Quantum information theory (MAT4430) **Spring 2021**

Lecture 11: The classical capacity of a quantum channel

Lecturer: Alexander Müller-Hermes

In the last lecture, we discussed Holevo's theorem quantifying how well classical information can be encoded into a quantum system. Holevo's theorem puts a bound on the capacity of classical channels obtained from preparing a quantum state $x \mapsto \sigma_x \in D(\mathcal{H})$ for each $x \in \Sigma_A$ followed by a measurement $\mu : \Sigma_B \to B(\mathcal{H})^+$. The resulting classical channel $N : \Sigma_A \to \mathcal{P}(\Sigma_B)$ is given by

$$N(y|x) = \langle \mu(y), \sigma_x \rangle.$$

To send classical information via a quantum channel $T : B(\mathcal{H}_A) \to B(\mathcal{H}_B)$ we could choose the following strategy motivated by Holevo's theorem. First, we choose some quantum states $\{\rho_x\}_{x\in\Sigma_A} \subset D(\mathcal{H}_A)$, i.e., we choose a quantum state for every symbol of the classical input alphabet. Sending these quantum states through the channel T yields the quantum states $\sigma_x = T(\rho_x) \in D(\mathcal{H}_B)$ on the output system. Finally, we can choose a measurement $\mu : \Sigma_B \to B(\mathcal{H}_B)^+$ on the output quantum system. In this way, we would obtain a classical channel $N_{\{\rho_x\},\mu,T} : \Sigma_A \to \mathcal{P}(\Sigma_B)$ given by

$$N_{\{\rho_x\},\mu,T}(y|x) = \langle \mu(y), T(\rho_x) \rangle,$$

and we could use this channel to transmit classical information (see Figure 1).

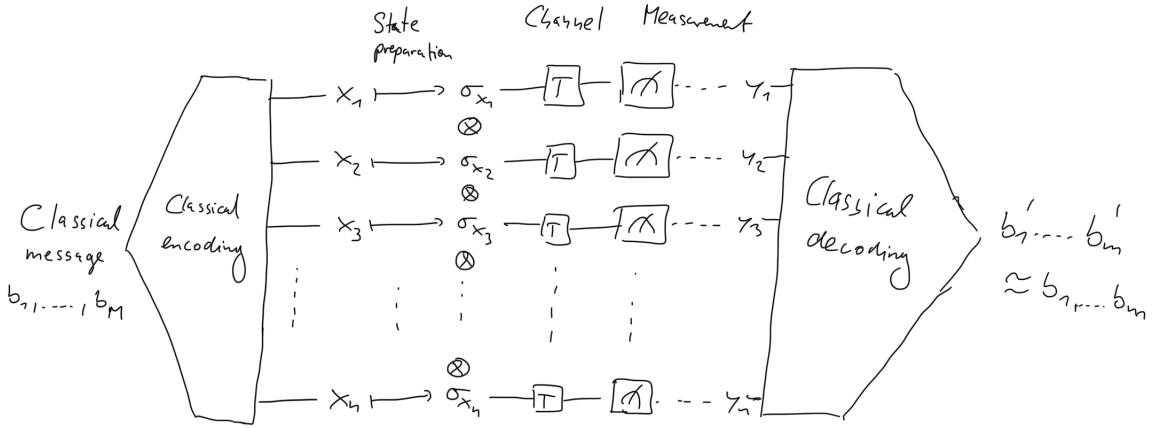

Figure 1: Naive approach to construct a coding scheme.

Optimizing the capacity $C(N_{\{\rho_x\},\mu,T})$ over all choices of $\{\rho_x\}_{x\in\Sigma_A} \subset D(\mathcal{H}_A)$ and measurements $\mu : \Sigma_B \to B(\mathcal{H})^+$ gives a lower bound on the classical capacity of the quantum channel T (which we will define precisely below). However, this bound could be pretty bad (spoiler: it is!) since it does not take into account entanglement. We could expect to get higher communication rates by using a general measurement $\mu : \Sigma_B \to B(\mathcal{H}_B^{\otimes n})^+$, i.e., measuring n systems together, instead of measuring each output system separately. Moreover, we could apply the above idea to the channel $T^{\otimes k}$ instead of T thereby allowing entangled quantum states to be inserted into many instances of the quantum channel. We will see that such strategies indeed lead to higher communication rates, and in general they can be used to achieve rates arbitrarily close to the classical capacity $C(T)$.

1 Definition of the classical capacity

Let us start with the definition of coding schemes for classical information transmission:

Definition 1.1 (Coding schemes)**.** *An* (n, m, δ)-coding scheme for classical information transmission over a quantum channel $T : B(\mathcal{H}_A) \to B(\mathcal{H}_B)$ *is given by an encoding map*

$$E : \{0,1\}^m \to D(\mathcal{H}_A^{\otimes n}),$$

and a measurement

$$\mu : \{0,1\}^m \to B(\mathcal{H}_B)^{\otimes n},$$

such that

$$\langle \mu(b_1, \ldots, b_m), T^{\otimes n} \circ E(b_1, \ldots, b_m) \rangle_{HS} \geqslant 1 - \delta,$$

for all $b_1, \ldots, b_m \in \{0,1\}$.

As always, having defined coding schemes leads to the definition of achievable rates and the classical capacity:

Definition 1.2 (Achievable rates and classical capacity)**.** *A rate* $R \geqslant 0$ *is called* achievable for classical communication over the quantum channel $T : B(\mathcal{H}_A) \to B(\mathcal{H}_B)$ *if either* $R = 0$ *or* $R > 0$ *and for any* $n \in \mathbb{N}$ *there exists an* (n, m_n, δ_n)*-coding scheme for classical information transmission over* T *such that*

$$R = \lim_{n\to\infty} \frac{m_n}{n} \quad \text{and} \quad \lim_{n\to\infty} \delta_n = 0.$$

We define the classical capacity *of the channel* T *to be*

$$C(T) = \sup\{R \geqslant 0 \text{ achievable rate for classical communication over } T\}.$$

The main result of this and the next lecture will be the quantum analogue of Shannon's capacity theorem. To state it, we need to define the Holveo information of a quantum channel:

Definition 1.3 (Holevo information of a quantum channel)**.** *For any quantum channel* $T : B(\mathcal{H}_A) \to B(\mathcal{H}_B)$ *we define*

$$\chi(T) = \sup_{\{p(x),\rho_x\}_{x\in\Sigma}} \chi\left(\{p(x), T(\rho_x)\}_{x\in\Sigma}\right),$$

where the supremum is over all ensembles $\{p(x), \rho_x\}_{x\in\Sigma}$ *with states in* $\rho_x \in D(\mathcal{H}_A)$ *and* $p \in \mathcal{P}(\Sigma)$ *and any alphabets* Σ.

We will show in the exercises that the supremum in the definition of $\chi(T)$ is actually achieved, and in general we may restrict the optimization in the previous definition to alphabets of the size $|\Sigma| \leqslant d_A^2$. Now, we can state the main theorem in this lecture:

Theorem 1.4 (Holevo-Schumacher-Westmoreland)**.** *For any quantum channel* $T : B(\mathcal{H}_A) \to B(\mathcal{H}_B)$ *we have*

$$C(T) = \lim_{k\to\infty} \frac{1}{k} \chi\left(T^{\otimes k}\right).$$

We need to make two comments about the previous theorem. First, it should be noted that unlike in the classical case the capacity $C(T)$ is not given by a so-called *single-letter formula* involving a single copy of the quantum channel T. Instead it is given by a *regularization* of the Holevo quantity, which involves many copies of T. In general, it is unclear how to compute such a formula. To this date, it is not known whether there is a simpler

expression for $C(T)$, and it is not even known if the capacity $C(T)$ is a Turing-computable function.

The second comment is about the converse of the previous theorem. We have seen similar theorems before, e.g., the compression theorem, where the converse was in the form that the probability of failure for a protocol goes to 1 if we try to exceed achievable rates. However, in the HSW-theorem we only claim that the failure probability does not vanish in the limit of $n \to \infty$. This is called a so-called *weak converse*, where the former is called a *strong converse*. To this date it is not known whether the strong-converse holds in the HSW-theorem.

In the next section, we will lay out the mathematical machinery needed for the proof of the HSW theorem.

2 Conditional typicality

We will need the following notion of *conditional typicality*:

Definition 2.1. *Let $p_{AB} \in \mathcal{P}(\Sigma_A \times \Sigma_B)$ denote a joint probability distribution and $p_A \in \mathcal{P}(\Sigma_A)$ the marginal distribution defined by*

$$p_A(x) = \sum_{y \in \Sigma_B} p_{AB}(x, y).$$

For every $\epsilon > 0$, every $n \in \mathbb{N}$ and every string $(x_1, \ldots, x_n) \in \Sigma_A^n$ such that

$$p_A(x_1) \cdots p_A(x_n) > 0,$$

*we say that a string $(y_1, \ldots, y_n) \in \Sigma_B^n$ is ϵ-*typical conditioned on *$(y_1, \ldots, y_n)$ if*

$$2^{-n(H(p_{AB}-H(p_A)+\epsilon)} < \frac{p_{AB}(x_1, y_1) \cdots p(x_n, x_n)}{p_A(x_1) \cdots p_A(x_n)} < 2^{-n(H(p_{AB}-H(p_A)-\epsilon)}.$$

We denote the set of ϵ-typical strings conditioned on $(x_1, \ldots, x_n) \in \Sigma_A^n$ by $\mathcal{T}_{n,\epsilon}(p_{AB}|x_1, \ldots, x_n)$ and for convenience we set $\mathcal{T}_{n,\epsilon}(p_{AB}|x_1, \ldots, x_n) = \emptyset$ if $p_A(x_1) \cdots p_A(x_n) = 0$.

The following lemma summarizes some properties of the conditional typical strings:

Lemma 2.2. *For any joint probability distribution $p_{AB} \in \mathcal{P}(\Sigma_A \times \Sigma_B)$ we have the following:*

1. *For all $\epsilon > 0$ we have*

$$\lim_{n\to\infty} \sum_{x_1,\ldots,x_n \in \Sigma_A^n} \sum_{y_1,\ldots,y_n \in \mathcal{T}_{n,\epsilon}(p_{AB}|x_1,\ldots,x_n)} p_{AB}(x_1, y_1) \cdots p_{AB}(x_n, y_n) = 1.$$

2. *For all $n \in \mathbb{N}$ and all $\epsilon > 0$ we have*

$$\sum_{x_1,\ldots,x_n \in \Sigma_A^n} p_A(x_1) \cdots p_A(x_n) |\mathcal{T}_{n,\epsilon}(p_{AB}|x_1, \ldots, x_n)| < 2^{n(H(p_{AB})-H(p_A)+\epsilon)}.$$

Proof. Exercises. □

Next, we need to define the notion of conditional typicality for ensembles of quantum states. For this consider an ensemble $\{p_A(x), \rho_x\}_{x \in \Sigma_A}$ and consider the spectral decomposition

$$\rho_x = \sum_{y \in \Sigma_B} p_B(y|x) |v_y^x\rangle\langle v_y^x|,$$

where we introduced a conditional probability distribution $p_B(\cdot|x) \in \mathcal{P}(\Sigma_B)$, for $\Sigma_B = \{1, \ldots, \dim(\mathcal{H})\}$, for each $x \in \Sigma_A$. We may define a joint probability distribution $p_{AB} \in \mathcal{P}(\Sigma_A \times \Sigma_B)$ by $p_{AB}(x, y) = p_A(x)p_B(y|x)$ for any $x \in \Sigma_A$ and any $y \in \Sigma_B$. With this, we state the following definition:

Definition 2.3. *Consider an ensemble* $\{p_A(x), \rho_x\}_{x\in\Sigma_A}$ *such that*

$$\rho_x = \sum_{y\in\Sigma_B} p_B(y|x)|v_y^x\rangle\langle v_y^x|,$$

for each $x \in \Sigma_A$*, and let* $p_{AB} \in \mathcal{P}(\Sigma_A \times \Sigma_B)$ *for* $\Sigma_B = \{1, \ldots, \dim(\mathcal{H})\}$ *be defined as above. For any* $\epsilon > 0$*, any* $n \in \mathbb{N}$ *and any string* $x_1, \ldots, x_n \in \Sigma_A^n$ *we define the* projection onto the ϵ-typical subspace conditioned on $x_1, \ldots, x_n$ *by*

$$\Lambda_{x_1,\ldots,x_n,\epsilon} = \sum_{\substack{y_1,\ldots,y_n \\ \in\mathcal{T}_{n,\epsilon}(p_{AB}|x_1,\ldots,x_n)}} |v_{y_1}^{x_1}\rangle\langle v_{y_1}^{x_1}| \otimes \cdots \otimes |v_{y_n}^{x_n}\rangle\langle v_{y_n}^{x_n}|.$$

The next lemma follows immediately from Lemma 2.2:

Lemma 2.4. *For any ensemble* $\{p_A(x), \rho_x\}_{x\in\Sigma_A}$ *we have the following:*

1. *For all* $\epsilon > 0$ *we have*

$$\lim_{n\to\infty} \sum_{x_1,\ldots,x_n\in\Sigma_A^n} p_A(x_1)\cdots p_A(x_n)\langle\Lambda_{x_1,\ldots,x_n,\epsilon}, \rho_{x_1}\otimes\cdots\otimes\rho_{x_n}\rangle = 1.$$

2. *For all* $n \in \mathbb{N}$ *and all* $\epsilon > 0$ *we have*

$$\sum_{x_1,\ldots,x_n\in\Sigma_A^n} p_A(x_1)\cdots p_A(x_n)\operatorname{Tr}\left[\Lambda_{x_1,\ldots,x_n,\epsilon}\right] < 2^{n\left(\sum_{x\in\Sigma_A} p_A(x)H(\rho_x)+\epsilon\right)}.$$

3 Achievable rates of product codes

Now, we will show how allowing for global measurements while restricting to product codes allows for achievable rates arbitrary close to the upper bound given by Holevo's theorem. Specifically, we will show the following:

Theorem 3.1 (Performance of product codes)**.** *Let* Σ *denote a finite alphabet,* $\mathcal{H}$ *a complex Euclidean space and*

$$\{\sigma_x \ : \ x \in \Sigma\} \subset D(\mathcal{H}),$$

a subset of quantum states. If $R < \chi\left(\{p(x), \sigma_x\}_{x\in\Sigma}\right)$ *for some* $p \in \mathcal{P}(\Sigma)$*, then for each* $n \in \mathbb{N}$ *there exists a function* $f_n : \{0,1\}^{m_n} \to \Sigma^n$ *where* $m_n = \lfloor Rn \rfloor$ *and a POVM* $\mu_n : \{0,1\}^{m_n} \to B(\mathcal{H}^{\otimes n})^+$ *such that*

$$\min_{b_1,\ldots,b_{m_n}\in\{0,1\}^{m_n}} \langle\mu(b_1,\ldots,b_{m_n}), \sigma_{f(b_1,\ldots,b_{m_n})}\rangle_{HS} \to 1,$$

as $n \to \infty$*. Here, we write* $\sigma_{x_1\cdots x_n} = \sigma_{x_1}\otimes\cdots\otimes\sigma_{x_n}$*.*

Proof. Fix some $p \in \mathcal{P}(\Sigma)$ and $R < \chi\left(\{p(x), \sigma_x\}_{x\in\Sigma}\right)$. Choose $\epsilon > 0$ such that

$$R < \chi\left(\{p(x), \sigma_x\}_{x\in\Sigma}\right) - 3\epsilon.$$

In the following, we will first fix $n, m \in \mathbb{N}$ and only in the end of the proof will we choose $m_n = \lfloor Rn \rfloor$. For $n, m \in \mathbb{N}$ we define the following objects:

- For each $x_1, \ldots, x_n \in \Sigma^n$ we denote by $\Lambda_{x_1,\ldots,x_n}$ the projection onto the ϵ-typical subspace conditioned on $x_1, \ldots, x_n$ with respect to the ensemble $\{p(x), \sigma_x\}_{x\in\Sigma}$.
- We denote by Π_n the projection onto the ϵ-typical subspace of $\mathcal{H}^{\otimes n}$ with respect to the average state $\sigma = \sum_{x\in\Sigma} p(x)\sigma_x$.

- For any function $g : \{0,1\}^{m+1} \to \Sigma^n$ we define

$$Q = \sum_{b_1,\ldots,b_{m+1} \in \{0,1\}^{m+1}} \Pi_n \Lambda_{g(b_1,\ldots,b_{m+1})} \Pi_n.$$

  and we define

$$Q_{b_1,\ldots,b_{m+1}} = Q^{-\frac{1}{2}} \Pi_n \Lambda_{g(b_1,\ldots,b_{m+1})} \Pi_n Q^{-\frac{1}{2}}.$$

  Note that $Q_{b_1,\ldots,b_{m+1}} \geqslant 0$ and

$$\sum_{b_1,\ldots,b_{m+1} \in \{0,1\}^{m+1}} Q_{b_1,\ldots,b_{m+1}} = \Pi_{\mathrm{Im}(Q)}.$$

- Finally, we define a POVM $\mu : \{0,1\}^{m+1} \to B(\mathcal{H}^{\otimes n})^+$ by

$$\mu(b_1,\ldots,b_{m+1}) = Q_{b_1,\ldots,b_{m+1}} + \frac{1}{2^{m+1}} \left(\mathbb{1} - \Pi_{\mathrm{Im}(Q)}\right),$$

  which also depends on the function $g : \{0,1\}^{m+1} \to \Sigma^n$.

We will now analyze the average probability of error when using the coding scheme given by the function g and the POVM μ from above. This quantity is given by

$$\overline{p}_{err}\left(g\right) = \frac{1}{2^{m+1}} \sum_{b_1,\ldots,b_{m+1} \in \{0,1\}^{m+1}} \langle \mathbb{1} - \mu(b_1,\ldots,b_{m+1}), \sigma_{g(b_1,\ldots,b_{m+1})} \rangle_{HS}.$$

First, we apply the Hayashi-Nagaoka inequality from the exercises to show that

$$\begin{aligned} \mathbb{1} - Q_{b_1,\ldots,b_{m+1}} &= \mathbb{1}_{\mathcal{H}} - Q^{-\frac{1}{2}} \Pi_n \Lambda_{g(b_1,\ldots,b_{m+1})} \Pi_n Q^{-\frac{1}{2}} \\ &\leqslant 2\left(\mathbb{1} - \Pi_n \Lambda_{g(b_1,\ldots,b_{m+1})} \Pi_n\right) + 4\left(Q - \Pi_n \Lambda_{g(b_1,\ldots,b_{m+1})} \Pi_n\right), \end{aligned}$$

for each $b_1,\ldots,b_{m+1} \in \{0,1\}^{m+1}$. This shows that

$$\overline{p}_{err}\left(g\right) \leqslant \frac{2}{2^{m+1}} \sum_{b_1,\ldots,b_{m+1} \in \{0,1\}^{m+1}} \langle \mathbb{1} - \Pi_n \Lambda_{g(b_1,\ldots,b_{m+1})} \Pi_n, \sigma_{g(b_1,\ldots,b_{m+1})} \rangle_{HS} \tag{1}$$

$$+ \frac{4}{2^{m+1}} \sum_{b_1,\ldots,b_{m+1} \in \{0,1\}^{m+1}} \langle Q - \Pi_n \Lambda_{g(b_1,\ldots,b_{m+1})} \Pi_n, \sigma_{g(b_1,\ldots,b_{m+1})} \rangle_{HS}. \tag{2}$$

We will now use the probabilistic method to show that for every $n \in \mathbb{N}$ and with $m_n = \lfloor Rn \rfloor$ there exists a function $g_n : \{0,1\}^{m+1} \to \Sigma^n$ such that $\overline{p}_{err}\left(g_n\right) \to 0$ as $n \to \infty$. For this we select functions g at random as follows: For each $b_1,\ldots,b_{m+1}$ we select $x_1,\ldots,x_n \in \Sigma^n$ i. i. d. at random according to the distribution $p^{\times n}$, i.e., we select $x_1,\ldots,x_n \in \Sigma^n$ as the value of $g(b_1,\ldots,b_{m+1})$ with probability $p(x_1)\cdots p(x_n)$. With this random model, we will estimate the expectation value $\mathbb{E}\left[\overline{p}_{err}\left(g\right)\right]$, and since the expectation value is linear, we can analyze the expectation values of the two sums in (1) and (2) separately.

Let us first consider the sum in (1). It is easy to verify the operator identity

$$ABA = AB + BA - B + (\mathbb{1} - A)B(\mathbb{1} - A),$$

for any $A, B \in B(\mathcal{H})$. For any fixed $x_1,\ldots,x_n \in \Sigma^n$ we have

$$\begin{aligned} \langle \Pi_n \Lambda_{x_1,\ldots,x_n} \Pi_n, \sigma_{x_1,\ldots,x_n} \rangle_{HS} &= \langle \Pi_n \Lambda_{x_1,\ldots,x_n}, \sigma_{x_1,\ldots,x_n} \rangle_{HS} + \langle \Lambda_{x_1,\ldots,x_n} \Pi_n, \sigma_{x_1,\ldots,x_n} \rangle_{HS} \\ &\quad - \langle \Lambda_{x_1,\ldots,x_n}, \sigma_{x_1,\ldots,x_n} \rangle_{HS} \\ &\quad + \langle \left(\mathbb{1} - \Pi_n\right) \Lambda_{x_1,\ldots,x_n} \left(\mathbb{1}_{\mathcal{H}} - \Pi_n\right), \sigma_{x_1,\ldots,x_n} \rangle_{HS} \\ &\geqslant \langle \Pi_n \Lambda_{x_1,\ldots,x_n}, \sigma_{x_1,\ldots,x_n} \rangle_{HS} + \langle \Lambda_{x_1,\ldots,x_n} \Pi_n, \sigma_{x_1,\ldots,x_n} \rangle_{HS} \\ &\quad - \langle \Lambda_{x_1,\ldots,x_n}, \sigma_{x_1,\ldots,x_n} \rangle_{HS} \\ &= \langle 2\Pi_n - \mathbb{1}_{\mathcal{H}}, \Lambda_{x_1,\ldots,x_n} \sigma_{x_1,\ldots,x_n} \rangle_{HS}, \end{aligned}$$

where we used in the last step that $[\Lambda_{x_1,\dots,x_n}, \sigma_{x_1,\dots,x_n}] = 0$. Now, observe that

$$\begin{aligned}
\langle \Pi_n \Lambda_{x_1,\dots,x_n} \Pi_n, \sigma_{x_1,\dots,x_n} \rangle_{HS} &\geqslant \langle 2\Pi_n - \mathbb{1}, \Lambda_{x_1,\dots,x_n} \sigma_{x_1,\dots,x_n} \rangle_{HS} \\
&= \langle 2\Pi_n - \mathbb{1}, \sigma_{x_1,\dots,x_n} \rangle_{HS} + \langle \mathbb{1} - 2\Pi_n, (\mathbb{1} - \Lambda_{x_1,\dots,x_n}) \sigma_{x_1,\dots,x_n} \rangle_{HS} \\
&\geqslant \langle 2\Pi_n - \mathbb{1}, \sigma_{x_1,\dots,x_n} \rangle_{HS} - \langle \mathbb{1} - \Lambda_{x_1,\dots,x_n}, \sigma_{x_1,\dots,x_n} \rangle_{HS} \\
&= 2\langle \Pi_n, \sigma_{x_1,\dots,x_n} \rangle_{HS} + \langle \Lambda_{x_1,\dots,x_n}, \sigma_{x_1,\dots,x_n} \rangle_{HS} - 2,
\end{aligned}$$

where we used that $\mathbb{1} - 2\Pi_n \geqslant -\mathbb{1}$. Applying these estimates to the expectation value of the term (1) for a randomly chosen function g (as described above) shows that

$$\begin{aligned}
&\mathbb{E}\left[\frac{2}{2^{m+1}} \sum_{b_1,\dots,b_{m+1} \in \{0,1\}^{m+1}} \langle \mathbb{1} - \Pi_n \Lambda_{g(b_1,\dots,b_{m+1})} \Pi_n, \sigma_{g(b_1,\dots,b_{m+1})} \rangle_{HS} \right] \\
&\quad = \frac{2}{2^{m+1}} \sum_{b_1,\dots,b_{m+1} \in \{0,1\}^{m+1}} \mathbb{E}\left[\langle \mathbb{1} - \Pi_n \Lambda_{g(b_1,\dots,b_{m+1})} \Pi_n, \sigma_{g(b_1,\dots,b_{m+1})} \rangle_{HS} \right] \\
&\quad = 1 - \sum_{x_1,\dots,x_n \in \Sigma^n} p(x_1) \cdots p(x_n) \langle \Pi_n \Lambda_{g(b_1,\dots,b_{m+1})} \Pi_n, \sigma_{g(b_1,\dots,b_{m+1})} \rangle_{HS} \\
&\quad \leqslant 3 - 2\langle \Pi_n, \sigma^{\otimes n} \rangle_{HS} - \sum_{x_1,\dots,x_n \in \Sigma^n} p(x_1) \cdots p(x_n) \langle \Lambda_{x_1,\dots,x_n}, \sigma_{x_1,\dots,x_n} \rangle_{HS} \\
&\quad \longrightarrow 0,
\end{aligned}$$

as $n \to \infty$ by Lemma 2.4 and the properties of the typical projections Π_n.

Next, we will analyze the expression in (2). Note that

$$Q - \Pi_n \Lambda_{g(b_1,\dots,b_{m+1}) \Pi_n} = \sum_{\substack{c_1,\dots,c_{m+1} \in \{0,1\}^{m+1} \\ c_1,\dots,c_{m+1} \neq b_1,\dots,b_{m+1}}} \Pi_n \Lambda_{g(c_1,\dots,c_{m+1})} \Pi_n.$$

Furthermore, we note that by the construction of the random function g (see above) we have

$$\begin{aligned}
&\mathbb{E}\left[\langle \Pi_n \Lambda_{g(c_1,\dots,c_{m+1})} \Pi_n, \sigma_{g(b_1,\dots,b_{m+1})} \rangle_{HS} \right] \\
&\quad = \sum_{\substack{x_1,\dots,x_n \in \Sigma^n \\ x'_1,\dots,x'_n \in \Sigma^n}} p(x_1) \cdots p(x_n) p(x'_1) \cdots p(x'_n) \langle \Lambda_{x_1,\dots,x_n}, \Pi_n \sigma_{x'_1,\dots,x'_n} \Pi_n \rangle_{HS} \\
&\quad = \sum_{x_1,\dots,x_n \in \Sigma^n} p(x_1) \cdots p(x_n) \langle \Lambda_{x_1,\dots,x_n}, \Pi_n \sigma^{\otimes n} \Pi_n \rangle_{HS},
\end{aligned}$$

for any $c_1, \dots, c_{m+1}, b_1, \dots, b_{m+1} \in \{0,1\}$. With this, we find that

$$\begin{aligned}
&\mathbb{E}\left[\langle Q - \Pi_n \Lambda_{g(b_1,\dots,b_{m+1})} \Pi_n, \sigma_{g(b_1,\dots,b_{m+1})} \rangle_{HS} \right] \\
&\quad = \sum_{\substack{c_1,\dots,c_{m+1} \in \{0,1\}^{m+1} \\ c_1,\dots,c_{m+1} \neq b_1,\dots,b_{m+1}}} \mathbb{E}\left[\langle \Pi_n \Lambda_{g(c_1,\dots,c_{m+1})} \Pi_n, \sigma_{g(b_1,\dots,b_{m+1})} \rangle_{HS} \right] \\
&\quad = (2^{m+1} - 1) \sum_{x_1,\dots,x_n \in \Sigma^n} p(x_1) \cdots p(x_n) \langle \Lambda_{x_1,\dots,x_n}, \Pi_n \sigma^{\otimes n} \Pi_n \rangle_{HS},
\end{aligned}$$

Recall from an exercise that

$$\Pi_n \sigma^{\otimes n} \Pi_n \leqslant 2^{-n(H(\sigma) - \epsilon)} \mathbb{1},$$

and therefore we have

$$\begin{aligned}
&\mathbb{E}\left[\langle Q - \Pi_n \Lambda_{g(b_1,\ldots,b_{m+1})} \Pi_n, \sigma_{g(b_1,\ldots,b_{m+1})} \rangle_{HS}\right] \\
&\quad \leqslant (2^{m+1}-1) 2^{-n(H(\sigma)-\epsilon)} \sum_{x_1,\ldots,x_n \in \Sigma^n} p(x_1)\cdots p(x_n) \operatorname{Tr}\left[\Lambda_{x_1,\ldots,x_n}\right] \\
&\quad \leqslant (2^{m+1}-1) 2^{-n(H(\sigma)-\epsilon)} 2^{n(\sum_x p(x) H(\sigma_x))} \\
&\quad \leqslant 2^{m+1-n(\chi(\{p(x),\sigma_x\})-2\epsilon)}.
\end{aligned}$$

With $m_n = \lfloor Rn \rfloor$ such that $R < \chi\left(\{p(x), \sigma_x\}_{x\in\Sigma}\right) - 3\epsilon$ we find that

$$\begin{aligned}
&\mathbb{E}\left[\frac{4}{2^{m+1}} \sum_{b_1,\ldots,b_{m+1}\in\{0,1\}^{m+1}} \langle Q - \Pi_n \Lambda_{g(b_1,\ldots,b_{m+1})} \Pi_n, \sigma_{g(b_1,\ldots,b_{m+1})} \rangle_{HS}\right] \\
&\leqslant 2^{m_n - n(\chi(\{p(x),\sigma_x\})-2\epsilon)+3} \longrightarrow 0,
\end{aligned}$$

as $n \to \infty$.

To finish the proof, we need to combine the previous observations. For every $n \in \mathbb{N}$ we denote

$$\delta_n = \mathbb{E}\left[\overline{p}_{err}\left(g\right)\right] \to 0,$$

as $n \to \infty$, by the previous computations. For every $n \in \mathbb{N}$, we can now argue as follows: Since $\mathbb{E}\left[\overline{p}_{err}\left(g\right)\right] = \delta_n$, there exists a (non-random) function $g_n : \{0,1\}^{m_n+1} \to \Sigma^n$ with $m_n = \lfloor Rn \rfloor$ such that

$$\overline{p}_{err}\left(g_n\right) \leqslant \delta_n.$$

Now, we define the set

$$B_n = \{b_1, \ldots, b_{m_n+1} \in \{0,1\}^{m_n+1} \ : \ \langle \mathbb{1} - \mu(b_1, \ldots, b_{m_n+1}), \sigma_{g(b_1,\ldots,b_{m_n+1})} \rangle_{HS} > 2\delta_n\}.$$

Since $\overline{p}_{err}\left(g_n\right) \leqslant \delta_n$, we have that

$$\frac{2\delta_n |B_n|}{2^{m_n+1}} \leqslant \delta_n,$$

and we conclude that $|B_n| \leqslant 2^{m_n}$. Therefore, there exists an injective function $h_n : \{0,1\}^{m_n} \to \{0,1\}^{m_n+1} \setminus B_n$ and we can define $f_n = g_n \circ h_n$. For this choice, we can verify that

$$\min_{b_1,\ldots,b_{m_n}\in\{0,1\}^{m_n}} \langle \mu(h_n(b_1,\ldots,b_{m_n})), \sigma_{f_n(b_1,\ldots,b_{m_n})} \rangle_{HS} \geqslant 1 - 2\delta_n \to 1,$$

as $n \to \infty$. □