Transfer Learning for Quantum Classifiers: An Information-Theoretic Generalization Analysis

Sharu Theresa Jose and Osvaldo Simeone

Abstract—A key component of a quantum machine learning model operating on classical inputs is the design of an embedding circuit mapping inputs to a quantum state. This paper studies a transfer learning setting in which classical-to-quantum embedding is carried out by an arbitrary parametric quantum circuit that is pre-trained based on data from a source task. At run time, a binary quantum classifier of the embedding is optimized based on data from the target task of interest. The average excess risk, i.e., the optimality gap, of the resulting classifier depends on how (dis)similar the source and target tasks are. We introduce a new measure of (dis)similarity between the binary quantum classification tasks via the trace distances. An upper bound on the optimality gap is derived in terms of the proposed task (dis)similarity measure, two Rényi mutual information terms between classical input and quantum embedding under source and target tasks, as well as a measure of complexity of the combined space of quantum embeddings and classifiers under the source task. The theoretical results are validated on a simple binary classification example.

I. INTRODUCTION

Quantum machine learning (QML) is an emerging paradigm for programming noisy, intermediate scale quantum (NISQ) computers [1]. In QML, the parameter vector $\theta$ defining the operation of a parametric quantum circuit (PQC) is optimized based on quantum or classical data. When input data are classical, it is necessary to design an embedding circuit to map classical inputs to a quantum state [2, Ch. 6]. This is illustrated in Fig. 1, in which the classical input vector $x$ is mapped, via a PQC, to a quantum state defined by a density matrix $\rho_\theta(x)$. We focus on the task of classifying the input $x$ by applying a quantum measurement $\{M_c\}$ to the density matrix $\rho_\theta(x)$ [2], [3]. Following the QML framework, both the parameter vector $\theta$ of the embedding circuit and the classifying quantum measurement are optimized based on supervised examples of the form $(c, x)$, where $c$ is a binary label. This paper analyzes the generalization properties of the trained quantum classifier.

Reference [4] has recently studied the generalization properties of the circuit in Fig. 1 for a fixed embedding parameter $\theta$ as a function of the number $N$ of examples $(c, x)$ used to optimize the measurement from an information-theoretic viewpoint. The authors have shown that the excess risk, i.e., the optimality gap, can be bounded as $O(\sqrt{2I_2(X; R_\theta)/N})$, where $I_2(X; R_\theta) = -\log \sum_c \rho_{c}^\theta |c\rangle \langle c|$, with $\rho_c^\theta$ being the marginal of the ground-truth distribution $p(c, x)$ [5].

In this paper, we consider the more practical case in which one needs to design both embedding parameter vector $\theta$ and classifying measurement $\{M_c\}$. Furthermore, we address the challenging scenario in which limited data is available from the target task. To this end, as in [6], we assume that the embedding circuit producing the state $\rho_\theta(x)$ is pre-trained based on, generally more abundant, data from a related source task. With the pre-trained embedding circuit $\rho_\theta(x)$, data from the target task is used only to adapt the classifying measurement. This corresponds to the quantum-quantum transfer learning scenario in the sense of [6]. Therefore, differently from [4], the average excess risk of the resulting classifier with respect to the target task depends crucially on how similar the source and target tasks are. Other related works that consider quantum-quantum transfer learning include reference [7], in which a PQC pre-trained using data from a source task is transferred to initialize the parameters for a different task so as to mitigate barren plateaus.

Our main contribution is a new measure of dissimilarity between binary quantum classification tasks that allows us to derive an upper bound on the aver-
age excess risk of the binary classifier. The derived bound scales as $O\left(\sqrt{\log \theta} + \frac{1}{\sqrt{N \tau}}\right)$ for $I_T^p$, where $I_T^p$ is the proposed (dis)similarity between the source and target tasks; and $\hat{\mathcal{R}}_{\theta, M}^T$ is the generalization error of the joint space of quantum embeddings and measurements that scales with the dimension of the Hilbert space of the quantum embedding.

Apart from the mentioned reference [4], generalization properties of variational quantum circuits as QML models have been characterized via an information geometric approach based on Fisher information [8]; via the Rademacher complexity of the space of PQC measured in terms of the complexity of the space of PQCs measured in terms of the Rademacher complexity of the joint space of quantum embeddings and measurements considered in this work, revealing the role of task similarity for transfer representation learning. To the best of our knowledge, ours is the first work that studies the generalization error incurred in transfer learning quantum embeddings from an information-theoretic perspective.

The rest of the paper is organized as follows. Sec. II details the two-stage transfer learning problem under study and defines the average excess risk. Sec. III introduces a similarity metric between source and target tasks based on information-theoretic upper bound on the average excess risk. Sec. IV presents an information-theoretic upper bound on the average excess risk. Theoretical conclusions are demonstrated via examples in Sec. V.

II. Problem Formulation

In this section, we first describe the quantum classification problem studied in [4] in which the quantum embedding parameter vector $\theta$ is fixed, and then we present the two-stage transfer learning problem illustrated in Fig. 1 in which the embedding circuit parameter $\theta$ is pre-trained based on data from a separate source task.

A. Quantum Classification with a Fixed Embedding

Let $x$ denote the classical input feature vector and $c \in \{0, 1\}$ be the corresponding binary class index. We take $x$ to assume values in an arbitrary discrete finite set, although extensions to continuous-valued inputs are direct [4]. The data sample $(c, x)$ is generated from an unknown underlying joint distribution $p_T(c, x)$ describing the target task. The embedding circuit maps the classical feature vector $x$ to a density matrix $\rho_0(x)$, i.e., to a positive semi-definite unit-trace matrix defined on some (finite-dimensional) Hilbert space. The embedding circuit is implemented by a PQC parameterized by a (classical) parameter vector $\theta \in \Theta$, where $\Theta$ is an arbitrary set. Note that we do not assume any specific structure for the embedding circuit. As such, the embedding ensembles as special cases one-shot data encoding [2] or data re-uploading [13]. As we will see later in Sec. IV, the structure of the embedding circuit determines the Rényi MI terms as well as the Rademacher complexity $\hat{\mathcal{R}}_{\theta, M}^T$ of the joint space of embeddings and measurements.

The classifier consists of a positive operator-valued measure (POVM) applied to the quantum state $\rho_0(x)$. The POVM is defined by two positive-semi-definite matrices $M = \{M_c\}_{c=0}^{1}$, of the same dimensions as the density matrix $\rho_0(x)$, that satisfy the conditions $M_c \geq 0$ and $\sum_{c=0}^{1} M_c = I$. By Born’s rule, the classifier chooses class $c$ with probability $\operatorname{Tr}(M_c \rho_0(x))$, where $\operatorname{Tr}(\cdot)$ represents the trace operation. We use $\mathcal{M} = \{M : M_c \geq 0, \sum_{c=0}^{1} M_c = I\}$ to denote the set of all binary POVMs for the given Hilbert space.

For a fixed embedding parameter $\theta$, quantum supervised classification [4] optimizes the POVM $M \in \mathcal{M}$ with the ideal goal of minimizing the expected probability of error, also known as the expected risk, i.e.,

$$\hat{\mathcal{R}}_{\mathcal{M}}^T = \mathbb{E}_{p_T(c, x)}[\ell_{\theta, M}(c, x)],$$

(1)

over $M \in \mathcal{M}$, where

$$\ell_{\theta, M}(c, x) = 1 - \operatorname{Tr}(M_c \rho_0(x))$$

(2)

is the probability of error evaluated on an example $(c, x)$. Accordingly, the minimum expected risk for parameter $\theta$ is given as

$$\mathcal{R}_\theta^T = \min_{M \in \mathcal{M}} \hat{\mathcal{R}}_{\theta, M}^T.$$

(3)

Since the ground-truth joint distribution $p_T(c, x)$ is unknown, the optimization of the POVM $M$ is done by using a training data set $D_T = \{(c_1, x_1), \ldots, (c_N, x_N)\}$ of $N_T$ samples, whose individual data points $(c_j, x_j)$ are assumed to be independent identically distributed (i.i.d.) according to $p_T(c, x)$. Specifically, the POVM is obtained by minimizing the training loss

$$\hat{\mathcal{R}}_{\theta, M}^T = \frac{1}{N_T} \sum_{(c, x) \in D_T} \ell_{\theta, M}(c, x).$$

(4)

The solution of this optimization can be obtained in closed form, yielding the so-called Helstrom measurement (see [3, Sec. III]). We write as

$$\hat{M}_\theta^T = \arg \min_{M \in \mathcal{M}} \hat{\mathcal{R}}_{\theta, M}^T$$

(5)

the optimal POVM and the corresponding minimized training loss for a fixed $\theta$, respectively.

The classifier obtained with the POVM (5) is considered to generalize well if it yields a low expected risk (2). In this regard, a key metric of interest is the excess risk

$$\Delta \mathcal{R}_\theta = \mathcal{R}_\theta^T - \hat{\mathcal{R}}_{\theta, \hat{M}_\theta^T}^T,$$

(6)

which is the difference between the expected risk (1) obtained via the outlined learning process and the genie-aided expected risk obtained with the optimal POVM. As described in Section I, an information-theoretic bound on the excess risk (6) was derived in [4] for a fixed parameter $\theta$. 533
B. Transfer Learning for Quantum Classification

In this work, as illustrated in Fig. 1, we consider a two-stage transfer learning problem, in which the embedding parameter vector \( \theta \) is pre-trained based on data from a source task with underlying true data distribution \( p^S(c, x) \), which is generally different from the distribution \( p^T(c, x) \) of the target task. To this end, we assume to have access to a training set \( D^S = \{(c_1, x_1), \ldots, (c_{NS}, x_{NS})\} \) of \( NS \) samples generated i.i.d. according to the source task distribution \( p^S(c, x) \). In a typical implementation, one uses source-task data to compensate for limitations in the availability of target-task data. Therefore, one may assume that the number of data samples \( N^S \) from the source task is larger than that for the target task, i.e., \( N^S \gg N^T \).

As illustrated in Fig. 1, in the pre-training phase, the source-task data set \( D^S \) is used to optimize the embedding parameter \( \theta \). In the training phase, the embedding parameter is fixed to the pre-trained value \( \hat{\theta} \) obtained from the first phase, and the POVM for the target task is optimized as described in the previous subsection.

To elaborate, in the pre-training phase, the embedding parameter vector \( \theta \) is obtained by minimizing the training loss on the source-task data as

\[
\hat{\theta} = \arg \min_{\theta \in \Theta} \hat{R}^S_{\theta},
\]

where we have defined the source-task training loss as \( \hat{R}^S_{\theta} = \arg \min_{M \in M} \hat{R}^S_{\theta, M} \) with \( \hat{R}^S_{\theta, M} = \sum_{(c, x) \in D^S} \ell_{\theta, M}(c, x)/N^S \).

In the training phase, the classification is maximized as in (5) using the target-task data for the pre-trained embedding parameter vector \( \hat{\theta} \). This yields the POVM \( \hat{M}^T_{\hat{\theta}} \) and the expected risk \( \hat{R}^T_{\hat{\theta}} = \hat{R}^T_{\hat{\theta}, \hat{M}^T_{\hat{\theta}}} \).

In order to evaluate the generalization properties of transfer learning, we adopt the transfer excess risk

\[
\Delta R^{S \rightarrow T} = R^T_{\hat{\theta}, \hat{M}^T_{\hat{\theta}}} - \min_{\theta \in \Theta} \hat{R}^T_{\theta},
\]

Unlike the excess risk in (6), the transfer excess risk captures the impact on generalization not only of the classifier, which is trained using target-task data, but also of the embedding parameter \( \theta \), which is pre-trained using source-task data. The transfer excess risk (8) thus depends intuitively on how “similar” the embedding parameter vectors \( \theta \) that minimize the losses on the source and target tasks are.

III. ON THE SIMILARITY OF SOURCE AND TARGET TASKS

In this section, we present a similarity metric for source and target tasks that will be shown in the next section to determine a bound on the transfer excess risk (8). To this end, we start with some preliminary background on quantum information.

A. Preliminaries

Let \( \rho \) and \( \sigma \) denote two square matrices defined on the same (finite-dimensional) Hilbert space. The trace distance \( T(\rho, \sigma) \) between the matrices \( \rho \) and \( \sigma \) is defined as [5]

\[
T(\rho, \sigma) = \frac{1}{2} \| \rho - \sigma \|_1,
\]

where \( \| A \|_1 = \text{Tr}(\sqrt{A^\dagger A}) \) is the trace norm of the matrix \( A \), with \( A^\dagger \) denoting the conjugate transpose of \( A \). The trace distance satisfies triangle inequality, and for two density matrices \( \rho \) and \( \sigma \), it is bounded as \( 0 \leq T(\rho, \sigma) \leq 1 \) [5].

B. Task-Induced Distance between Embedding Parameters

We start by defining a distance measure \( d^A(\theta, \theta') \) between two embedding parameter vectors \( \theta \) and \( \theta' \) in \( \Theta \) induced by a task \( A \in \{S, T\} \). The distance \( d^A(\theta, \theta') \) is given by the difference between the minimum expected risks (3) obtained with embedding parameters \( \theta \) and \( \theta' \).

**Definition 3.1 (Task-based Distance):** For task \( A \), with \( A \in \{S, T\} \), the task-based distance between any two embedding parameters \( \theta \) and \( \theta' \in \Theta \) is defined as

\[
d^A(\theta, \theta') = |R^A_{\theta} - R^A_{\theta'}|,
\]

where the minimum expected risk is defined in (3). We have the inequalities \( 0 \leq d^A(\theta, \theta') \leq 0.5 \).

The task-based distance (10) can be computed explicitly by introducing the class-c average density matrix

\[
\rho^A_{\theta}(c) = \mathbb{E}_{p^A_S(x)}[p^A_S(x)]
\]

for \( c \in \{0, 1\} \) and task \( A \in \{S, T\} \). As mentioned in Sec. II-A, the expected risk (3) for task \( A \) is minimized by the Helstrom POVM, and the resulting minimal expected risk can be obtained in closed form as [5, Ex. 9.1.7],

\[
R^A_{\theta} = \frac{1}{2} - T(p^A_{\theta}((0)) \rho^A_{\theta}(0), p^A_{\theta}(1) \rho^A_{\theta}(1))
\]

**C. Measure of Similarity between Source and Target Tasks**

Of particular interest is the task-based distance \( d^A(\theta, \theta^A) \) between any embedding parameter \( \theta \in \Theta \) and the embedding parameter \( \theta^A \) that minimizes the expected risk (3) for task \( A \), i.e., \( \theta^A = \arg \min_{\theta \in \Theta} R^A_{\theta} \), for \( A \in \{S, T\} \). This distance measures the sub-optimality of the parameter \( \theta \) with respect to the optimal embedding parameter \( \theta^A \) for task \( A \). This is because, by Definition 3.1, a small distance \( d^A(\theta, \theta^A) \) implies that the expected risks with embedding parameters \( \theta \) and \( \theta^A \) are close. Using this idea, and inspired by [12, Def. 3], we introduce the following definition of task dissimilarity.

**Definition 3.2:** Tasks \( T \) and \( S \) are \( D^{ST}-\text{dissimilar} \) if we have the inequality

\[
d^T(\theta, \theta^T) \leq d^S(\theta, \theta^S) + D^{ST}
\]

for all embedding parameters \( \theta \in \Theta \).

Hence, the two tasks are \( D^{ST}-\text{dissimilar} \) if the suboptimality of each embedding parameter \( \theta \) on the target task differs from the suboptimality for the source task by no more than a scalar constant \( D^{ST} \). Intuitively, a small constant \( D^{ST} \) should result in a positive transfer of information from source to target task during pre-training.

The following theorem provides two explicit task dissimilarity measures satisfying (13) (see [14] for proof). To this
end, we define $TV(p, q) = 0.5 \sum_{x \in X} |p(x) - q(x)|$ as the total variation distance between discrete distributions $p$ and $q$.

**Theorem 3.1:** For source task, with data distribution $p^S(c, x)$, and target task, with data distribution $p^T(c, x)$, the following quantities

$$D^{ST}_{\text{trace}} = 2 \sup_{\theta \in \Theta} |R^S_{\theta} - R^T_{\theta}|,$$

and

$$\sum_{x \in X} p^S(c, x) - \sum_{x \in X} p^T(c, x)$$

satisfy the inequality (13). Furthermore, we have the inequality $D^{ST}_{\text{trace}} \leq D^{ST}_{TV}$.

**IV. AN INFORMATION-THEORETIC BOUND ON THE TRANSFER EXCESS RISK**

In this section, we leverage the measure of dissimilarity $D^{ST}$ between source and target tasks introduced in Section III-C to obtain an information-theoretic upper bound on the transfer excess risk (8). Detailed proofs of all results can be found in [14].

To start, consider the bipartite quantum system comprising of the classical register $X$ reporting the value of input vector $x$ and the quantum register $R_\theta$ corresponding to the embedding $\rho_\theta(x)$. For a given embedding parameter vector $\theta$, the classical-quantum state of the above bipartite system for task $A \in \{S, T\}$ is described by the density matrix $\rho^A_{X,R_\theta} = \sum_{x} |x\rangle\langle x| \otimes \rho_\theta(x)$, where $\otimes$ is the Kronecker product. $p^A(x)$ is the relevant marginal of the joint distribution $p^A(c, x)$, and $\{x\}$ is an orthonormal basis for the Hilbert space of register $X$, which has dimension equal to the number of possible values for $x$ (see, e.g., [5]). Then, the 2-Rényi mutual information (MI) between the subsystems $X$ and $R_\theta$ is defined as [4]

$$I^2_A(X; R_\theta) = 2 \log_2 \text{Tr} \left( \sum_x p^A(x) \rho_\theta(x)^2 \right).$$

We also define the Rademacher complexity of the space $M$ of POVM measurements as

$$\mathbb{R}^A_M = \sup_{\theta \in \Theta} \mathbb{E}_{p^A(c, x), p^{(\sigma)}} \left[ \sup_{M \in M} \sum_{j=1}^N \frac{\sigma_j \ell_M \theta(c_j, x_j)}{\sqrt{N}} \right],$$

and the joint Rademacher complexity of the space $\Theta$ of embedding parameters and of the space $M$ as

$$\mathbb{R}^A_{\Theta, M} = \mathbb{E}_{p^A(c, x), p^{(\sigma)}} \left[ \sup_{\theta \in \Theta, M \in M} \sum_{j=1}^N \frac{\sigma_j \ell_M \theta(c_j, x_j)}{\sqrt{N}} \right],$$

where the expectation is taken over i.i.d. variables $(c, x) \sim p^A(c, x)$ and over i.i.d. zero-mean and equiprobable Rademacher variables $\sigma = (\sigma_1, \ldots, \sigma_N) \sim p(\sigma)$ with $\sigma_j \in \{+1, -1\}$. The following lemma presents an upper bound on the Rademacher complexity measures (17)-(18).

**Lemma 4.1:** Assume that the embedding circuit defines quantum states of the form $\rho_\theta(x) = U(\theta, x)|0\rangle\langle 0| U(\theta, x)^\dagger$, where $U(\theta, x) = \prod_{l=1}^L U_l(\theta_l) S_l(x)$ consists of parameterized unitary gates $U_l(\theta_l)$ as well as encoding gates $S_l(x)$ [2]. The Rademacher complexity measures (17) and (18) can be upper bounded as

$$\mathbb{R}^A_M \leq \mathbb{R}^A_{\Theta, M} \leq n (c(p^A(x))),$$

where $n$ is the dimension of the Hilbert space, and $c(p^A(x)) \leq 1$ is a constant that depends on the marginal distribution $p^A(x)$.

**A. Upper Bound on Target-Task Excess Risk With No Source-Task Data**

We first consider a baseline scenario when no data from the source task is available. Only data from the target task is used to jointly optimize quantum embedding $\rho_\theta(x)$ and measurement $M$. In this case, the excess risk (6) for the target task evaluates as $R_{\theta, T}$, where $\theta = \arg \min_{\theta \in \Theta} \min_{M \in M} \mathbb{R}^T_M$. The following theorem presents an upper bound on the excess risk.

**Theorem 4.1:** The following upper bound on the excess risk holds with probability at least $1 - \delta$, for $\delta \in (0, 1)$, with respect to the i.i.d. random draws of data set $D^T$ from the joint distribution $p^T(c, x)$

$$\Delta R^T_{\theta, T} \leq 2 \left( \mathbb{R}^T_M + \mathbb{R}^T_{\Theta, M} \right) + \sqrt{2 \frac{2}{N^T} \log \frac{3}{\delta}},$$

where $\mathbb{R}^T_M$ is bounded as

$$\mathbb{R}^T_M \leq 0.5 \sqrt{\sup_{\theta \in \Theta} I^2_T(X; R_\theta)},$$

with $I^2_T(X; R_\theta)$ denoting the 2-Rényi MI in (16).

The upper bound in (20) shows that, in the absence of source-task data, the sample complexity scales (at most) proportionally to the sum $\mathbb{R}^T_M + \mathbb{R}^T_{\Theta, M}$.

**B. Upper Bound on Transfer Excess Risk**

We now present an upper bound on the transfer excess risk (8) for the case in which source-task data is available.

**Theorem 4.2:** For any constant $D^{ST}$ satisfying (13), the following upper bound on the transfer excess risk holds with probability at least $1 - \delta$, for $\delta \in (0, 1)$, with respect to the i.i.d. random draws of data sets $D^T$ and $D^S$ from the respective joint distributions $p^T(c, x)$ and $p^S(c, x)$

$$\Delta R^{S\rightarrow T} \leq \frac{2 \mathbb{R}^S_M}{\sqrt{N^S}} + \sqrt{2 \frac{2}{N^T} \log \frac{3}{\delta}} + D^{ST}$$

$$+ \frac{2 \mathbb{R}^T_M + \mathbb{R}^T_{\Theta, M}}{\sqrt{N^T}} + \sqrt{2 \frac{2}{N^S} \log \frac{3}{\delta}},$$

where $\mathbb{R}^A_M$, for $A \in \{S, T\}$, is bounded as in (21).

The bound (22) illustrates the advantage of transfer learning in reducing the sample complexity for the target task. In fact, if abundant data is available from the source task (i.e., if $N^S \rightarrow \infty$) and if the source and target tasks are sufficiently similar so that $D^{ST}$ is small, the sample complexity of the target task scales proportional to $\mathbb{R}^T_M$. By inequality (19), this is smaller than the scaling $2 (\mathbb{R}^T_M + \mathbb{R}^T_{\Theta, M})$ obtained in Theorem 4.1 when no source-task data is available.
In this section, we consider a source task and a target task with equiprobable class label $c \in \{0, 1\}$. For each class $c \in \{0, 1\}$, we obtain the discrete-valued input $x$ by finely quantizing a continuous-valued feature input $\hat{x} \in \mathbb{R}$ so that the discrete sum in (16) can be evaluated via numerical integration [4]. For the source task, the feature $\hat{x}$ is Gaussian distributed as $\mathcal{N}(\hat{x} | \mu_c^S, \sigma^2)$ with mean $\mu_c^S \in \mathbb{R}$ and variance $\sigma^2$; while, for the target task, we have the per-class Gaussian distribution $\mathcal{N}(x | \mu_c^T, \sigma^2)$ with mean $\mu_c^T \in \mathbb{R}$, generally different from that of source task, and the same variance $\sigma^2$.

The embedding circuit maps the classical input $x$ to the rank-1 density matrix $\rho_0(x) = |x\rangle\langle x|$, with the pure quantum state $|x\rangle$ given as

$$|x\rangle = U_\theta(x)|0\rangle, \quad \text{with} \quad U_\theta(x) = R_X(x)R_{\theta_0}R_X(x),$$

where $U_\theta(x)$ is a unitary matrix parameterized by the angles $\theta = (\theta_1, \theta_2, \theta_3) \in [0, 2\pi]^3$, which constitutes the embedding PQC (see, e.g., [2]). The operation of the embedding circuit involves the Pauli-X rotation $R_X(x)$ defined as in [2, Eq. (3.45)], and the general rotation $R_{\theta_0}$ defined as in [2, Eq. (3.48)].

In Figure 2, we plot the transfer excess risk $\Delta R_{S \rightarrow T}$ (top figure), along with the corresponding upper bound derived in (22) (bottom figure) as a function of the number of target task samples $N^T$ for varying values of source task samples $N^S$. Note that $N^S = 0$ corresponds to the excess risk $\Delta R_{S \rightarrow T}^T$ and the upper bound (20). Other parameters are set as $\delta = 0.5$, $\sigma^2 = 0.11$, $\mu_0^S = 1$, $\mu_1^S = -1$, $\mu_0^T = 1.5$, and $\mu_1^T = -0.5$. The transfer excess risk $\Delta R_{S \rightarrow T}$ is a random variable, which is evaluated by drawing multiple pairs of data sets $(D^S, D^T)$ from their respective joint distributions $p^S(c, x)$ and $p^T(c, x)$. The thick lines in the top figure correspond to the median of the resulting empirical distribution, while the shaded areas represent its spread.

The figure shows that the upper bound (22), while numerically loose (as is common for related information-theoretic bounds in classical machine learning [15], [16]), predicts well the regime where transfer learning is advantageous. Comparing the case when no source task is available (i.e., $N^S = 0$) to when abundant source-task data is available for transfer learning (i.e., $N^S = 100$), Fig 2 shows that transfer learning can achieve a smaller excess risk when limited data are available from the target task ($N^T < 8$). This advantage vanishes when target task-data become increasingly available, in which case the contribution of the task dissimilarity measure $D_{ST}$ to transfer excess risk outweighs the other terms in (20). This also explains the non-vanishing behaviour of transfer excess risk in Fig 2 in the limit as $N^S$ and $N^T$ grow large.

The impact of the dissimilarity between the two tasks is further elaborated on in Figure 3, which illustrates the transfer excess risk $\Delta R_{S \rightarrow T}$ (top) and the corresponding upper bound in (22) (bottom) as a function of the difference $\mu_c^T - \mu_c^S$ between the means of the input data under the target and source tasks for both classes $c \in \{0, 1\}$. We fix $\mu_0^S = 1$ and $\mu_1^S = -2$, $N^T = 4$, $N^S = 10$, $\sigma^2 = 1$, and $\delta = 0.9$. As can be seen, when the difference between the means is zero, i.e., when the source and target tasks coincide, the transfer excess risk, and the corresponding upper bound, are minimized.

**REFERENCES**

[1] O. Simeone, “An introduction to quantum machine learning for engineers,” Foundations and Trends® in Signal Processing, vol. 16, no. 1-2, pp. 1–223, 2022.

[2] M. Schuld and F. Petruccione, Machine Learning with Quantum Computers. Springer Nature, 2021.

[3] C. W. Helstrom, “Quantum detection and estimation theory,” Journal of Statistical Physics, vol. 1, no. 2, pp. 231–252, 1969.

[4] L. Banchi, J. Pereira, and S. Pirandola, “Generalization in quantum machine learning: A quantum information standpoint,” Quantum, vol. 2, no. 4, p. 040321, 2021.

[5] M. M. Wilde, Quantum information theory. Cambridge University Press, 2013.

[6] A. Mari, T. R. Bromley, J. Izaac, M. Schuld, and N. Killoran, “Transfer learning in hybrid classical-quantum neural networks,” Quantum, vol. 4, p. 340, 2020.

[7] H.-Y. Liu, T.-P. Sun, Y.-C. Wu, Y.-J. Han, and G.-P. Guo, “Mitigating barren plateaus with transfer-learning-inspired parameter initializations,” New Journal of Physics, vol. 25, no. 1, p. 013039, 2023.

[8] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, “The power of quantum neural networks,” Nature Computational Science, vol. 1, no. 6, pp. 403–409, 2021.

[9] M. C. Caro, H.-Y. Huang, M. Cerezo, K. Sharma, A. Sornborger, L. Cincio, and P. J. Coles, “Generalization in quantum machine learning from few training data,” arXiv preprint arXiv:2111.05292, 2021.
[10] Y. Du, Z. Tu, X. Yuan, and D. Tao, “Efficient measure for the expressivity of variational quantum algorithms,” *Physical Review Letters*, vol. 128, no. 8, p. 080506, 2022.

[11] M. C. Caro, E. Gil-Fuster, J. J. Meyer, J. Eisert, and R. Sweke, “Encoding-dependent generalization bounds for parametrized quantum circuits,” *Quantum*, vol. 5, p. 582, 2021.

[12] N. Tripuraneni, M. I. Jordan, and C. Jin, “On the theory of transfer learning: The importance of task diversity,” *arXiv preprint arXiv:2006.11650*, 2020.

[13] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, “Data re-uploading for a universal quantum classifier,” *Quantum*, vol. 4, p. 226, Feb. 2020. [Online]. Available: https://doi.org/10.22331/q-2020-02-06-226

[14] S. T. Jose and O. Simeone, “Transfer learning for quantum classifiers: An information-theoretic generalization analysis,” *arXiv preprint arXiv:2201.06297*, 2022.

[15] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, “Information-theoretic analysis for transfer learning,” in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2819–2824.

[16] S. T. Jose and O. Simeone, “Information-theoretic bounds on transfer generalization gap based on jensen-shannon divergence,” in *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 1461–1465.