# Efficient algorithms for quantum information bottleneck

Masahito Hayashi[1,2,3,4] and Yuxiang Yang[5]

[1]Shenzhen Institute for Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen,518055, China
[2]International Quantum Academy (SIQA), Futian District, Shenzhen 518048, China
[3]Guangdong Provincial Key Laboratory of Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, China
[4]Graduate School of Mathematics, Nagoya University, Nagoya, 464-8602, Japan
[5]QICI Quantum Information and Computation Initiative, Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong

**The ability to extract relevant information is critical to learning. An ingenious approach as such is the information bottleneck, an optimisation problem whose solution corresponds to a faithful and memory-efficient representation of relevant information from a large system. The advent of the age of quantum computing calls for efficient methods that work on information regarding quantum systems. Here we address this by proposing a new and general algorithm for the quantum generalisation of information bottleneck. Our algorithm excels in the speed and the definiteness of convergence compared with prior results. It also works for a much broader range of problems, including the quantum extension of deterministic information bottleneck, an important variant of the original information bottleneck problem. Notably, we discover that a quantum system can achieve strictly better performance than a classical system of the same size regarding quantum information bottleneck, providing new vision on justifying the advantage of quantum machine learning.**

## 1 Introduction

Learning is a task of eminent importance to the contemporary world. As such, it has always been of top priority to quest powerful tools for learning information. Information bottleneck [32] stands as an excellent example, with many useful applications including deep learning [8, 28, 33], video processing [16], clustering [29] and polar coding [30]. Concretely, information bottleneck is a method to extract a piece of information $T$ with respect to the system $Y$ from the system $X$, and is formulated as the minimization problem of the difference $I(T:X) - \beta I(T:Y)$ with a positive parameter $\beta$, where $I(T:X)$ is the mutual information between $T$ and $X$. In particular, we are interested in the case when $X$ is classical. By design, information bottleneck achieves an irreversible compression, by extracting essential information about $Y$ and simultaneously removing unessential information contained in $X$.

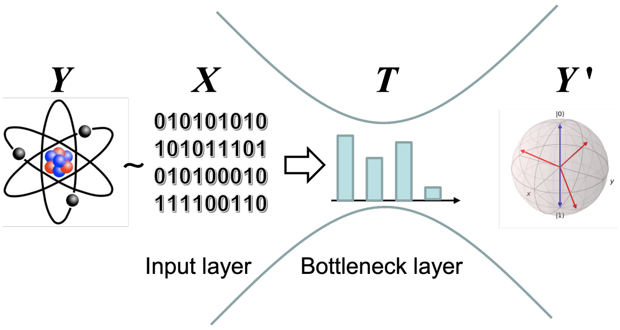

Figure 1: **Visualization of quantum information bottleneck.** In a prototypical setting of quantum information bottleneck, the task is to compress a classical system into a smaller system $T$, which can be either classical or quantum, by extracting its useful information about a quantum system $Y$ and removing the useless information. It is expected that more relevant information $Y'$ about $Y$, instead of the entire $X$, can be recovered from $T$.

As we are stepping into the age of quantum information, the demand is growing for a method that efficiently learns information on a quantum system. For this purpose, let us consider the setup of quantum information bottleneck (QIB), demonstrated in Fig. 1. Similar as its classical counterpart, the aim of QIB is to compress $X$ into a smaller system $T$ while preserving the correlation with $Y$ when some of these systems are quantum systems. Prior to this work, QIB has been discussed in several recent works [2, 6, 9, 14, 24] and has been applied to quantum information theory [6, 14] and quantum machine learning [2]. On the other hand, the fundamental properties of QIB such as convergence have not been analysed, which hinders its application in more practical tasks. Quantum information bottleneck is first proposed as a quantum extension of information bottleneck method in [9]. It also derived a necessary condition for the solution of the minimization problem (see [9, Appendix A]) by

Masahito Hayashi: hayashi@sustech.edu.cn
Yuxiang Yang: yuxiang@cs.hku.hk

using Lagrange multiplier method in the same way as [1, 4]. Using the obtained condition, it also proposed an iterative algorithm to find a solution to satisfy the necessary condition [9, Appendix C]. Then, the reference [24] considered QIB in the quantum communication scenario. [1] However, no study discusses the behaviour of the iterative algorithm, i.e., it is not known whether the algorithm monotonically reduces the objective function [9, 24, 31, 32]. It was also claimed in [9, Appendix B] that there is no advantage of using a quantum $T$ if $X, Y$ are both classical.

In this work, we conduct a systematic study on quantum information bottleneck, focusing on the case when the system $X$ is classical. Compared to existing works [2, 6, 9, 14, 24], our work makes significant contributions in several directions:

First, we provide throughout analyses on two critical properties – efficiency and convergence – of QIB. Motivated by a recent generalization [22] of the Arimoto-Blahut algorithm [1, 4], we introduce a new quantum information bottleneck algorithm with an acceleration parameter $\gamma$ that can make the value of QIB converges much faster than before when chosen properly. We prove rigorous criteria for our algorithm to converge and to achieve a minimum. In particular, we prove that the choice of $\beta$ plays an important role in convergence.

Second, in contrast to the claim in Refs. [9, 24], we provide concrete examples where using a quantum instead of classical $T$ could reduce the minimal value of QIB. Notably, our result justifies a genuine quantum advantage in quantum machine learning [3, 27, 34], where the employment of quantum circuits has been prevalent [5, 11, 17, 20, 25, 26] but the quantum advantage was rarely justified.

Last but not least, we generalise QIB by considering a general target function $(1-\alpha)H(T) + \alpha I(T:X) - \beta I(T:Y)$ with parameters $\alpha, \beta \geq 0$, which reduces to the standard QIB when $\alpha = 1$. By doing so, the generalised QIB contains QDIB, i.e., the quantum version of deterministic information bottleneck [31], by setting $\alpha = 0$. We show that our analyses and our algorithm hold for this generalised setting and, in particular, to QDIB. Then, we clarify that QDIB can be used to find a good approximate sufficient statistics $T$ for $X$ for $Y$, which requires a smaller entropy $H(T)$ and larger mutual information $I(T:Y)$. We justify our finding via a numerical example, where QDIB extracts a good approximate sufficient statistics over information about a quantum ensemble.

In summary, our work addresses several critical issues of QIB, including convergence, efficiency, choice of parameters, and the quantum advantage. We also extend QIB to a generalised setting and introduce the notion of QDIB. Our results consist of both rigorous analytical analyses and numerical experiments that justifies the importance of QIB and QDIB in fundamental tasks of learning.

The remaining part of this paper is organized as follows. Section 2 introduces our algorithm for quantum information bottleneck, and discusses its convergence and dependence of the parameter $\beta$. Section 3 discusses our algorithm when our memory system $T$ is classical. Section 4 presents examples that realizes a smaller value of the target function by quantum memory $T$ than by classical memory $T$. Section 5 discusses an application of our QIB algorithm in data classification. Section 6 proposes our algorithm for quantum deterministic information bottleneck, and studies its properties. Section 7 applies it to the extraction of approximate sufficient statistics, and numerically verifies its efficiency in an example. Section 8 makes discussion and conclusion.

## 2 The quantum information bottleneck (QIB) problem

### 2.1 Problem definition

Consider a classical-quantum joint system composed of $X$ and $Y$ with the joint state

$$\rho_{XY} := \sum_x P_X(x)|x\rangle\langle x| \otimes \rho_{Y|x}, \tag{1}$$

where $X$ is a classical system and $Y$ is a quantum system. Our quantum information bottleneck (QIB) problem aims at constructing an information processor, modelled by a c-q channel $\sigma_{T|X}$ from $X$ to $T$ (which prepares a quantum state $\sigma_{T|x}$ when the classical register is $x$), that extracts efficient information from $X$ with respect to the quantum system $Y$. After the action of the information processor, the joint state becomes:

$$\rho_{XYT} := \sum_x P_X(x)|x\rangle\langle x| \otimes \rho_{Y|x} \otimes \sigma_{T|x}. \tag{2}$$

To this aim, the QIB problem concerns constructing a classical-quantum channel $\sigma_{T|X} : X \to T$ that minimizes the information bottleneck function, consisting of entropic quantities defined with respect to the joint state $\rho_{XYT}$:

$$\begin{aligned} f_\alpha(\sigma_{T|X}) &:= H(T) - \alpha H(T|X) - \beta I(T:Y) \\ &= (1-\alpha)H(T) + \alpha I(T:X) - \beta I(T:Y), \end{aligned} \tag{3}$$

where $H(T)$ denotes the entropy of $T$ [2], $H(T|X)$ denotes the conditional entropy of $T$ on $X$, and $I(T:Y)$

[1] The reference [24, Appendix A] derived a necessary condition for the solution of the minimization problem by using Lagrange multiplier method in the same way as [1, 4]. Using the obtained condition, it also proposed an iterative algorithm to find a solution to satisfy the necessary condition [24, The end of Appendix C].

[2] For convenience, the notation $H(A)$ stands for the Shannon entropy when the system $A$ is classical and for the von Neumann entropy when $A$ is quantum.

stands for the mutual information between $T$ and $Y$.

That is, our aim is the calculation of the following value:

$$\mathcal{I}_{\alpha,\beta} := \min_{\sigma_{T|X}} f_\alpha(\sigma_{T|X}). \tag{4}$$

In the information bottleneck (3), $\alpha$ and $\beta$ are positive real variables modelling the objective of the task. In the original proposal of information bottleneck [32] $\alpha = 1$. Another common choice of $\alpha$ is $\alpha = 0$, and the task is called a deterministic QIB (whose classical counterpart was discussed in Ref. [31]). The parameter $\beta$ controls the tradeoff between faithfulness and compression. For instance, in a deterministic information bottleneck, a larger $\beta$ would make $I(T : Y)$ more prominent in the objective function, forcing the information processor to preserve more information about $Y$, whereas a smaller $\beta$ would signify the role of $I(T : X)$, prompting the information processor to do more compression in $X$.

Although this section addresses the case with quantum systems $Y$ and $T$, the case with a classical system $Y$ and a quantum $T$ can be contained as a special case by considering the diagonal densities $\rho_{Y|x}$. On the other hand, the case with a classical system $T$ is a different problem from the case with a quantum system $T$ because we need to discuss a different minimization problem, which has a different range for the minimizing variable. Fortunately, our algorithm for a quantum system $T$, presented in the next subsection, can be applied to the case with a classical system $T$. Section 3 discusses the case of $T$ being classical. We remark that the case where both $T$ and $Y$ are classical has been widely studied in classical information theory and machine learning; see, e.g., Refs. [28, 31–33].

## 2.2 QIB algorithm for $\alpha = 1$

The paper [9] discussed this problem when $X, Y, T$ are quantum systems and $\alpha = 1$, extending the classical information bottleneck [32] to the quantum regime. It derived a necessary condition for $\sigma_{X|T}$ to achieve the minimum (4). The necessary condition with quantum systems $T, Y$ and a classical system $X$ is written as

$$\begin{aligned}\log \sigma_{T|x} =& (1-\beta)\log \sigma_T[\sigma_{T|X}] \\ &- \beta \operatorname{Tr}_Y \rho_{Y|x}\Big(\log \rho_Y - \log \sigma_{YT}[\sigma_{T|X}]\Big) - C_x,\end{aligned} \tag{5}$$

where $C_x$ is a normalizing constant and

$$\rho_Y := \sum_x P_X(x)\rho_{Y|x} \tag{6}$$

$$\sigma_T[\sigma_{T|X}] := \sum_x P_X(x)\sigma_{T|x} \tag{7}$$

$$\sigma_{YT}[\sigma_{T|X}] := \sum_x P_X(x)\sigma_{T|x} \otimes \rho_{Y|x}. \tag{8}$$

Since this condition is self-consistent, using this condition, the paper [9] proposed the following iterative algorithm with the following update rule:

$$\begin{aligned}\sigma_{T|x}^{(n+1)} :=& \frac{1}{e^{C_x}} \exp\Big((1-\beta)\log \sigma_T[\sigma_{T|X}^{(n)}] \\ &- \beta \operatorname{Tr}_Y \rho_{Y|x}\Big(\log \rho_Y - \log \sigma_{YT}[\sigma_{T|X}^{(n)}]\Big)\Big).\end{aligned} \tag{9}$$

## 2.3 The acceleration parameter $\gamma$

Next, we propose an extension of the iterative algorithm in [9]. First, we introduce a new parameter $\gamma > 0$ and rewrite the condition (5) as:

$$\begin{aligned}&\log \sigma_{T|x} = (1-\frac{1}{\gamma})\log \sigma_{T|x} + \frac{1}{\gamma}\log \sigma_{T|x} \\ =&(1-\frac{1}{\gamma})\log \sigma_{T|x} + \frac{1}{\gamma}(1-\beta)\log \sigma_T[\sigma_{T|X}] \\ &- \frac{1}{\gamma}\beta \operatorname{Tr}_Y \rho_{Y|x}\Big(\log \rho_Y - \log \sigma_{YT}[\sigma_{T|X}]\Big) - \frac{1}{\gamma}C_x \\ =&\log \sigma_{T|x} - \frac{1}{\gamma}\mathcal{F}_1[\sigma_{T|X}](x) - \frac{1}{\gamma}C_x,\end{aligned} \tag{10}$$

where

$$\begin{aligned}&\mathcal{F}_1[\sigma_{T|X}](x) \\ :=& -\log \sigma_T[\sigma_{T|X}] + \log \sigma_{T|x} \\ &+ \beta \operatorname{Tr}_Y \Big(\rho_{Y|x}\Big(\log(\sigma_T[\sigma_{T|X}] \otimes \rho_Y) - \log \sigma_{YT}[\sigma_{T|X}]\Big)\Big).\end{aligned} \tag{11}$$

Using (10), we can derive another iterative algorithm as

$$\sigma_{T|x}^{(n+1)} := \frac{1}{e^{\frac{1}{\gamma}C_x}} \exp\Big(\log \sigma_{T|x}^{(n)} - \frac{1}{\gamma}\mathcal{F}_1[\sigma_{T|X}^{(n)}](x)\Big). \tag{12}$$

In this way, we can easily generalize the iterative algorithm (9) by [9]. However, it is not trivial to find the suitable value for $\frac{1}{\gamma}$, which, as we show later, is critical to the efficiency of our iterative algorithm. Although many papers [9, 24, 31, 32] discussed the iterative algorithm given by (9) including the classical case, no preceding study showed the convergence of the iterative algorithm by (9). In addition, the discussion above focuses on the case of $\alpha = 1$ and does not include the case of deterministic information bottleneck ($\alpha = 0$). Therefore, to make an efficient algorithm, we need to discuss the choice of the parameter $\gamma$ for generic $\alpha$.

## 2.4 QIB algorithm with general $\alpha$ and convergence

To analyze the convergence of the algorithm (12), we introduce a two-input variable function based on the idea in Ref. [22, Section III-B], whereas the method in Ref. [22, Section III-B] was obtained as a generalization of the Arimoto-Blahut algorithm [1, 4]. The idea is that, instead of directly solving the minimization of $f_\alpha(\sigma_{T|X})$, which is often too difficult, we find a continuous function $J(\sigma_{T|X},\sigma'_{T|X})$ with two variables $\sigma_{T|X},\sigma'_{T|X}$. Then we can update these two input variables $\sigma_{T|X},\sigma'_{T|X}$ alternately to decrease $J(\sigma_{T|X},\sigma'_{T|X})$. Finally, if the function satisfies

$$f_\alpha(\sigma_{T|X}) = J(\sigma_{T|X},\sigma_{T|X}), \tag{13}$$

the minimum of $J(\sigma_{T|X},\sigma'_{T|X})$ will be close to the minimum of the IB function.

The above type of functions can be constructed if we find an operator $\mathcal{F}_\alpha[\sigma_{T|X}](x)$ to satisfy

$$f_\alpha(\sigma_{T|X}) = \sum_x P_X(x)\operatorname{Tr}_T \sigma_{T|x}\mathcal{F}_\alpha[\sigma_{T|X}](x), \tag{14}$$

In this paper, we employ the following function:

$$\begin{aligned}&\mathcal{F}_\alpha[\sigma_{T|X}](x)\\ :=& -\log\sigma_T[\sigma_{T|X}] + \alpha\log\sigma_{T|x}\\ &+\beta\operatorname{Tr}_Y\Big(\rho_{Y|x}\Big(\log(\sigma_T[\sigma_{T|X}]\otimes\rho_Y) - \log\sigma_{YT}[\sigma_{T|X}]\Big)\Big).\end{aligned} \tag{15}$$

Then, the condition (14) is satisfied.

Using this function, we can define $J_0(\sigma_{T|X},\sigma'_{T|X}) := \operatorname{Tr}_T\sum_x \sigma_{T|x}P_X(x)\mathcal{F}_\alpha[\sigma'_{T|X}](x)$, which satisfies the condition (13). However, it is difficult to optimize two input variables alternately in the function $J_0(\sigma_{T|X},\sigma'_{T|X})$. Instead, for $\gamma > 0$, we introduce the following function

$$\begin{aligned}&J_{\gamma,\alpha}(\sigma_{T|X},\sigma'_{T|X}) &(16)\\ :=&\gamma D(\sigma_{T|X}\|\sigma'_{T|X}) + \sum_x P_X(x)\operatorname{Tr}_T\sigma_{T|x}\mathcal{F}_\alpha[\sigma'_{T|X}](x),\end{aligned} \tag{17}$$

where $D(\sigma_{T|X}\|\sigma'_{T|X}) := \sum_x P_X(x)D(\sigma_{T|x}\|\sigma'_{T|x})$ and $D(\sigma_{T|x}\|\sigma'_{T|x})$ denotes the relative entropy.

Next, we need to specify the rules of the alternatively updating $\sigma_{T|X},\sigma'_{T|X}$. Crucially, we need to ensure that $J_{\gamma,\alpha}(\sigma_{T|X},\sigma'_{T|X})$ is non-increasing under the updating rules. To this purpose, we first introduce the following condition:

**(A1)** $\sigma_{T|X}$ and $\sigma'_{T|X}$ satisfy the relation

$$\begin{aligned}&\gamma\sum_x P_X(x)D(\sigma_{T|x}\|\sigma'_{T|x})\\ \ge&\sum_x P_X(x)\operatorname{Tr}_T\sigma_{T|x}(\mathcal{F}_\alpha[\sigma_{T|X}](x) - \mathcal{F}_\alpha[\sigma'_{T|X}](x)).\end{aligned} \tag{18}$$

In fact, the condition (A1) is rewritten as $\gamma \ge \gamma(\sigma_{T|X},\sigma'_{T|X})$ by defining $\gamma(\sigma_{T|X},\sigma'_{T|X})$ as

$$\begin{aligned}&\gamma(\sigma_{T|X},\sigma'_{T|X})\\ :=&\frac{\sum_x P_X(x)\operatorname{Tr}_T\sigma_{T|x}(\mathcal{F}_\alpha[\sigma_{T|X}](x) - \mathcal{F}_\alpha[\sigma'_{T|X}](x))}{\sum_x P_X(x)D(\sigma_{T|x}\|\sigma'_{T|x})}.\end{aligned} \tag{19}$$

This quantity is evaluated as

$$\gamma(\sigma_{T|X},\sigma'_{T|X}) \le \alpha \tag{20}$$

because the relation

$$\begin{aligned}&D(\rho_{YT}[\sigma_{T|X}]\|\rho_{YT}[\sigma'_{T|X}]) \ge D(\sigma_T[\sigma_{T|X}]\|\sigma_T[\sigma'_{T|X}])\\ =&D(\sigma_T[\sigma_{T|X}]\otimes\rho_Y\|\sigma_T[\sigma'_{T|X}]\otimes\rho_Y)\end{aligned} \tag{21}$$

implies the relation

$$\begin{aligned}&\sum_x P_X(x)\operatorname{Tr}_T\sigma_{T|x}(\mathcal{F}_\alpha[\sigma_{T|X}](x) - \mathcal{F}_\alpha[\sigma'_{T|X}](x))\\ =&-D(\sigma_T[\sigma_{T|X}]\|\sigma_T[\sigma'_{T|X}])\\ &+\alpha\sum_x P_X(x)D(\sigma_{T|x}\|\sigma'_{T|x})\\ &+\beta D(\sigma_T[\sigma_{T|X}]\otimes\rho_Y\|\sigma_T[\sigma'_{T|X}]\otimes\rho_Y)\\ &-\beta D(\rho_{YT}[\sigma_{T|X}]\|\rho_{YT}[\sigma'_{T|X}])\\ \le&-D(\sigma_T[\sigma_{T|X}]\|\sigma_T[\sigma'_{T|X}])\\ &+\alpha\sum_x P_X(x)D(\sigma_{T|x}\|\sigma'_{T|x})\\ \le&\alpha\sum_x P_X(x)D(\sigma_{T|x}\|\sigma'_{T|x}).\end{aligned} \tag{22}$$

To state our updating rules, we define

$$\hat\sigma_{\gamma,\alpha,T}[\sigma_{T|X}](x) := \exp\Big(\log\sigma_{T|x} - \frac{1}{\gamma}\mathcal{F}_\alpha[\sigma_{T|X}](x)\Big) \tag{23}$$

$$\hat\eta_{\gamma,\alpha|x}[\sigma_{T|X}] := \operatorname{Tr}\hat\sigma_{\gamma,\alpha,T}[\sigma_{T|X}](x) \tag{24}$$

$$\hat\sigma_{\gamma,\alpha,T|x}[\sigma_{T|X}] := \frac{1}{\hat\eta_{\gamma,\alpha}[\sigma_{T|X}](x)}\hat\sigma_{\gamma,\alpha,T}[\sigma_{T|X}](x). \tag{25}$$

In particular, when $\gamma = \alpha$, the operator $\hat\sigma_{\gamma,\alpha,T}[\sigma_{T|X}](x)$ is simplified as

$$\begin{aligned}&\hat\sigma_{\alpha,T}[\sigma_{T|X}](x)\\ =&\exp\Big(\frac{1-\beta}{\alpha}\log\sigma_T[\sigma_{T|X}]\\ &-\frac{\beta}{\alpha}\operatorname{Tr}_Y\rho_{Y|x}\Big(\log\rho_Y - \log\sigma_{YT}[\sigma_{T|X}]\Big)\Big).\end{aligned} \tag{26}$$

**Theorem 1** *Under the condition (A1), we have*

$$J_{\gamma,\alpha}(\sigma_{T|X},\sigma'_{T|X}) \ge J_{\gamma,\alpha}(\sigma_{T|X},\sigma_{T|X}) \tag{27}$$

$$J_{\gamma,\alpha}(\sigma_{T|X},\sigma'_{T|X}) \ge J_{\gamma,\alpha}(\hat\sigma_{\gamma,\alpha,T|X}[\sigma'_{T|X}],\sigma'_{T|X}). \tag{28}$$

*Proof of Theorem 1:* The condition (A1) yields

$$
\begin{aligned}
&J_{\gamma,\alpha}(\sigma_{T|X},\sigma_{T|X})\\
=&\sum_t \operatorname{Tr}\sigma_{T|x}P_X(x)\mathcal{F}_\alpha[\sigma_{T|X}](x)\\
\le&\sum_x \operatorname{Tr}\sigma_{T|x}P_X(x)\mathcal{F}_\alpha[\sigma'_{T|X}](x,t)\\
&+\gamma\sum_x P_X(x)D(\sigma_{T|x}\|\sigma'_{T|x})\\
=&J_{\gamma,\alpha}(\sigma_{T|X},\sigma'_{T|X}). \qquad (29)
\end{aligned}
$$

Hence, we obtain (27).

Also, we have

$$
\begin{aligned}
&J_{\gamma,\alpha}(\sigma_{T|X},\sigma'_{T|X})\\
\overset{(a)}{=}&\gamma\sum_x P_X(x)\operatorname{Tr}\sigma_{T|x}\Big(\log\sigma_{T|x}-\log\sigma'_{T|x}\\
&+\frac{1}{\gamma}\mathcal{F}_\alpha[\sigma'_{T|X}](x)\Big)\\
\overset{(b)}{=}&\gamma\sum_x P_X(x)\operatorname{Tr}\sigma_{T|x}\Big(\log\sigma_{T|x}-\hat{\sigma}_{\gamma,\alpha,T}[\sigma_{T|X}](x)\Big)\\
\overset{(c)}{=}&\gamma\sum_x P_X(x)\Big(\operatorname{Tr}\sigma_{T|x}\Big(\log\sigma_{T|x}-\log\hat{\sigma}_{\gamma,\alpha,T|x}[\sigma'_{T|X}]\Big)\\
&-\log\hat{\eta}_{\gamma,\alpha}[\sigma'_{T|X}](x)\Big)\\
=&\gamma\sum_x P_X(x)\big(D(\sigma_{T|x}\|\hat{\sigma}_{\gamma,\alpha,T|x}[\sigma'_{T|X}])\big)\\
&-\gamma\sum_x P_X(x)\log\hat{\eta}_{\gamma,\alpha|x}[\sigma'_{T|X}], \qquad (30)
\end{aligned}
$$

where $(a)$, $(b)$, and $(c)$ follow from (17), (23), and (25), respectively. Finally, from Eq. (30) we can see that the minimum of $J_{\gamma,\alpha}(\sigma_{T|X},\sigma'_{T|X})$ is achieved when $\sigma_{T|X}=\hat{\sigma}_{\gamma,\alpha,T|x}[\sigma'_{T|X}]$, since the first term of (30) is non-negative (with equality achieved when $\sigma_{T|X}=\hat{\sigma}_{\gamma,\alpha,T|x}[\sigma'_{T|X}]$) and the second term is independent of $\sigma_{T|X}$. Hence, we obtain (28). ■

**Corollary 2** *Assume that* $\gamma \ge \sup_{\sigma_{T|X},\sigma'_{T|X}}\gamma(\sigma_{T|X},\sigma'_{T|X})$. *When* $\sigma_{T|X}$ *is a local minimizer, we have*

$$\hat{\sigma}_{\gamma,\alpha,T|x}[\sigma_{T|X}]=\sigma_{T|X}, \qquad (31)$$

*which is equivalent to* (5) *when* $\alpha=1$.

When $\gamma \ge \gamma(\hat{\sigma}_{\gamma,\alpha,T|X}[\sigma_{T|X}],\sigma_{T|X})$, the following chain of inequalities hold: $f_\alpha(\sigma_{T|X})=J_{\gamma,\alpha}(\sigma_{T|X},\sigma_{T|X}) \ge J_{\gamma,\alpha}(\hat{\sigma}_{\gamma,\alpha,T|X}[\sigma_{T|X}],\sigma_{T|X}) \ge J_{\gamma,\alpha}(\hat{\sigma}_{\gamma,\alpha,T|X}[\sigma_{T|X}],\hat{\sigma}_{\gamma,\alpha,T|X}[\sigma_{T|X}]) = f_\alpha(\hat{\sigma}_{\gamma,\alpha,T|X}[\sigma_{T|X}])$. Hence, the monotonicity of the information bottleneck under the updating rules is also guaranteed, as long as $\gamma$ is sufficiently large.

Finally, we propose the following algorithm with a fixed $\gamma$ and general $\alpha$:

**Algorithm 1** QIB algorithm

```
Input: A joint state $\rho_{XY}$ [as in Eq. (1)].
Randomly choose an initial c-q channel $\sigma^{(1)}_{T|X}$;
Create a counter $n$ as the number of iterations; initialize $n$ to 1.
repeat
    Choose $\sigma^{(n+1)}_{T|X}$ as $\hat{\sigma}_{\gamma,\alpha,T|X}[\sigma^{(n)}_{T|X}]$ [cf. Eqs. (23) and (25)]; set $n$ as $n+1$.
until convergence.
Output: A c-q channel $\sigma^{(n+1)}_{T|X}$
```

As mentioned, when $\gamma$ satisfies the condition (A1) in all iteration steps, i.e., when $\gamma$ is sufficiently large, Theorem 1 guarantees the monotonicity of the information bottleneck function:

$$f_\alpha(\sigma^{(n+1)}_{T|X}) \le J_{\gamma,\alpha}(\sigma^{(n+1)}_{T|X},\sigma^{(n)}_{T|X}) \le f_\alpha(\sigma^{(n)}_{T|X}). \qquad (32)$$

Since $f_\alpha$ consists of bounded entropic quantities (assuming the system to be finite), it is a bounded quantity. Therefore, the sequence $\{f_\alpha(\sigma^{(n)}_{T|X})\}$ in our Algorithm converges. In addition, we can show that the sequence of c-q channels $\{\sigma^{(n)}_{T|X}\}$ converges as well:

**Theorem 3** *When* $\gamma \ge \sup_{\sigma_{T|X},\sigma'_{T|X}}\gamma(\sigma_{T|X},\sigma'_{T|X})$, *the sequence* $\{\sigma^{(n)}_{T|X}\}$ *converges.*

In particular, since $\alpha \ge \sup_{\sigma_{T|X},\sigma'_{T|X}}\gamma(\sigma_{T|X},\sigma'_{T|X})$, the sequence $\{\sigma^{(n)}_{T|X}\}$ converges with $\gamma=\alpha$.

*Proof:* Since $\{f_\alpha(\sigma^{(n)}_{T|X})\}$ is monotonically decreasing for $n$, we have

$$\lim_{n\to\infty} f_\alpha(\sigma^{(n)}_{T|X})-f_\alpha(\sigma^{(n+1)}_{T|X})=0. \qquad (33)$$

Using (30), we have

$$
\begin{aligned}
&f_\alpha(\sigma^{(n)}_{T|X})=J_{\gamma,\alpha}(\sigma^{(n)}_{T|X},\sigma^{(n)}_{T|X})\\
=&\gamma\sum_x P_X(x)D(\sigma^{(n)}_{T|x}\|\sigma^{(n+1)}_{T|x})+J_{\gamma,\alpha}(\sigma^{(n+1)}_{T|X},\sigma^{(n)}_{T|X})\\
\ge&\gamma\sum_x P_X(x)D(\sigma^{(n)}_{T|x}\|\sigma^{(n+1)}_{T|x})+f_\alpha(\sigma^{(n+1)}_{T|X}). \qquad (34)
\end{aligned}
$$

Thus, we have

$$\gamma\sum_x P_X(x)D(\sigma^{(n)}_{T|x}\|\sigma^{(n+1)}_{T|x}) \le f_\alpha(\sigma^{(n)}_{T|X})-f_\alpha(\sigma^{(n+1)}_{T|X}). \qquad (35)$$

Since due to (33) and (35), the sequence $\{\sigma^{(n)}_{T|X}\}$ is a Cauchy sequence, it converges. ■

We remark that it is free to choose the convergence criterion in Algorithm 1.

In Algorithm 1, $\gamma$ is fixed to be a large enough value. Intuitively (see the next paragraph for more detailed discussion), $\gamma$ (or, more precisely, $1/\gamma$) is an acceleration parameter that makes the algorithm converge faster if chosen to be a smaller value.

To begin with, we show the role of $\gamma$ in convergence of the algorithm. Denote by $\sigma^*_{T|X}$ the convergence point of $\{\sigma^{(n)}_{T|X}\}$. The performance of our algorithm can be characterized by the decreasing speed of the average divergence between $\sigma^*_{T|X}$ and $\sigma^{(n)}_{T|X}$, which is evaluated as

$$
\begin{aligned}
&\sum_x P_X(x) D(\sigma^*_{T|x}\|\sigma^{(n)}_{T|x}) - \sum_x P_X(x) D(\sigma^*_{T|x}\|\sigma^{(n+1)}_{T|x}) \\
=&\sum_x P_X(x) \operatorname{Tr} \sigma^*_{T|x}\Big(\log \sigma^*_{T|x} - \log \sigma^{(n)}_{T|x}\Big) \\
&-\sum_x P_X(x) \operatorname{Tr} \sigma^*_{T|x}\Big(\log \sigma^*_{T|x} - \log \sigma^{(n+1)}_{T|x}\Big) \\
=&\sum_x P_X(x) \operatorname{Tr} \sigma^*_{T|x}\Big(\log \sigma^{(n+1)}_{T|x} - \log \sigma^{(n)}_{T|x}\Big) \\
\overset{(a)}{=}&\sum_x P_X(x) \operatorname{Tr} \sigma^*_{T|x}\Big( -\frac{1}{\gamma}\mathcal{F}_\alpha[\sigma^{(n)}_{T|X}](x) - \log \hat{\eta}_{\gamma,\alpha}[\sigma^{(n)}_{T|X}](x)\Big) \\
\overset{(b)}{=}&\frac{1}{\gamma} J_{\gamma,\alpha}(\sigma^{(n+1)}_{T|X}, \sigma^{(n)}_{T|X}) - \frac{1}{\gamma}\sum_x P_X(x) \operatorname{Tr} \sigma^*_{T|x} \mathcal{F}_\alpha[\sigma^{(n)}_{T|X}](x) \\
\overset{(c)}{=}&\frac{1}{\gamma}\Big( (J_{\gamma,\alpha}(\sigma^{(n+1)}_{T|X}, \sigma^{(n)}_{T|X}) - f_\alpha(\sigma^*_{T|X})) \\
&+\sum_x P_X(x) \operatorname{Tr} \sigma^*_{T|x}\Big(\mathcal{F}_\alpha[\sigma^*_{T|X}](x) - \mathcal{F}_\alpha[\sigma^{(n)}_{T|X}](x)\Big)\Big),
\end{aligned}
\tag{36}
$$

where $(a)$, $(b)$, and $(c)$ follow from the combination of (23) and (25), (30), and (27), respectively.

The above discussion manifests that if $\frac{1}{\gamma}\Big((J_{\gamma,\alpha}(\sigma^{(n+1)}_{T|X}, \sigma^{(n)}_{T|X}) - f_\alpha(\sigma^*_{T|X})) + \sum_x P_X(x) \operatorname{Tr} \sigma^*_{T|x}\Big(\mathcal{F}_\alpha[\sigma^*_{T|X}](x) - \mathcal{F}_\alpha[\sigma^{(n)}_{T|X}](x)\Big)\Big) > 0$, making $\gamma$ smaller makes the average divergence between $\sigma^*_{T|X}$ and $\sigma^{(n)}_{T|X}$ decrease faster. On the other hand, making $\gamma$ too small leads to a risk of violating the condition (18) (and, consequently, breaking the monotonicity of $J_{\gamma,\alpha}$).

**Remark 1** *The reference [22, Section III] considered a general setting. If $\sigma_{T|X}$ is a single density matrix, our method can be considered as a special case of their setting. However, since $\sigma_{T|X}$ is classical-quantum channel in our case, our analysis is not a special case of their setting.*

**Remark 2** *The references [9, Appendix A] [24, Appendix A] considered the case when the systems $X, Y, T$ are quantum systems and $\alpha = 1$. They derived a necessary condition for the solution of the minimization problem by using Lagrange multiplier method in the same way as [1, 4]. Using the obtained condition, they [9, Appendix C] [24, Appendix C] also proposed an iterative algorithm to find a solution to satisfy the necessary condition. It seems that their necessary condition is the same as (31) with $\gamma = \alpha = 1$. However, they did not discuss the convergence to a local minimizer in their algorithm.*

## 2.5 Numerics on the effects of different $\gamma$

To see the effect of different $\gamma$, let us take a look at a concrete example: Consider a single-qubit quantum system $Y$ and a classical register $X$ with size $2^8$. Then, we assume that $P_X$ is the uniform distribution over $\mathcal{X} = \{0, \ldots, 2^8 - 1\}$, and the density $\rho_{Y|x}$ is given as $\rho_{Y|x} = \rho(\theta_x, \lambda_x)$, where

$$
\rho(\theta, \lambda) := \exp\left(i\theta\sigma_x\right)\begin{pmatrix} 1-\lambda & 0 \\ 0 & \lambda \end{pmatrix}\exp\left(-i\theta\sigma_x\right), \tag{37}
$$

where $\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is the Pauli-$X$ matrix. The parameters $\theta_x$ and $\lambda_x$ are randomly chosen.

Then, the ensemble we consider admits the following joint density matrix:

$$
\hat{\rho}_{XY} = \sum_x P_X(x)|\pi(x)\rangle\langle\pi(x)| \otimes \rho\left(\theta_x, \lambda_x\right) \tag{38}
$$

with $\rho\left(\theta_x, \lambda_x\right)$ given by Eq. (37).

Now, we apply our QIB algorithm (i.e., Algorithm 1) to the ensemble (38). We consider a classical $T$ whose size is the square root of $|\mathcal{X}|$ (i.e., $|\mathcal{T}| = 2^4$). We set $\alpha = 1$, and $\beta = 10$. Our focus will be the effects of different choices of the acceleration parameter $\gamma$. As shown in Fig. 2, the choice of $\gamma$ is crucial for the performance, more specifically, the efficiency and the convergence, of the QIB algorithm.

Two interesting phenomena are manifested by our numerics: For one thing, choosing a smaller $\gamma$ will accelerate the course of convergence. As shown in Fig. 2, by choosing a suitably smaller value of $\gamma$ (e.g., 0.8 or 0.5), our QIB algorithm achieves convergence faster than the existing QIB algorithm [9, 24], which corresponds to Algorithm 1 with $\gamma = 1$. For the other, choosing a too small $\gamma$ will ruin the convergence property of the QIB algorithm. For instance, when $\gamma$ is chosen to be 0.4, $f_\alpha$ jumps up after a few iterations and ends up in a much larger value than its initial value.

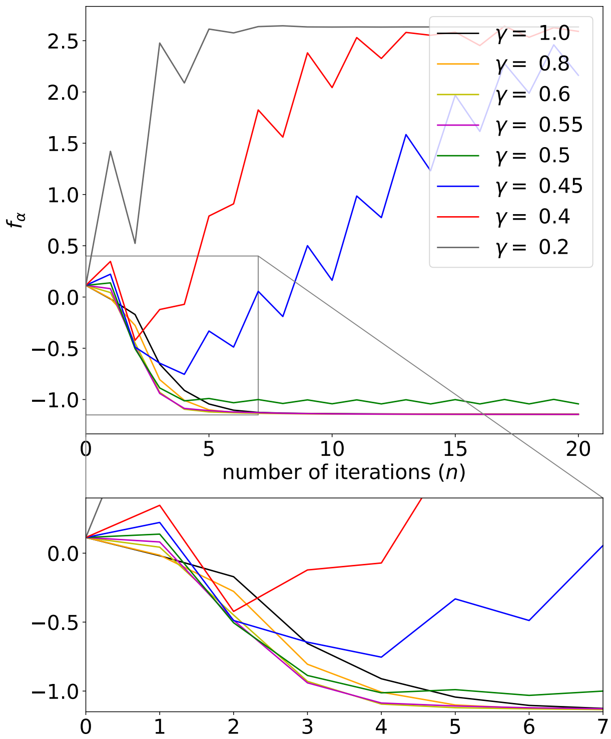

Figure 2: **Performance of Algorithm 1 for different $\gamma$.** We apply Algorithm 1 ($|\mathcal{T}| = 16$, $\alpha = 1$, and $\beta = 10$) to the joint state (38). The information bottleneck $f_\alpha$ is plotted as a function of the number of iterations for different values of $\gamma$. The green curve with $\gamma = 0.55$ converges most quickly. It significantly improves the convergence speed in comparison with the black line with $\gamma = 1$. The blue curve with $\gamma = 0.45$ goes down even faster in the beginning but gets overtaken after a few iterations. Finally, it goes up around $n = 7$. It shows that $\gamma = 0.45$ does not satisfy the condition (A1) for $n \geq 7$.

In conclusion, the numerics has justified our theoretical analysis (see Section 2.4) on the importance of choosing a suitable $\gamma$. We emphasize that our contribution in this direction is twofold:

1. We proposed a method of accelerating the QIB algorithm, making it converge within fewer rounds of iteration, by introducing a new parameter $\gamma$ and setting it to be smaller than one.
2. We showed that the QIB algorithm cannot achieve the desired minimal value of $f_\alpha$ if $\gamma$ is too small.

## 2.6 Choice of $\beta$

The output of our QIB algorithms depend not only on $\rho_{XY}$ [cf. (1)] but also on the choice of $\alpha$ and $\beta$. Intuitively, a larger $\beta$ improves the faithfulness (as it makes $I(Y:T)$ more significant in $f_\alpha$), while a smaller $\beta$ leads to more compression (as it makes $I(X:T)$ more significant in $f_\alpha$). Somehow surprisingly, the choice of $\beta$ is not completely free: In the following, we show that the QIB algorithm will yield a trivial $\sigma_{T|X}$ if $\beta$ is too small.

To consider the relation between the choice of $\beta$ and the resultant information on $T$, we introduce the following condition for a subset $\mathcal{S} \subset \mathcal{S}_{X\to T}$, where $\mathcal{S}_{X\to T}$ is the the set of all c-q channels from $X$ to $T$, i.e., the set $\{\sigma_{T|X} = (\sigma_{T|x})_{x\in\mathcal{X}}\}$:

**(A2)** For any two distinct elements $\sigma_{T|X}, \sigma'_{T|X} \in \mathcal{S}$, $\sum_x P_X(x) \operatorname{Tr}_T \sigma_{T|x}(\mathcal{F}_\alpha[\sigma_{T|X}](x) - \mathcal{F}_\alpha[\sigma'_{T|X}](x)) > 0$

The condition (A2) is unitarily invariant, i.e., the pair $(\sigma_{T|X}, \sigma'_{T|X})$ satisfies the condition (A2), if and only if the pair $(U\sigma_{T|X}U^\dagger, U\sigma'_{T|X}U^\dagger)$ satisfies the condition (A2) for any unitary $U$ on $T$. Hence, we choose $\mathcal{S}$ as a unitarily invariant subset.

**Theorem 4** *Assume that a unitarily invariant subset $\mathcal{S}$ satisfies (A2). Let $\sigma^M_{T|X} := \operatorname{argmin}_{\sigma_{T|X}} f_\alpha(\sigma_{T|X})$ be the solution to the QIB problem. When $\sigma^M_{T|X}$ belongs to $\mathcal{S}$, $\sigma^M_{T|x}$ is the maximally mixed state on $T$ for any $x$.*

If $\sigma^M_{T|x}$ is the maximally mixed state for every $x$, $T$ is uncorrelated with $Y$ and does not contain any meaningful information. In other words, when the assumption for Theorem 4 holds, the solution of the QIB problem is not useful. Hence, we need to choose the parameters $\alpha, \beta$ such that condition (A2) does not hold.

Now we discuss how to avoid the condition (A2). The LHS of (A2) is evaluated as

$$
\begin{aligned}
&\sum_x P_X(x) \operatorname{Tr}_T \sigma_{T|x}(\mathcal{F}_\alpha[\sigma_{T|X}](x) - \mathcal{F}_\alpha[\sigma'_{T|X}](x)) \\
=& \operatorname{Tr}_{TY} \sum_x P_X(x)(\sigma_{T|x} \otimes \rho_{Y|x})\Big( - (\log \sigma_T[\sigma_{T|X}] - \log \sigma_T[\sigma'_{T|X}]) + \alpha(\log \sigma_{T|x} - \log \sigma'_{T|x}) \\
&+ \beta\Big((\log(\sigma_T[\sigma_{T|X}] \otimes \rho_Y) - \log(\sigma_T[\sigma'_{T|X}] \otimes \rho_Y)) - (\log \sigma_{YT}[\sigma_{T|X}] - \log \sigma_{YT}[\sigma'_{T|X}])\Big)\Big) \\
=& \operatorname{Tr}_{TY} \sum_x P_X(x)(\sigma_{T|x} \otimes \rho_{Y|x})\Big( - (\log \sigma_T[\sigma_{T|X}] - \log \sigma_T[\sigma'_{T|X}]) + \alpha(\log P_X(x)\sigma_{T|x} - \log P_X(x)\sigma'_{T|x}) \\
&+ \beta\Big((\log(\sigma_T[\sigma_{T|X}] \otimes \rho_Y) - \log(\sigma_T[\sigma'_{T|X}] \otimes \rho_Y)) - (\log \sigma_{YT}[\sigma_{T|X}] - \log \sigma_{YT}[\sigma'_{T|X}])\Big)\Big) \\
=& - D(\sigma_T[\sigma_{T|X}]\|\sigma_T[\sigma'_{T|X}]) + \alpha D(\sigma_{XT}[\sigma_{T|X}]\|\sigma_{XT}[\sigma'_{T|X}]) \\
&- \beta\big(D(\sigma_{YT}[\sigma_{T|X}]\|\sigma_{YT}[\sigma'_{T|X}]) - D(\sigma_T[\sigma_{T|X}]\|\sigma_T[\sigma'_{T|X}])\big), \qquad (39)
\end{aligned}
$$

where $\sigma_{XT}[\sigma_{T|X}] := \sum_x P_X(x)\sigma_{T|x}[\sigma_{T|X}] \otimes |x\rangle\langle x|$. Since $D(\sigma_{YT}[\sigma_{T|X}]\|\sigma_{YT}[\sigma'_{T|X}]) \geq D(\sigma_T[\sigma_{T|X}]\|\sigma_T[\sigma'_{T|X}])$, the coefficient of $\beta$ is a negative value. Hence, a smaller $\beta$ has a possibility to satisfy the condition (A2). That is, to obtain a useful solution, we need to choose $\beta$ to be a sufficiently large value.

*Proof of Theorem 4:* Let $U$ be an arbitrary unitary on $\mathcal{T}$. We define $\sigma^{M'}_{T|X}$ by $\sigma^{M'}_{T|x} = U\sigma^{M}_{T|x}U^\dagger$. Substituting $\sigma^{(n)}_{T|x}$ with $\sigma^{M'}_{T|x}$ in (36), we have

$$
\begin{aligned}
0 =& \sum_x P_X(x) D(\sigma^M_{T|x}\|\sigma^{M'}_{T|x}) \\
&- \sum_x P_X(x) D(\sigma^M_{T|x}\|\hat{\sigma}_{\gamma,\alpha,T|x}[\sigma^{M'}_{T|X}]) \\
=& \frac{1}{\gamma}(f_\alpha(\sigma^{M'}_{T|X}) - f_\alpha(\sigma^{M}_{T|X})) \\
&+ \frac{1}{\gamma}\sum_x P_X(x) \operatorname{Tr} \sigma^M_{T|x}\Big(\mathcal{F}_\alpha[\sigma^M_{T|X}](x) - \mathcal{F}_\alpha[\sigma^{M'}_{T|X}](x)\Big) \qquad (40) \\
=& \frac{1}{\gamma}\sum_x P_X(x) \operatorname{Tr} \sigma^M_{T|x}\Big(\mathcal{F}_\alpha[\sigma^M_{T|X}](x) - \mathcal{F}_\alpha[\sigma^{M'}_{T|X}](x)\Big). \qquad (41)
\end{aligned}
$$

Thus, the condition (A2) implies $\sigma^M_{T|X} = \sigma^{M'}_{T|X}$. $\sigma^M_{T|x}$ is the completely mixed state on $T$ for any $x$. ■

## 3 Classical system $T$

Next, we consider the case when $T$ is *constrained* to be a classical system. We stress that this is a different minimization from the previously discussed one with a quantum system $T$, whose minimum may not be attainable with a classical $T$. Instead, our objective function now is

$$
\mathcal{I}^c_{\alpha,\beta} := \min_{\sigma_{T|X}:diagonal} f_\alpha(\sigma_{T|X}). \qquad (42)
$$

Therefore, we need to re-examine the validity of our previous analyses.

Let us start with the form of QIB algorithm. Fortunately, our algorithm with a quantum system $T$ can be applied to this case, simply with the adaptation that the states $\sigma_{T|x}$ are limited to diagonal density matrices with respect to the basis $\{|t\rangle\}$ of $T$. Under this condition, the states $\hat{\sigma}_{\gamma,\alpha,T|x}[\sigma_{T|X}]$ are also diagonal density matrices. Therefore, when we set the initial state as diagonal density matrices, Algorithm 1 works for this case.

The above discussion leads to an interesting observation as follows. The convergent $\sigma^*_{T|X}$ with initial diagonal $\sigma_{T|X}$ satisfies the condition (10) and it is also diagonal. That is, if the minimum with classical $T$ is strictly larger than the minimum with quantum $T$, the minimum with classical $T$ is an example for the following statement: A solution of the condition (10) does not necessarily give the minimum of $f_\alpha$ with quantum $T$. This fact shows the possible risk that a solution to (10) might be a saddle point or a local minimum rather than the global minimum for $f_\alpha$ with quantum $T$.

When the states $\sigma_{T|x}$ are limited to diagonal density matrices with respect to the basis $\{|t\rangle\}$ of $T$, $\sigma_{TY}[\sigma_{T|X}]$ is commutative with $\sigma_T[\sigma_{T|X}]$ so that we can define $\sigma_{Y|T}[\sigma_{T|X}] := \sigma_{TY}[\sigma_{T|X}]\sigma_T[\sigma_{T|X}]^{-1}$. Then, $\hat{\sigma}_{\gamma,\alpha,T}[\sigma_{T|X}](x)$ is simplified as follows.

$$
\begin{aligned}
&\log \hat{\sigma}_{\gamma,\alpha,T}[\sigma_{T|X}](x) \\
=&(1 - \frac{\alpha}{\gamma}) \log \sigma_{T|x} + \frac{1}{\gamma} \log \sigma_T[\sigma_{T|X}] \\
&- \frac{\beta}{\gamma} \operatorname{Tr}_Y \Big(\rho_{Y|x}(\log \rho_Y - \log \sigma_{Y|T}[\sigma_{T|X}])\Big). \qquad (43)
\end{aligned}
$$

The notion of unitary invariance is reduced to invariance under permutations on $T$, and the condition (A2) is invariant under permutations on $T$. Then, Theorem 4 can be rewritten as follows.

**Theorem 5** *Assume that a subset $\mathcal{S}$ satisfies (A2) and is invariant under any permutation on $T$. Let $\sigma^*_{T|X}$*

*be the minimizer of* $\min_{\sigma_{T|X}:diagonal} f_\alpha(\sigma_{T|X})$*. When* $\sigma^*_{T|X}$ *belongs to* $\mathcal{S}$*,* $\sigma^*_{T|x}$ *is the uniform distribution over* $T$ *for any* $x$*.*

Theorem 5 can be shown in the same way as Theorem 4.

In this case, we can make a more precise discussion for the condition (A2). For this purpose, we consider the maximum ratio

$$\kappa := \max_{Q_X, Q'_X} \frac{D(\sum_x Q_X(x)\rho_{Y|x} \| \sum_x Q'_X(x)\rho_{Y|x})}{D(Q_X \| Q'_X)}. \tag{44}$$

The inequality $\kappa \leq 1$ follows from the information processing inequality for the map $Q_X \mapsto \sum_x Q_X(x)\rho_{Y|x}$. In this condition, $\sigma_T[\sigma_{T|X}]$ is written as $\sum_t Q_T[\sigma_{T|X}](t)|t\rangle\langle t|$ by using a distribution $Q_T[\sigma_{T|X}]$. Then, the LHS of (A2) is simplified as

$$\begin{aligned}
&\sum_x P_X(x) \operatorname{Tr}_T \sigma_{T|x}(\mathcal{F}_\alpha[\sigma_{T|X}](x) - \mathcal{F}_\alpha[\sigma'_{T|X}](x)) \\
=&(\beta-1)D(\sigma_T[\sigma_{T|X}]\|\sigma_T[\sigma'_{T|X}]) \\
&+ \alpha D(\sigma_{XT}[\sigma_{T|X}]\|\sigma_{XT}[\sigma'_{T|X}]) \\
&- \beta D(\sigma_{YT}[\sigma_{T|X}]\|\sigma_{YT}[\sigma'_{T|X}]) \\
=&(\alpha-1)D(\sigma_T[\sigma_{T|X}]\|\sigma_T[\sigma'_{T|X}]) \\
&+ \sum_t Q_T[\sigma_{T|X}](t)\Big(\alpha D(\sigma_{X|T=t}[\sigma_{T|X}]\|\sigma_{X|T=t}[\sigma'_{T|X}]) \\
&- \beta D(\sigma_{Y|T=t}[\sigma_{T|X}]\|\sigma_{Y|T=t}[\sigma'_{T|X}])\Big) \\
\geq&(\alpha-1)D(\sigma_T[\sigma_{T|X}]\|\sigma_T[\sigma'_{T|X}]) \\
&+ (\alpha - \beta\kappa)\sum_t Q_T[\sigma_{T|X}](t) \\
&\cdot D(\sigma_{X|T=t}[\sigma_{T|X}]\|\sigma_{X|T=t}[\sigma'_{T|X}]).
\end{aligned} \tag{45}$$

When the condition $\alpha \geq 1, \frac{\alpha}{\kappa} > \beta$ holds, the LHS of (A2) is positive for $\sigma_{T|X} \neq \sigma'_{T|X}$. Hence, to extract useful $\sigma_{T|X}$, we need to choose $\beta$ to satisfy the condition $\beta > \frac{\alpha}{\kappa}$ with $\alpha = 1$. In fact, even when $\beta > \frac{\alpha}{\kappa}$, there is a possibility that a permutation-invariant subset $\mathcal{S}$ satisfies (A2). Due to Theorem 5, when a permutation-invariant subset $\mathcal{S}$ satisfies (A2), a useful solution does not belong to the subset $\mathcal{S}$. Hence, to obtain a useful solution, we need to choose $\beta$ sufficiently large beyond the above condition $\beta > \frac{\alpha}{\kappa}$ with $\alpha = 1$.

**Remark 3** *We consider the case with classical* $Y$ *and* $\gamma = \alpha$*. The operator* $\hat{\sigma}_{\alpha,T}[\sigma_{T|X}](x)$ *is simplified as follows.*

$$\begin{aligned}
&\hat{\sigma}_{\alpha,T}[\sigma_{T|X}](x) \\
=&\exp\Big(\frac{1}{\alpha}\log\sigma_T[\sigma_{T|X}] \\
&-\frac{\beta}{\alpha}\operatorname{Tr}_Y\Big(\rho_{Y|x}(\log\rho_Y - \log\sigma_{Y|T}[\sigma_{T|X}])\Big)\Big).
\end{aligned} \tag{46}$$

*In this case, the reference [31, (14) Section 3] proposed the following update rule:*

$$\hat{\tau}_{T|x}[\sigma_{T|X}] := \frac{1}{\operatorname{Tr}\hat{\tau}_T[\sigma_{T|X}](x)}\hat{\tau}_T[\sigma_{T|X}](x), \tag{47}$$

*where the operator* $\hat{\tau}_T[\sigma_{T|X}](x)$ *is defined as*

$$\begin{aligned}
&\hat{\tau}_T[\sigma_{T|X}](x) \\
:=&\exp\Big(\frac{1}{\alpha}\log\sigma_T[\sigma_{T|X}] \\
&-\frac{\beta}{\alpha}\operatorname{Tr}_Y\Big(\rho_{Y|x}(\log\rho_{Y|x} - \log\sigma_{Y|T}[\sigma_{T|X}])\Big)\Big).
\end{aligned} \tag{48}$$

*Since*

$$\log\hat{\tau}_T[\sigma_{T|X}](x) - \log\hat{\sigma}_T[\sigma_{T|X}](x) = \frac{\beta}{\alpha}D(\rho_{Y|x}\|\rho_Y), \tag{49}$$

*we have*

$$\begin{aligned}
&\hat{\tau}_{T|x}[\sigma_{T|X}] \\
=&\frac{1}{\operatorname{Tr} e^{\frac{\beta}{\alpha}D(\rho_{Y|x}\|\rho_Y)}\hat{\sigma}_T[\sigma_{T|X}](x)} e^{\frac{\beta}{\alpha}D(\rho_{Y|x}\|\rho_Y)}\hat{\sigma}_T[\sigma_{T|X}](x) \\
=&\hat{\sigma}_{T|x}[\sigma_{T|X}].
\end{aligned} \tag{50}$$

*That is, the update rule (47) by [31, (14) Section 3] is the same as ours of this special case. In particular, the update rule (47) with* $\alpha = 1$ *coincides with the update rule by the reference [32].*

**Remark 4** *When the system* $Y$ *is classical and* $\alpha = 1$*, the reference [9, Appendix B] claimed that there is no difference between the optimal value with quantum* $T$ *and the optimal value with classical* $T$*. Since their algorithm works with* $T$ *of a fixed size, it can be considered that they claimed the above statement when the size of* $T$ *is fixed. However, their proof (see [9, Appendix B II]) contains a gap: The statement under Eq. (B23) that "the Lagrangian is invariant under a measurement of the memory* $M$ *in a chosen basis* $|m\rangle$*" is not backed by a rigorous mathematical proof. It is thus unclear whether this statement and, consequently, the claim that there is no quantum advantage are correct. On the other hand, as we show next, the optimal value with quantum* $T$ *can be strictly smaller than the optimal value with classical* $T$*. That is, the claim in [9, Appendix B] contradicts with our result of the next section.*

## 4 Quantum advantage for $T$

To see the advantage of quantum system $T$ over classical system $T$, we discuss several examples with the strict inequality

$$\mathcal{I}_{\alpha,\beta} < \mathcal{I}^c_{\alpha,\beta}. \tag{51}$$

We provide an analytical example in this section and a numerical example with application in quantum machine learning in Section 5.2 when the size of the system $T$ is fixed. Generally, to achieve the optimal performance, we need to choose the system $T$ as a sufficiently large dimensional system. However, in this section, to provide analytical examples, we fix the size of the system $T$ to a certain value.

Assume that $\mathcal{Y}$ is a classical system of size $d$. The size of $\mathcal{X}$ is $k$ times of the size $d$ of $\mathcal{Y}$. We assume that $\mathcal{X}$ is given as $\mathcal{X}_1 \times \mathcal{X}_2$ with $\mathcal{X}_1 = \mathcal{Y}$ and $|\mathcal{X}_2| = k$. The distribution of $X$ is assumed to be uniform. We focus on the quantum system $T$ with the dimension $n < d$.

**Lemma 6** *When $\beta \geq 1$ and $\beta \geq \alpha$, we have*

$$\mathcal{I}_{\alpha,\beta} = (1-\beta)\log n \tag{52}$$

*Proof:* First, we show a bound on the QIB for generic (quantum) $T$. For any $\sigma_{T|x}$, we have $H(T) \geq I(T:X) \geq I(T:Y)$. Hence, the relation $\beta - \alpha \geq 0$ implies $-(\beta-\alpha)I(T:Y) \geq -(\beta-\alpha)H(T)$. Hence, we have

$$\begin{aligned} f_\alpha(\sigma_{T|x}) &= (1-\alpha)H(T) + \alpha I(T:X) - \beta I(T:Y) \\ &\geq (1-\alpha)H(T) - (\beta-\alpha)I(T:Y) \geq (1-\beta)H(T). \end{aligned} \tag{53}$$

Since $H(T) \leq \log n$ and $1 - \beta \leq 0$, we obtain

$$\mathcal{I}_{\alpha,\beta} \geq (1-\beta)\log n. \tag{54}$$

The above bound is tight. Indeed, we choose $\sigma_{T|x_1,x_2}$ as the pure state $\sum_{t=1}^n \frac{1}{\sqrt{n}} e^{\frac{2\pi x_1}{n} i}|t\rangle$. Then, we have $H(T) = \log n$. Also, $H(T) = I(T:X) = I(T:Y)$. Therefore, $f_\alpha(\sigma_{T|x}) = (1-\beta)\log n$. ■

Next, we focus on the case when $T$ is a classical system of dimension $n < d$.

**Lemma 7** *Assume that $d = mn + l$ with $0 \leq l < n$. When $\beta \geq 1 \geq \alpha$, we have*

$$\mathcal{I}^c_{\alpha,\beta} = (1-\beta)\Big(\frac{l(m+1)}{d}\log\frac{d}{m+1} + \frac{(n-l)m}{d}\log\frac{d}{m}\Big) \tag{55}$$

*Proof:* Any channel $\sigma_{T|x}$ can be written as a probabilistic mixture of deterministic channels $\sigma^j_{T|x}$. That is, we have

$$\sigma_{T|x} = \sum_j p_j \sigma^j_{T|x}. \tag{56}$$

Since $Y$ is independent of $X_2$ and the random variable $J$ describing the choice of $j$, we have

$$\begin{aligned} I(T:Y|JX_2) =& I(T:Y|JX_2) + I(Y:JX_2) \\ =& I(TJX_2:Y) \geq I(T:Y). \end{aligned} \tag{57}$$

Also, we have

$$H(T) \geq H(T|JX_2). \tag{58}$$

Then, we have

$$\begin{aligned} & f_\alpha(\sigma_{T|x}) \overset{(a)}{\geq} (1-\alpha)H(T) - (\beta-\alpha)I(T:Y) \\ \overset{(b)}{\geq} & (1-\alpha)H(T|JX_2) - (\beta-\alpha)I(T:Y|JX_2), \end{aligned} \tag{59}$$

where $(a)$ follows from (53), and $(b)$ follows from (57) and (58). The minimization of $(1-\alpha)H(T|JX_2) - (\beta-\alpha)I(T:Y|JX_2)$ equals the minimization of the same function under the condition that $\sigma_{T|X}$ is a deterministic channel and $\sigma_{T|x_1x_2}$ depends only on $x_1$.

Under this condition, we have $I(T:X) = I(T:X_1) = I(T:Y)$, which implies the equality in $(a)$ at (59). Therefore, for the minimization, we can impose this condition, i.e., the variable $T$ is determined only by $X_1 = Y$, which implies $I(T:Y) = H(T)$. In this case, we have $f_\alpha(\sigma_{T|x}) = (1-\beta)H(T)$. In the classical case, the maximum entropy $H(T)$ among deterministic channels is achieved when the distribution $(P_T(t))_{t=1}^n$ as close as possible to the uniform distribution, i.e., $P_T = (\overbrace{\frac{m+1}{d}, \ldots, \frac{m+1}{d}}^{l}, \overbrace{\frac{m}{d}, \ldots, \frac{m}{d}}^{n-l})$. Hence, the maximum entropy $H(T)$ is $\frac{l(m+1)}{d}\log\frac{d}{m+1} + \frac{(n-l)m}{d}\log\frac{d}{m}$. Therefore, we obtain the desired statement. ■

When the conditions of Lemma 7 hold, $d$ cannot be divided by $n$. In this case, since $\frac{l(m+1)}{d}\log\frac{d}{m+1} + \frac{(n-l)m}{d}\log\frac{d}{m}$ is strictly smaller than $\log n$, when the state $\rho_{XY}$ is close to the state $\sum_x \frac{1}{d}|x,x\rangle\langle x,x|$, the strict inequality (51) holds. There is clearly an advantage of using a quantum $T$.

# 5 Quantum feature maps with QIB

## 5.1 Information bottleneck in supervised learning

Supervised learning is a cornerstone of machine learning. Given a dataset $\{(x,y)\}$ sampled from an unknown probability distribution $P_{XY}$, a general supervised learning task is to find a classifier such that, for any testing data $(x', y')$ sampled from the same distribution $P_{XC}$, it predicts the label $y'$ with as high accuracy as possible given $x'$.

Remarkably, recent studies [8, 28, 33] on the information bottleneck theory showed evidences that the training phase of deep learning can be divided into two stages. In the first stage, a representation $T$ of $X$ that faithfully encodes its correlation with $Y$ is found, featured by increasing $I(T:Y)$. In the second stage, the size of $T$ is compressed, featured by decreasing $I(T:X)$. This result suggests that finding an efficient and compressed representation of $X$ facilitates data classification.

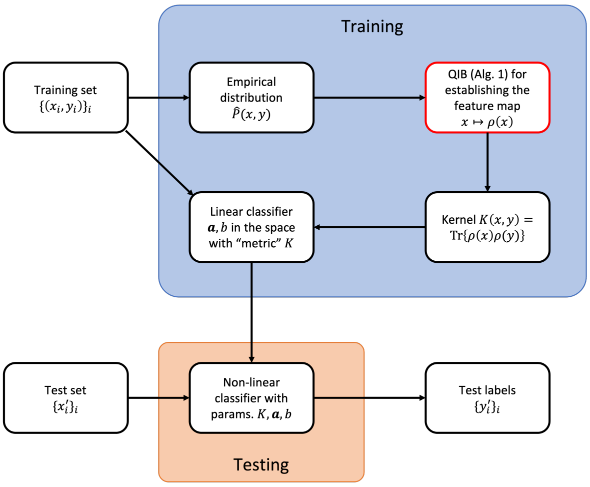
Training
Training set $\{(x_i, y_i)\}_i$
Empirical distribution $\hat{P}(x,y)$
QIB (Alg. 1) for establishing the feature map $x \mapsto \rho(x)$
Linear classifier $\boldsymbol{a}, b$ in the space with “metric” $K$
Kernel $K(x,y) = \mathrm{Tr}\{\rho(x)\rho(y)\}$
Test set $\{x_i'\}_i$
Non-linear classifier with params. $K, \boldsymbol{a}, b$
Test labels $\{y_i'\}_i$
Testing

Figure 3: **Data classification with quantum feature maps.** The flowchart illustrates the training phase and the testing phase of data classification using the technique of quantum feature maps. The part where our QIB algorithm is applied is highlighted.

## 5.2 Quantum feature maps

Following the above intuition, we propose a classical-quantum hybrid algorithm of data classification, by combining the QIB algorithm with the kernel method. The idea is illustrated in the flowchart in Fig. 3. Given a training dataset $\mathcal{S}_{\mathrm{train}}$, the algorithm first identifies an efficient representation $T$ of $X$ by minimising the information bottleneck $f_\alpha := H(T) - \alpha H(T|X) - \beta I(T:Y)$. Then a classifier is constructed that yields a prediction $\hat{Y}$ based on the state in $T$ corresponding to the value of $X$. For simplicity, we consider for now the case when $Y \in \{1, -1\}$ is binary. In the first step, we set the representation $T$ to be a quantum state $\rho(x)$ that depends on the data $x$, and we obtain $\rho(x)$ via Algorithm 1. In the second step, we use a linear classifier

$$c_{\mathrm{QIB}}\left(\rho(\tilde{x})\right) = \mathrm{sgn}\left(\mathrm{Tr}[A\rho(\tilde{x})] + b\right) \tag{60}$$

where $A$ is a Hermitian operator and $b \in \mathbb{R}$. We further consider $A$ that can be expressed as a linear combination $A = \sum_{x:(x,y)\in\mathcal{S}_{\mathrm{train}}} a_x \rho(x)$, and the classifier has the reduced form

$$c_{\mathrm{QIB}}\left(\rho(\tilde{x})\right) = \mathrm{sgn}\left(\sum_{x:(x,c)\in\mathcal{S}_{\mathrm{train}}} a_x K(x,\tilde{x}) + b\right), \tag{61}$$

where $K(x,\tilde{x})$ is the *kernel* function, in our case given by the Hilbert-Schmidt (HS) inner product of quantum states and can be evaluated by performing the SWAP test on a quantum computer:

$$K(x,y) = \mathrm{Tr}\{\rho(x)\rho(y)\}. \tag{62}$$

The algorithm is summarised as follows:

**Algorithm 2** QIB for data classification

```
input: A training data set $\mathcal{S}_{\mathrm{train}} = \{(x,y)\}$; configuration $(\alpha,\beta,\gamma)$.
input: A classifier $c_{\mathrm{QIB}} : X \to \hat{Y}$.
1) Generate an empirical distribution $\hat{P}(x,y)$ from $\mathcal{S}_{\mathrm{train}}$.
2) Run Algorithm 1 with $\hat{P}(x,y)$ as input and certain (adjustable) parameters $\alpha$,$\beta$,$\gamma$.
3) Compute the kernel $K$ in Eq. (61) using the output of Step 2).
4) Train the classifier (61) with $\mathcal{S}_{\mathrm{train}}$ and output the trained classifier.
```

We remark that the quantum kernel method, where a mapping $x \to \rho(x)$ is constructed for better classification, has been a hot topic recently (see, e.g., [5, 11, 17, 20, 25, 26]). The key distinction between existing works and our present method is the following: In existing works, the parameter $x$ is passed to a parametrised (a.k.a. variational) quantum circuit that prepares the state $\rho(x)$. One needs to train the circuit parameters on a quantum computer to obtain a good mapping $x \mapsto \rho(x)$, which is called a feature map. In the near term, this method might be subject to the physical limitations of quantum devices. In contrast, in our present method $\rho(x)$ is directly computed via a simple iterative algorithm. Therefore, there are two possible ways of realizing our present method, i.e., Algorithm 2. In the near term, we can regard Algorithm 2 as a “quantum-inspired” classical algorithm, and evaluate everything on a classical computer. When large-scale quantum computing becomes feasible, Algorithm 2 can be readily “quantised”. Indeed, the evaluation of $\rho(x)$ in each iteration requires subroutines that compute matrix powers and logarithm and solve linear systems, which have already been developed in Refs. [7, 10, 18, 19].

## 5.3 Numerical experiments

As a proof-of-principle experiment, we tested the performance of our QIB classifier on a dataset on $\mathbb{R}^2$, generated in the following way: First, we define the discrete sets $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ and $\mathcal{Y}$, with $\mathcal{X}_1 = \mathcal{Y} = \{0,1,2\}$ and $\mathcal{X}_2 = \{0,1,\ldots,9\}$. To apply our classification method, we arbitrarily choose permutation $\pi$, and generate $n' = 400$ independent and identically distributed data $(\tilde{X}_{1,i}, \tilde{X}_{2,i}, Y_i)$ for $i = 1,\ldots,n'$ as follows. We independently generate $(X_{1,i}, X_{2,i}, Y_i)$ according to the following distribution

$$P_{XY}(x_1,x_2,y) := P_Y(y) Q_{X_1|Y}(x_1', y) Q_{X_2|X_1}(x_2', x_1'), \tag{63}$$

where $P_Y$ is the uniform distribution over $Y$, $Q_{X_1|Y}(x_1,y) = \delta(x_1,y)$, $Q_{X_2|X_1}(x_2,x_1) = \frac{\delta(x_1,x_2)+1}{|\mathcal{X}_2|+1}$, and $(x_1', x_2') = \pi(x_1,x_2)$. Next, we generate the random variables $\tilde{X}_{j,i} := X_{j,i} + R_{j,i}$, where the ran-

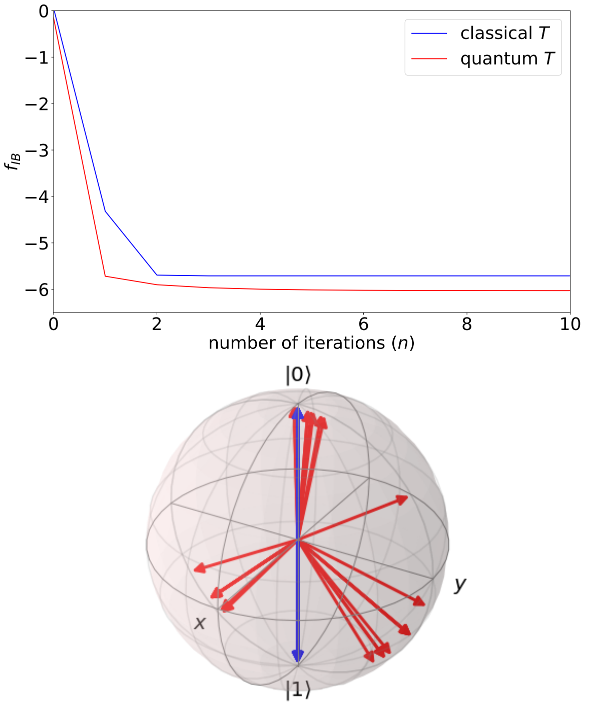

Figure 4: **Quantum vs classical feature maps.** We run Algorithm 1 with $\alpha = \gamma = 1$, $\beta = 15$ on the distribution $\tilde{P}_{XY}$ based on the training data, and compare the converging values of QIB when $T$ is classical (i.e., a probabilistic bit) and when $T$ is quantum (i.e., a single qubit). This numerics shows the advantage of use of quantum $T$ over classical $T$. The final feature maps with quantum $T$ (plotted in red) and with the classical-$T$ (plotted in blue) are visualised in the Bloch ball.

dom variable $R_{j,i}$ is subject to the uniform distribution in the interval $[0, 1.2)$ unless $i = 1, X_i = 2$ nor $i = 2, X_i = 9$, it is subject to the uniform distribution in the interval $[0, 1)$ otherwise. Then, using the obtained data $(\lfloor \tilde{X}_{1,i} \rfloor, \lfloor \tilde{X}_{2,i} \rfloor, Y_i)$ with $i = 1, \ldots, n$, we define its empirical distribution $\tilde{P}_{XY}$. We apply Algorithm 1 to the distribution $\tilde{P}_{XY}$ as Fig. 4. In the case with the distribution $\tilde{P}_{XY}$, Algorithm 1 with quantum $T$ can realize a smaller $f_\alpha$ than Algorithm 1 with classical $T$, which shows the advantage of quantum $T$ over classical $T$.

In the classification experiment, 50% of the data are used as the training set and the rest are used as the testing set. The kernel is constructed with Algorithm 2 with $\alpha = 1, \beta = 15, \gamma = 1$, a single-qubit register $T$, and 10 iterations. We consider both when $T$ is a generic qubit system and when $T$ is restricted to a binary classical system, and we compare their performance. As can be seen from Fig. 4, the case of quantum $T$ has lower IB value than the case of classical $T$. The final feature map $\sigma_{T|X}$ for the quantum $T$ case suffers from certain degree of dispersion due to the random noise $r_1, r_2$, but the quantum features still form 3 clusters. In contrast, the final $\sigma_{T|X}$ in the classical $T$ case maps different values of $X$ into two clusters.

The effect of the above distinction is made apparent in the classification performance. In Fig. 5, the performance of the classifiers constructed from the kernels are illustrated via their decision regions. It can be seen that, since the classical-$T$ feature map groups $X$ into two clusters, its resultant classifier gives a binary prediction on any input data, giving up the least possible label. In stark contrast, the quantum-$T$ feature map utilizes the full Bloch ball to generate 3 clusters, leading to a much higher accuracy of prediction. The advantage of a genuinely quantum feature map is thus manifested by this numerical example.

For reference, in Fig. 5, we also plot the performance of two standard methods of classical feature maps. The referential methods (linear kernel and polynomial kernel) achieve accuracies (defined by the ratio of correct predictions in the testing set) 0.64 and 0.62, which is slightly higher than the classical-$T$ information bottleneck kernel (0.565) but much lower than the QIB kernel (0.92). This further justifies the superior performance of our QIB method in classification.

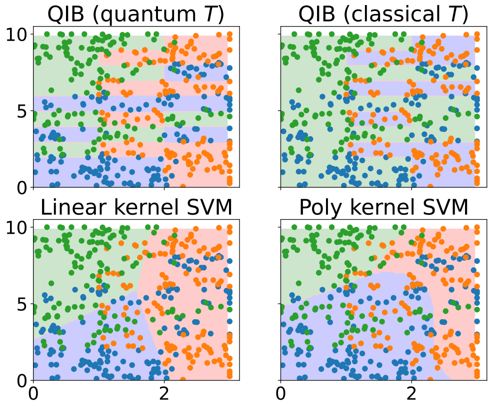

Figure 5: **Decision regions of the QIB classifier and reference classifiers.** The decision regions of the QIB classifier, the classical-$T$ IB classifier, and two reference classifiers are plotted together with the test data. The different dot colors correspond to data with different labels, and the color of each region corresponds to the prediction made by the classifier for data in that region.

## 6 Quantum deterministic information bottleneck (QDIB)

Considering the limit $\alpha \to +0$, the paper [31] proposed deterministic IB, which minimize $f_0$. Now, we consider this minimization with quantum systems

$T, Y$ and classical system $X$. First, we define

$$\begin{aligned}&\hat{\sigma}_{0,T|x}[\sigma_{T|X}]\\ :=&\frac{1}{\mathrm{Tr}\,\sigma_{T|x}P_{T|x}[\sigma_{T|X}]}P_{T|x}[\sigma_{T|X}]\sigma_{T|x}P_{T|x}[\sigma_{T|X}],\end{aligned} \tag{64}$$

where $P_{T|x}[\sigma_{T|X}]$ is the projection to the maximum eigenvalue of the operator $(1-\beta)\log\sigma_T[\sigma_{T|X}] + \beta\,\mathrm{Tr}_Y\,\rho_{Y|x}(\log\sigma_{YT}[\sigma_{T|X}] - \log\rho_Y)$.

Given an initial point $\sigma^{(1)}_{T|X}$, we propose the following update rule

$$\sigma^{(n+1)}_{T|X} := \hat{\sigma}_{0,T|X}[\sigma^{(n)}_{T|X}]. \tag{65}$$

As shown below, each step of this algorithm improves the value of the target function $f_0$.

The operator $\hat{\sigma}_{0,T|x}[\sigma_{T|X}]$ is characterized as

$$\hat{\sigma}_{0,T|x}[\sigma_{T|X}] = \lim_{\alpha\to 0}\hat{\sigma}_{\alpha,\alpha,T|x}[\sigma_{T|X}]. \tag{66}$$

Since Theorem 1 and (20) guarantee

$$\begin{aligned}&f_\alpha(\hat{\sigma}_{\alpha,\alpha,T|X}[\sigma_{T|X}])\\ =&J_{\alpha,\alpha}(\hat{\sigma}_{\alpha,\alpha,T|X}[\sigma_{T|X}],\hat{\sigma}_{\alpha,\alpha,T|X}[\sigma_{T|X}])\\ \le&J_{\alpha,\alpha}(\hat{\sigma}_{\alpha,\alpha,T|X}[\sigma_{T|X}],\sigma_{T|X})\\ \le&J_{\alpha,\alpha}(\sigma_{T|X},\sigma_{T|X}) = f_\alpha(\sigma_{T|X}),\end{aligned} \tag{67}$$

the limit $\alpha\to 0$ in (67) implies

$$f_{\alpha\to 0}(\hat{\sigma}_{0,T|X}[\sigma_{T|X}]) \le f_{\alpha\to 0}(\sigma_{T|X}]), \tag{68}$$

which shows that each step of this algorithm improves the value of the target function $f_{\mathrm{DIB}} := f_{\alpha\to 0}$.

**Algorithm 3** Quantum deterministic information bottleneck (QDIB) algorithm

```
Input: A joint state $\rho_{XY}$ [cf. (1)].
Create a counter $n$ as the number of iterations, initialized to 1.
repeat
    Choose $\sigma^{(n+1)}_{T|X}$ as
        $$\sigma^{(n+1)}_{T|x} = \frac{P_{T|x}[\sigma^{(n)}_{T|X}]\sigma^{(n)}_{T|x}P_{T|x}[\sigma^{(n)}_{T|X}]}{\mathrm{Tr}\left(\sigma^{(n)}_{T|x}P_{T|x}[\sigma_{T|X}]\right)} \tag{69}$$
    where $P_{T|x}[\sigma^{(n)}_{T|X}]$ is the projection on the space spanned by the eigenvectors of $\mathcal{F}_{\alpha=0}[\sigma^{(n)}_{T|X}](x)$ [cf. (15)] corresponding to the minimum eigenvalue.
    Set $n$ as $n+1$.
until convergence.
Output: A c-q channel $\sigma^{(n+1)}_{T|X}$
```

# 7 Approximate sufficient statistics from DIB

## 7.1 Task formulation

Next, we discuss how DIB can be used for the extraction of useful information under a classical-quantum (c-q) joint system composed of $X$ and $Y$ with the joint state $\rho_{XY} := \sum_x P_X(x)|x\rangle\langle x| \otimes \rho_{Y|x}$, where $X$ is a classical system and $Y$ is a quantum system. For example, assume that our interest is in the quantum phenomena in the quantum system $Y$. This quantum system $Y$ is correlated to the classical system $X$. However, there is a possibility that the classical system $X$ contains redundant information. In this case, it is useful to extract essential information from $X$ to describe the behavior of the quantum phenomena in the quantum system $Y$. To discuss the essential information, we introduce the concept of $\epsilon$-(approximate) sufficient statistics of the classical system $X$ with respect to the quantum system $Y$ while the papers [12, 36] discussed this concept when system $Y$ is a classical system.

A function $f$ from $X$ to $T$ is called a sufficient statistics of $X$ for the quantum system $Y$ when there exists a conditional distribution $P_{X|T}$ such that

$$\rho_{XY} = \sum_t P_{X|T}(x|t)|x\rangle\langle x| \otimes \sum_{x'\in f^{-1}(t)} P_X(x')\rho_{Y|x'}. \tag{70}$$

The above condition is equivalent to the condition

$$I(X:Y) = I(T:Y) \tag{71}$$

while in general we have the inequality $I(X:Y) \ge I(T:Y)$.

However, when we use sufficient statistics, we cannot remove a small correlation generated by a noise. As an example, suppose that the classical system $X$ is composed of two classical systems $X_1$ and $X_2$. Assume that we have a c-q state $\rho_{X_1X_2Y} = \sum_{x_1}\sum_{x_2} P_{X_1,X_2}(x_1,x_2)|x_1,x_2\rangle\langle x_1,x_2| \otimes \rho_{Y|x_1}$ with two classical systems $X_1$ and $X_2$.

We assume that we have already known the distribution $P_{X_1X_2}$ but we do not know $\rho_{Y|x}$. Also, we assume that we generate this state several times and apply the state estimation to the generated state. As a result, we obtain our estimate

$$\hat{\rho}_{X_1X_2Y} = \sum_{x_1}\sum_{x_2} P_{X_1X_2}(x_1,x_2)|x_1,x_2\rangle\langle x_1,x_2| \otimes \hat{\rho}_{Y|x_1,x_2}. \tag{72}$$

Since our estimate always has small error, $\hat{\rho}_{Y|x_1,x_2}$ is not exactly the same as $\rho_{Y|x_1}$, but it is close to $\rho_{Y|x_1}$. In this case, this difference should be considered as a noise. That is, the dependence of $X_2$ is not essential. It is better to consider that the correlation is given as $\hat{\rho}_{Y|x_1} := \sum_{x_2} P_{X_2|X_1}(x_2|x_1)\hat{\rho}_{Y|x_1,x_2}$

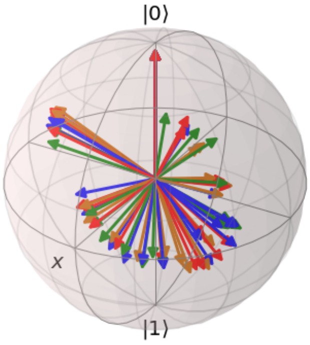

Figure 6: **Bloch representation of the estimated ensemble $\{\rho(\theta_{x_1,x_2},\lambda_{x_1,x_2})\}$.** As can be seen in the figure, the qubit states, especially those with higher purity, form several clusters in the Bloch ball. In each cluster, the states have the same value of $x_1$ and different values of $x_2$. This shows that the correlation between $X_1$ and $Y$ is higher than the correlation between $X_2$ and $Y$.

so that our estimate of $\rho_{X_1X_2Y}$ is given as $\sum_{x_1}\sum_{x_2} P_{X_1,X_2}(x_1,x_2)|x_1,x_2\rangle\langle x_1,x_2|\otimes\hat\rho_{Y|x_1}$.

For $\epsilon>0$, a function $f:X\to T$ is called an $\epsilon$-sufficient statistics when the inequality

$$I(X:Y)-\epsilon\le I(T:Y) \tag{73}$$

holds. Hence, a sufficient statistics with $T$ of small size and an $\epsilon$-sufficient statistics can be considered as compressed data of $X$ with respect to $Y$.

In the above example, $X_1X_2$ is a sufficient statistics for $Y$. When $\delta$ is sufficiently small for $\epsilon$, $I(X_1:Y)$ is close to $I(X_1X_2:Y)$, i.e., $X_1$ is an $\epsilon$-sufficient statistics. Hence, we can remove non-essential information $X_2$. In fact, if $\mathcal{X}=\mathcal{X}_1\times\mathcal{X}_2$ is disturbed by a random permutation $\pi$, it will be non-trivial to extract essential information. To cover such a non-trivial case, we need a systematic approach to find such a function with a small-size $T$. For this aim, we can use the information bottleneck algorithm.

To extract approximate sufficient statistics $T$, we focus on two requirements. The mutual information $I(T:Y)$ should be larger, and the entropy $H(T)$ should be smaller. To handle these requirements, we simply minimize $H(T)-\beta I(T:Y)$ by using deterministic information bottleneck algorithm with $|\mathcal{T}|=|\mathcal{X}|$. Since the algorithm minimizes $H(T)-\beta I(T:Y)$, and the conditional distribution $P_{T|X}$ in the solution is deterministic, the support of $P_T$ in the solution is expected to be smaller than the original set $\mathcal{T}$.

## 7.2 Numerics

To demonstrate the above idea, let us take a look at a concrete example, which is a modification of the example in Section 2.5. Consider a single-qubit quantum system $Y$ and a classical register $X$ that encodes information about $Y$. The register $X$ is further split into two sub-registers $X_1$ and $X_2$ that take values in the sets $\mathcal{X}_1=\{0,1,\dots,4\}$ and $\mathcal{X}_2=\{0,1,\dots,19\}$. Then, we assume that $P_X$ is the uniform distribution over $\mathcal{X}_1\times\mathcal{X}_2$, and the density $\rho_{Y|x_1}$ is given as $\rho(\theta_{x_1},\lambda_{x_1})$ with (37). The parameters $\theta$ and $\lambda$ depend on $x_1$ as

$$\theta_{x_1}:=\pi\cdot\frac{x_1}{|\mathcal{X}_1|}\qquad\lambda_{x_1}:=\frac{x_1}{4|\mathcal{X}_1|}. \tag{74}$$

Obviously, the quantum system depends only on $X_1$ and $X_2$ contains no information about the quantum system. An experimentalist who has access to the ensemble, however, does not know this. To extract information about the quantum system, for each pair of $(x_1,x_2)$, the experimentalist estimates its density matrix by repetitively (for $\nu<\infty$ times) making a suitable measurement on $\rho\left(\theta_{x_1},\lambda_{x_1}\right)$. According to quantum state estimation theory [13, 15], the estimate has an inaccuracy proportional to $1/\sqrt{\nu}$. Taking this into account, we model the estimated density matrix as $\rho\left(\theta_{x_1,x_2},\lambda_{x_1,x_2}\right)$ when the actual density matrix is $\rho\left(\theta_{x_1},\lambda_{x_1}\right)$, where

$$\theta_{x_1,x_2}:=\pi\cdot\frac{x_1}{|\mathcal{X}_1|}\left(1+r_\nu(x_1,x_2)\right) \tag{75}$$

$$\lambda_{x_1,x_2}:=\frac{x_1}{4|\mathcal{X}_1|}\left(1+r'_\nu(x_1,x_2)\right) \tag{76}$$

and $r_\nu(x_1,x_2),r'_\nu(x_1,x_2)=O(1/\sqrt{\nu})$ characterise the estimation errors. The estimated ensemble then admits the density matrix given in (72) with $\hat\rho_{Y|x_1,x_2}=\rho\left(\theta_{x_1,x_2},\lambda_{x_1,x_2}\right)$, which is given by Eqs. (37), (75), and (76). Notice that now the register $X_2$ is correlated with $Y$ in the estimated joint state $\hat\rho_{XY}$, even if the estimation-induced noise follows a distribution that does not depend on the value of $X_2$.

Now, the task is to compress the register $X$, by constructing a map from $X$ to a smaller classical register $T$. Here we take $T$ to be the same size as $X$. One intuitive approach is to discard the $X_2$ register because $X_1$ contains much more information about the qubit state than $X_2$. Nevertheless, such a simple map does not exist in more general cases. For instance, if the values of $(x_1,x_2)$ in Eq. (72) are permuted, discarding $X_2$ will not result in faithful compression. To see this, we further apply a arbitrary chosen unknown reshuffling $\pi:\mathcal{X}\to\mathcal{X}$ to the classical register $\mathcal{X}=\mathcal{X}_1\times\mathcal{X}_2$ in Eq. (72). The ensemble then admits the following joint density matrix:

$$\hat\rho'_{XY}=\sum_{x_1,x_2}\Big(P_X(x_1,x_2)|\pi(x_1,x_2)\rangle\langle\pi(x_1,x_2)|\otimes\rho\left(\theta_{x_1,x_2},\lambda_{x_1,x_2}\right)\Big) \tag{77}$$

with $\rho\left(\theta_{x_1,x_2},\lambda_{x_1,x_2}\right)$ given by Eqs. (75) and (76). The goal is to extract an approximate sufficient statistics by constructing a map $Q:\mathcal{X}\to\mathcal{T}$.

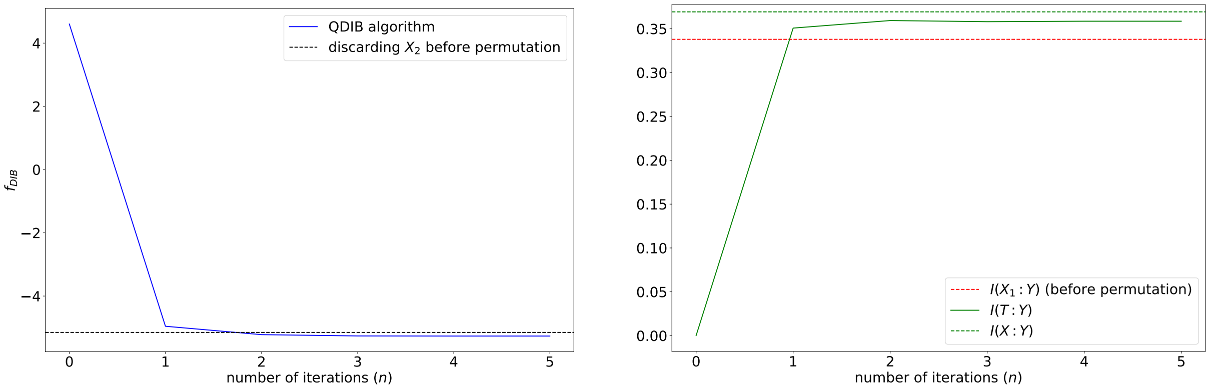

Figure 7: **Performance of QDIB algorithm in constructing approximate sufficient statistics.** We apply our quantum deterministic information bottleneck (QDIB) algorithm on the state (77) (see also Fig. 6). For the joint state, we choose $|\mathcal{X}_1| = 5$ and $|\mathcal{X}_2| = 20$, and $P_X$ to be the uniform distribution over $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. The noise $r_\nu(x_1, x_2)$ and $r'_\nu(x_1, x_2)$ are drawn randomly and uniformly from the interval $(-1/\sqrt{\nu}, 1/\sqrt{\nu})$ with $\nu = 20$ for any $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$. In the QDIB algorithm (Algorithm 3), we choose $\beta = 20$ and $|\mathcal{T}| = |\mathcal{X}| = 100$. In the figure above, the information bottleneck $f_{\mathrm{DIB}} := f_{\alpha \to 0}$ is plotted as a function of the number of iterations. As can be seen from the plot, the QDIB value of our algorithm becomes lower than that of the fictional protocol of "discarding $X_2$ after the inverse permutation $\pi^{-1}$" after only 3 iterations. In the figure below, the faithfulness $I(T : Y)$ is plotted as a function of the number of iterations, and $I(X_1 : Y)$ after the inverse permutation $\pi^{-1}$ (corresponding to the performance of the fictional protocol of "discarding $X_2$ after the inverse permutation $\pi^{-1}$") as well as $I(X : Y)$ (corresponding to the upper bound of $I(T : Y)$) are plotted for reference. Both plots justify that our QDIB algorithm performs well in the task of constructing approximate sufficient statistics.

Our QDIB algorithm works as a more systematic and more efficient method to extract essential information and discard non-essential information, even in the presence of an arbitrary permutation. In the QDIB algorithm (Algorithm 3), we choose $\beta = 20$ and $|\mathcal{T}| = |\mathcal{X}| = |\mathcal{X}_1||\mathcal{X}_2|$. First, we consider the case when the ensemble admits the form (72), and the performance is summarised in Fig. 7. As one can see from the numerics, $f_{\mathrm{DIB}} := f_{\alpha \to 0}$ of applying our QDIB algorithm to $\hat{\rho}_{XY}$ drops lower than that of the "discarding $X_2$ after the inverse permutation $\pi^{-1}$" approach within 5 iterations, and converges to a much lower value, suggesting a better compression performance. This is further justified in the second plot, where the faithfulness $I(T : Y)$ and the residual information $I(T : X)$ are plotted. We can see that since our QDIB algorithm preserves almost as much information about $Y$ as the original variable $X$, it compresses a considerably larger portion of information about the original register $X$.

# 8 Discussion and conclusion

We have proposed a generalized algorithm for QIB with an acceleration parameter $\gamma$ and an additional parameter $\alpha$, and have derived a necessary condition for the monotonic decrease of the objective function $f_\alpha = H(T) - \alpha H(T|X) - \beta I(T : Y)$ with quantum systems $Y, T$ and classical system $X$ when we extract information $T$ with respect to $Y$ from $X$. We have also showed its convergence under the same condition and that a wisely-chosen parameter $\gamma$ can accelerate the convergence. Our numerical calculation has further justified the above analysis as follows. In our numerical experiment, making $\gamma$ smaller accelerates the convergence, but if $\gamma$ is made smaller than a threshold the algorithm will fail to converge. In addition, we have provided examples that quantum system $T$ have an advantage over classical system $T$ even when $Y$ and $X$ are classical.

Next, taking the limit $\alpha \to +0$, we have proposed an iterative algorithm for QDIB that minimizes the objective function $f_{\mathrm{DIB}} = H(T) - \beta I(T : Y)$. We have shown that this iterative algorithm always decreases the objective function monotonically. QDIB can be applied to find an approximate sufficient statistics because it realizes a smaller entropy $H(T)$ and a larger mutual information $I(T : Y)$. Then, we have numerically demonstrated that our QDIB algorithm works well as an approximate sufficient statistics.

An important application we show in this work is that our QIB algorithm yields a new approach of constructing quantum feature maps for classification. In our numerical example, quantum system $T$ realizes a smaller value of the objective function than classical system $T$. This numerical analysis shows the advantage of using quantum memory $T$ for the classification. Despite significant recent progress [3, 5, 11, 17, 20, 25–27, 34], the advantage of quantum machine learning over its classical counterpart has not been discussed much. Our work provides a new angle of attacking this issue, shedding light on a new proposal to rigorously justify and quantify quantum supremacy in the world of learning.

An open question left for future study is how to extend our result to the case where $X$ is also a quantum system, which covers, for instance,the scenario of compressing a quantum system while keeping its correlation with a classical label [21, 23, 35–38]. Remarkably, in such a scenario, it has been shown that, if $T$ is classical, some correlation will be lost regardless of its size [37]. Therefore, we anticipate that the advantage of a quantum $T$ might persist or grow even stronger for QIB with a quantum $X$.

Finally, we remark that currently there is no efficient method to compute the restriction on $\gamma$ in Theorem 3. Resolving this important issue in a future work will accelerate the convergence of our information bottleneck algorithm.

## Acknowledgement

MH is supported in part by the National Natural Science Foundation of China (Grant No. 62171212) and Guangdong Provincial Key Laboratory (Grant No. 2019B121203002). YY is supported by Guangdong Basic and Applied Basic Research Foundation (Project No. 2022A1515010340), and by the Hong Kong Research Grant Council (RGC) through the Early Career Scheme (ECS) grant 27310822.

## References

[1] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972. DOI: 10.1109/TIT.1972.1054753.

[2] Leonardo Banchi, Jason Pereira, and Stefano Pirandola. Generalization in quantum machine learning: A quantum information standpoint. *PRX Quantum*, 2:040321, Nov 2021. DOI: 10.1103/PRXQuantum.2.040321.

[3] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017. DOI: 10.1038/nature23474.

[4] R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972. DOI: 10.1109/TIT.1972.1054855.

[5] Carsten Blank, Daniel K Park, June-Koo Kevin Rhee, and Francesco Petruccione. Quantum classifier with tailored quantum kernel. *npj Quantum Information*, 6(1):1–7, 2020. DOI: 10.1038/s41534-020-0272-6.

[6] Nilanjana Datta, Christoph Hirche, and Andreas Winter. Convexity and operational interpretation of the quantum information bottleneck function. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1157–1161, 2019. DOI: 10.1109/ISIT.2019.8849518.

[7] András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 193–204, 2019. DOI: 10.1145/3313276.3316366.

[8] Ziv Goldfeld and Yury Polyanskiy. The information bottleneck problem and its applications in machine learning. *IEEE Journal on Selected Areas in Information Theory*, 1(1):19–38, 2020. DOI: 10.1109/JSAIT.2020.2991561.

[9] Arne L. Grimsmo and Susanne Still. Quantum predictive filtering. *Phys. Rev. A*, 94:012338, Jul 2016. DOI: 10.1103/PhysRevA.94.012338.

[10] Aram W Harrow, Avinatan Hassidim, and Seth Lloyd. Quantum algorithm for linear systems of equations. *Physical review letters*, 103(15):150502, 2009. DOI: 10.1103/PhysRevLett.103.150502.

[11] Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019. DOI: 10.1038/s41586-019-0980-2.

[12] Masahito Hayashi and Vincent Y. F. Tan. Minimum rates of approximate sufficient statistics. *IEEE Transactions on Information Theory*, 64(2):875–888, 2018. DOI: 10.1109/TIT.2017.2775612.

[13] Carl W Helstrom. Quantum detection and estimation theory. *Journal of Statistical Physics*, 1 (2):231–252, 1969. DOI: 10.1007/BF01007479.

[14] Christoph Hirche and Andreas Winter. An alphabet-size bound for the information bottleneck function. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2383–2388, 2020. DOI: 10.1109/ISIT44484.2020.9174416.

[15] Alexander S Holevo. *Probabilistic and statistical aspects of quantum theory*, volume 1. Springer Science & Business Media, 2011. DOI: 10.1007/978-88-7642-378-9.

[16] Winston H. Hsu, Lyndon S. Kennedy, and Shih-Fu Chang. Video search reranking via information bottleneck principle. MM '06, pages 35–44, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595934472. DOI: 10.1145/1180639.1180654.

[17] Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. Quantum embeddings for machine learning. *arXiv preprint arXiv:2001.03622*, 2020. DOI: 10.48550/arXiv.2001.03622.

[18] Guang Hao Low and Isaac L Chuang. Hamiltonian simulation by uniform spectral amplification. *arXiv preprint arXiv:1707.05391*, 2017. DOI: 10.48550/arXiv.1707.05391.

[19] Guang Hao Low and Isaac L Chuang. Hamiltonian simulation by qubitization. *Quantum*, 3: 163, 2019. DOI: 10.22331/q-2019-07-12-163.

[20] Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4: 226, 2020. DOI: 10.22331/q-2020-02-06-226.

[21] Martin Plesch and Vladimír Bužek. Efficient compression of quantum information. *Physical Review A*, 81(3):032317, 2010. DOI: 10.1103/PhysRevA.81.032317.

[22] Navneeth Ramakrishnan, Raban Iten, Volkher B. Scholz, and Mario Berta. Computing quantum channel capacities. *IEEE Transactions on Information Theory*, 67(2):946–960, 2021. DOI: 10.1109/TIT.2020.3034471.

[23] Lee A Rozema, Dylan H Mahler, Alex Hayat, Peter S Turner, and Aephraim M Steinberg. Quantum data compression of a qubit ensemble. *Physical Review Letters*, 113(16):160504, 2014. DOI: 10.1103/PhysRevLett.113.160504.

[24] Sina Salek, Daniela Cadamuro, Philipp Kammerlander, and Karoline Wiesner. Quantum rate-distortion coding of relevant information. *IEEE Transactions on Information Theory*, 65(4):2603–2613, 2019. DOI: 10.1109/TIT.2018.2878412.

[25] Maria Schuld. Supervised quantum machine learning models are kernel methods. *arXiv preprint arXiv:2101.11020*, 2021. DOI: 10.48550/arXiv.2101.11020.

[26] Maria Schuld and Nathan Killoran. Quantum machine learning in feature Hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019. DOI: 10.1103/PhysRevLett.122.040504.

[27] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185, 2015. DOI: 10.1080/00107514.2014.964942.

[28] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017. DOI: 10.48550/arXiv.1703.00810.

[29] Noam Slonim and Naftali Tishby. Document clustering using word clusters via the information bottleneck method. SIGIR '00, pages 208–215, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132263. DOI: 10.1145/345508.345578.

[30] Maximilian Stark, Aizaz Shah, and Gerhard Bauch. Polar code construction using the information bottleneck method. In *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pages 7–12, 2018. DOI: 10.1109/WCNCW.2018.8368978.

[31] DJ Strouse and David J. Schwab. The Deterministic Information Bottleneck. *Neural Computation*, 29(6):1611–1630, 06 2017. ISSN 0899-7667. DOI: 10.1162/NECO˙a˙00961.

[32] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*, pages 368–377. Univ. Illinois Press, 1999. DOI: 10.48550/arXiv.physics/0004057.

[33] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE information theory workshop (ITW)*, pages 1–5. IEEE, 2015. DOI: 10.1109/ITW.2015.7133169.

[34] Peter Wittek. *Quantum machine learning: what quantum computing means to data mining*. Academic Press, 2014. DOI: 10.1016/C2013-0-19170-2.

[35] Yuxiang Yang, Giulio Chiribella, and Daniel Ebler. Efficient quantum compression for ensembles of identically prepared mixed states. *Physical Review Letters*, 116(8):080501, 2016. DOI: 10.1103/PhysRevLett.116.080501.

[36] Yuxiang Yang, Giulio Chiribella, and Masahito Hayashi. Optimal compression for identically prepared qubit states. *Phys. Rev. Lett.*, 117:090502, Aug 2016. DOI: 10.1103/PhysRevLett.117.090502.

[37] Yuxiang Yang, Ge Bai, Giulio Chiribella, and Masahito Hayashi. Compression for quantum population coding. *IEEE Transactions on Information Theory*, 64(7):4766–4783, 2018. DOI: 10.1109/TIT.2017.2788407.

[38] Yuxiang Yang, Giulio Chiribella, and Masahito Hayashi. Quantum stopwatch: how to store time in a quantum memory. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2213):20170773, 2018. DOI: 10.1098/rspa.2017.0773.