trivial observation, however it should be noted that when we move from classical probabilities to quantum states, in general such an equality does not exist, leading to major complications in quantum information theory [17], [18]. In the following we will discuss an intermediate setting where the random variables X, Y are quantum and only W is restricted to be classical. In this setting, the generalization of Eq. (9) still holds and we will be able to apply Caratheodory's theorem as well as the new framework developed in the first part of this work.

Let us briefly recall the definition of the quantum information bottleneck. In [10] it was defined as the natural generalization of the classical case as

$$I^{q}_{XY}(R) := \sup_{\substack{\mathcal{N}^{X\to W} \\ I(X';W)_{\tilde{\tau}} \leq R}} I(Y;W)_{\sigma}, \tag{10}$$

with

$$\tilde{\tau}_{X'W} := (\mathrm{id}_{X'} \otimes \mathcal{N}^{X\to W})\tau_{X'X}, \tag{11}$$

where $\tau_{X'X}$ is a purification of ρ_X, and

$$\sigma_{WY} := (\mathcal{N}^{X\to W} \otimes \mathrm{id}_Y)\rho_{XY}. \tag{12}$$

Hence $\rho_X = \mathrm{Tr}_{X'}\, \tau_{X'X} = \mathrm{Tr}_Y\, \rho_{XY}$. This formula unfortunately proved difficult to handle analytically. It was however shown in [12] that the following formulation is equivalent to the one above:

$$I^{q}_{XY}(R) = \sup_{\substack{\mathcal{N}^{X\to W} \\ I(YR;W)_{\sigma} \leq R}} I(Y;W)_{\sigma}, \tag{13}$$

where $\sigma_{WYR} := (\mathcal{N}^{X\to W} \otimes \mathrm{id}_{YR})\psi_{XYR}$ and ψ_{XYR} a purification of ρ_{XY}.

Defining the quantum information bottleneck via purifications is in general necessary, due to the well known fact that quantum information cannot be copied. In the intermediate case where only Y is quantum and X, W are classical, this reduces to a definition that more closely resembles the fully classical case:

$$I^{cq}_{XY}(R) := \sup_{\substack{\mathcal{N}^{X\to W} \\ I(X;W)_{\sigma} \leq R}} I(Y;W)_{\sigma}, \tag{14}$$

where purifying the state ρ_{XY} becomes equivalent to copying the classical X.

First, we will argue that the classical bound proven by using Caratheodory's theorem can be generalized to the classical-quantum setting where X and Y are quantum, and W is classical, leading to the following lemma.

Lemma 5 *For X and Y quantum, and W classical, an optimal solution for the quantum information bottleneck can be achieved with $|\mathcal{W}| \leq |\mathcal{Y}|^2|\mathcal{R}|^2 + 1$.*

Proof: The proof idea in the classical case comes originally from [8], see also [9] for an application to generalized classical information bottleneck functions. Starting from a state ψ_{XYR} we consider the setting where the channel $\mathcal{N}$ has a quantum input but a classical output. The resulting state σ_{WYR} will then have the following classical-quantum form

$$\sigma_{WYR} = \sum_{w} p(w)|w\rangle\langle w| \otimes \rho^{w}_{YR}, \tag{15}$$

with $\{|w\rangle\}_w$ some orthonormal basis on W. Now, optimizing over the channel is equivalent to optimizing over all $\{p(w), |w\rangle\}$. The feasible solutions for the quantum information bottleneck in this setting are now completely determined by the set

$$\begin{aligned}
&\mathcal{C}(\Psi_{XYR}) \\
&= \left\{ \big(\rho_{YR}, I(YR\!:\!W), I(Y\!:\!W)\big) : \sum_{w} p(w)\rho^{w}_{YR} = \rho_{YR} \right\} \\
&= \Bigg\{ \Bigg(\rho_{YR}, \sum_{w} p(w)[H(YR) - H(YR|W=w)], \\
&\qquad \sum_{w} p(w)[H(Y) - H(Y|W=w)] \Bigg) : \sum_{w} p(w)\rho^{w}_{YR} = \rho_{YR} \Bigg\}.
\end{aligned} \tag{16}$$

Following the argument in [8], $\mathcal{C}(\Psi_{XYR})$ is a convex and compact set. Also, the state ρ_{YR} can be interpreted as living on a $|\mathcal{Y}|^2|\mathcal{R}|^2$ real vector space, therefore a direct application of Caratheodory's theorem proves the above lemma. ■

Note that ψ can always be chosen to be a minimal purification of ρ with $|\mathcal{R}| \leq |\mathcal{X}||\mathcal{Y}|$, and therefore we have $|\mathcal{W}| \leq |\mathcal{Y}|^4|\mathcal{X}|^2 + 1$.

One can also consider a setting in between the fully classical case and the one considered before, that leads to a bound closer to what is familiar in the classical case.

Lemma 6 *For Y quantum, but X and W classical, an optimal solution for the quantum information bottleneck can be achieved with $|\mathcal{W}| \leq |\mathcal{X}| + 1$.*

The proof is almost identical to that of the fully classical case and omitted here for brevity.

In this second intermediate scenario we can also find a classical-quantum version of the cardinality bounds achieved with the recoverability strategy. Therefore, we first need a generalization of Lemma 1.

Lemma 7 *Given a classical-quantum state*

$$\rho_{XY} = \sum_x p(x)|x\rangle\langle x| \otimes \rho_Y^x, \tag{17}$$

and assume that there exist N quantum states $\sigma_Y^1, \dots, \sigma_Y^N$ and a function $f : \mathcal{X} \longrightarrow [N]$ with the property that

$$\forall x \quad \frac{1}{2}\|\rho_Y^x - \sigma_Y^{f(x)}\|_1 \leq \epsilon, \tag{18}$$

for given $\epsilon > 0$. Then there exists a recovery channel $S : [N] \longrightarrow \mathcal{X}$ such that the Markov chain $Y - X - X' - \widehat{X}$ defined by $X' = f(X)$ and $P_{\widehat{X}|X'} = S$ satisfies $P_X = P_{\widehat{X}}$ and $\frac{1}{2}\|\rho_{XY} - \rho_{\widehat{X}Y}\|_1 \leq \epsilon' = 2\epsilon$.

Proof: The proof is similar to the completely classical case. Again, the function f defines a partition of $\mathcal{X} = \dot{\bigcup}_i \mathcal{X}_i$ into the pre-images $\mathcal{X}_i = f^{-1}(i)$. Now, it can be shown that the function

$$S(x|i) = \begin{cases} \frac{1}{P_X(\mathcal{X}_i)} P_X(x) & \text{if } x \in \mathcal{X}_i, \\ 0 & \text{otherwise,} \end{cases}$$

achieves the desired result. The above strategy again has the property that it gives a Markov chain $Y - X - X' - \widehat{X} - X'$, since $X' = f(X) = f(\widehat{X})$ by construction. ■

Again, we shall ask what is the necessary N for the assumptions to hold in general. Following the classical argument it is always possible to satisfy the assumptions by choosing an ϵ-net, but now on the set of quantum states. Pure quantum states correspond to the unit sphere in $\mathbb{C}^{|\mathcal{Y}|} \cong \mathbb{R}^{2|\mathcal{Y}|}$. Mixed quantum states can be represented by pure states on a purifying system $\mathcal{Y} \otimes \mathcal{Y}'$, whose Hilbert space dimension is $|\mathcal{Y}|^2$. We can now use the previously presented volume bound to get

$$N_Q \leq \left(\frac{2}{\epsilon} + 1\right)^{2|\mathcal{Y}|^2} \leq \left(\frac{3}{\epsilon}\right)^{2|\mathcal{Y}|^2}. \tag{19}$$

Using the result in Lemma 7, we immediately get the following generalizations.

Lemma 8 *Let $Y - X - \widetilde{X}$ be a Markov chain. Then the IB function of ρ_{XY} dominates the IB function of $\rho_{\widetilde{X}Y}$ pointwise:*

$$I_{XY}^{cq}(R) \geq I_{\widetilde{X}Y}^{cq}(R) \quad \forall R.$$

Corollary 9 *Under the assumptions of Lemma 7,*

$$I_{X'Y}^{cq}(R) \leq I_{XY}^{cq}(R) \leq I_{X'Y}^{cq}(R) + \delta(\epsilon, |\mathcal{Y}|),$$

where $\delta(\epsilon, |\mathcal{Y}|) := \epsilon' \log |\mathcal{Y}| + (1+\epsilon')h\left(\frac{\epsilon'}{1+\epsilon'}\right)$.

Corollary 10 *Under the assumptions of Lemma 7,*

$$I_{XY}^{cq}(R, N) \leq I_{XY}^{cq}(R) \leq I_{XY}^{cq}(R, N) + \delta(\epsilon, |\mathcal{Y}|),$$

where $\delta(\epsilon, |\mathcal{Y}|)$ is as before and $N \leq \left(\frac{3}{\epsilon}\right)^{2|\mathcal{Y}|^2}$.

The proofs of the last three results are very similar to those in the fully classical case and therefore omitted here for the sake of brevity.

V. Conclusion

In the present paper we addressed the question whether the information bottleneck function can be well-approximated with an auxiliary random variable whose alphabet size $|\mathcal{W}|$ only depends on $|\mathcal{Y}|$ and the allowed error. We proved that this is indeed the case, using an approach based on approximate sufficiency, which itself relates to approximate Markov chains through (approximate) recovery maps. We showed that for many practical problems this approach drastically reduces the cardinality needed and therefore the computational complexity of evaluating the bottleneck function. In the second part, we discussed extensions of these and other known bounds to settings with quantum information and showed that in some special cases dimension bounds exist and make it possible to evaluate the quantum information bottleneck.

Furthermore our results, in particular Lemma 1, could provide a possible road towards classifying problems which are suitable to be solved by neural networks. The general approach of this work could also lead to new cardinality bounds for other information theoretic quantities.

An important question left open is to find dimension bounds for the fully quantum case, a setting where the usual tools like Caratheodory's theorem do not seem to be applicable. We discuss the application of our new recoverability approach to the fully quantum setting in [19].

Acknowledgments. The authors thank Axel Foley for discussions on information recovery. CH was supported by the VILLUM FONDEN via the QMATH Centre of Excellence (Grant no. 10059). AW was supported by the Spanish MINECO, projects FIS2016-86681-P, with the support of FEDER funds; and by the Generalitat de Catalunya, CIRIT project 2014-SGR-966.

References

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," arXiv:physics/0004057, 2000.

[2] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. 2015 IEEE Information Theory Workshop (ITW)*, IEEE, 2015, pp. 1–5.

[3] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," arXiv[cs.IT]:1703.00810, 2017.

[4] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *Proc. 14th ACM International Conference on Multimedia*, ACM, 2006, pp. 35–44.

[5] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2000, pp. 208–215.

[6] M. Stark, A. Shah, and G. Bauch, "Polar code construction using the information bottleneck method," in *Proc. 2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, IEEE, Apr 2018, pp. 7–12.

[7] R. T. Rockafeller, *Convex Analysis*, Princeton Mathematics Series, vol. 28, Princeton University Press, 1970.

[8] H. S. Witsenhausen and A. D. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 5, pp. 493–501, Sept 1975.

[9] H. Hsu, S. Asoodeh, S. Salamatian, and F. P. Calmon, "Generalizing bottleneck problems," in *Proc. 2018 IEEE International Symposium on Information Theory (ISIT)*, IEEE, July 2018, pp. 531–535.

[10] A. L. Grimsmo and S. Still, "Quantum predictive filtering," *Physical Review A*, vol. 94, 012338, Jul 2016.

[11] S. Salek, D. Cadamuro, P. Kammerlander, and K. Wiesner, "Quantum Rate-Distortion Coding of Relevant Information," *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 2603–2613, Apr 2019; arXiv[quant-ph]:1704.02903.

[12] N. Datta, C. Hirche, and A. Winter, "Convexity and Operational Interpretation of the Quantum Information Bottleneck Function," in *Proc. 2019 IEEE International Symposium on Information Theory (ISIT)*, IEEE, July 2019, pp. 1157–1161.

[13] G. Aubrun and S. J. Szarek, *Alice and Bob Meet Banach: The Interface of Asymptotic Geometric Analysis and Quantum Information Theory*, Mathematical Surveys and Monographs, vol. 223, American Mathematical Society, 2017.

[14] A. Winter, "Tight Uniform Continuity Bounds for Quantum Entropies: Conditional Entropy, Relative Entropy Distance and Energy Constraints," *Communications in Mathematical Physics*, vol. 347, pp. 291–313, Oct 2016.

[15] A. Kolchinsky, B. D. Tracey, and S. Van Kuyk, "Caveats for information bottleneck in deterministic scenarios," in *Proc. 7th International Conference on Learning Representations (ICLR) 2019*, 2019; arXiv[stat.ML]:1808.07593v4. Discussion at this URL: https://openreview.net/forum?id=rke4HiAcY7

[16] M. M. Wilde, *Quantum Information Theory*. Cambridge University Press, Cambridge, UK, 2013.

[17] S. Beigi and A. A. Gohari, "On dimension bounds for auxiliary quantum systems," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 368–387, Jan 2013.

[18] C. Hirche and D. Reeb, "Bounds on Information Combining With Quantum Side Information," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 4739–4757, July 2018.

[19] M. Christandl, C. Hirche, and A. Winter, "Dimension size bounds in quantum information theory from recoverability," in preparation, 2020.