Quantum Data Compression and Quantum Cross Entropy

Zhou Shangnan[a][*]

[a]*Stanford Institute for Theoretical Physics,*
Stanford University, Stanford, CA 94305, USA

Abstract

Quantum machine learning is an emerging field at the intersection of machine learning and quantum computing. A central quantity for the theoretical foundation of quantum machine learning is the quantum cross entropy. In this paper, we present one operational interpretation of this quantity, that the quantum cross entropy is the compression rate for sub-optimal quantum source coding. To do so, we give a simple, universal quantum data compression protocol, which is developed based on quantum generalization of variable-length coding, as well as quantum strong typicality. Moreover, since data compression can be viewed as a machine learning task, quantum cross entropy also serves as a loss function, whose minimum is the von Neumann entropy. This is consistent with the result that von Neumann entropy is the optimal compression rate.

1 Introduction

Machine learning has attracted interests from various fields as a powerful tool for finding patterns in data. With the advancement of quantum information science and technology, there is increasing interest in developing machine learning algorithms that are suitable for quantum data and quantum computers [1]. Despite many attempts in designing quantum machine learning architectures [2, 3, 4, 5], there are many unanswered questions and unresolved issues, partly due to a lack of theoretical foundation. A central concept of classical machine learning is the classical cross entropy, and its quantum generalization, the quantum cross entropy [6], is a building block of the theoretical foundation of quantum machine learning.

One interesting and important question on quantum cross entropy is its operational meaning from an information-theoretic perspective. Classical cross entropy $H(p,q) = -\sum_i p_i \log q_i$ is a measure of the compression rate when we mistakenly use probability distribution q for source coding, instead of the true distribution p. To demonstrate that this

---

[*]snzhou@stanford.edu.

holds in the quantum case as well, we develop a protocol that can do lossless data compression when we don't have perfect knowledge of the quantum information source. Moreover, suppose that the true source is ρ, and we carry out the protocol under the assumption that the source is σ, the compression rate is the quantum cross entropy $S(\rho, \sigma) = -\operatorname{tr}(\rho \log \sigma)$. When our knowledge of the source is perfect, which means $\rho = \sigma$, the compression rate agrees with the optimal compression rate $S(\rho) = -\operatorname{tr}(\rho \log \rho)$. Furthermore, since data compression extracts the key features of the original source, it can be seen as a machine learning task. In this case, the quantum cross entropy serves as a loss function, whose minimum is the von Neumann entropy $S(\rho)$. This is consistent with the result that von Neumann entropy is the optimal compression rate.

There have been mainly two approaches in doing quantum data compression with perfect information of the source: one is to only encode quantum states that are in the typical subspace [7, 8], the other is to do quantum indeterminate-length coding [9], a quantum version of classical variable-length coding. Other related works include [10, 11, 12]. Our protocol is inspired by a combination of these two ideas, in which we define the length observable [9], and project $\rho^{\otimes N}$ to the subspace where the expectation length of the states are typical, which is an extension of the former approach.

In section 2, we review the definition of classical and quantum information source, as well as classical and quantum typicality, with a focus on strong typicality. In section 3, we show a simple protocol that does lossless quantum data compression even when our knowledge of the quantum source is wrong, and unveil the physical meaning of the quantum cross entropy.

2 Information source and strong typicality

There are different models in defining an information source. We start from one simple but fruitful model for a classical information source [13]. The source emits a letter from a finite alphabet $\mathcal{I}$ with D letters at each single use. We assume that different uses of the source are independent and identically distributed. A possible output from N consecutive uses is a sequence $i^N = i_1, i_2, ..., i_n, ..., i_N$ sampled from N random variables $I_1, I_2, ..., I_n, ..., I_N$. We denote the probability of emitting letter i on any given use of the source as $P(i) = p_i$. Typically, the frequency of occurrence of any given letter i in a sequence output is close to p_i. To formalize this intuition, we first define the empirical probability mass function of i^N (also referred to as its type) as

$$\pi(i|i^N) = \frac{|\{n : i_n = i\}|}{N}, \text{ for } i \in \mathcal{I}. \tag{1}$$

For example, if $i^N = (0, 1, 1, 0, 0, 1, 0)$, then $\pi(i|i^N) = \frac{4}{7}$ for $i = 0$, and $\pi(i|i^N) = \frac{3}{7}$ for $i = 1$.

When N is large, by the law of large numbers, for each $i \in \mathcal{I}$,

$$\pi(i|i^N) \rightarrow P(i) \text{ in probability.} \tag{2}$$

We can then define the set of ϵ-strong-typical N-sequences i^N (or the strong typical set in short) as

$$\mathcal{T}_\epsilon^{(N)}(I) = \{i^N : |\pi(i|i^N) - P(i)| \leq \epsilon P(i) \text{ for all } i \in \mathcal{I}\}. \tag{3}$$

Another useful definition is the set of ϵ-weak-typical N-sequences i^N (or the weak typical set in short), which is

$$\mathcal{U}_\epsilon^{(N)}(I) = \left\{i^N : \left|\frac{1}{N} \log \frac{1}{P(i^N)} - H(p)\right| \leq \epsilon\right\}, \tag{4}$$

where $H(p) = -\sum_{i \in \mathcal{I}} P(i) \log P(i)$ is the Shannon entropy, and $P(i^N)$ is the probability that a certain sequence i^N occurs.

A sequence that is ϵ-strong-typical is definitely ϵ-weak-typical, while the reverse doesn't always hold. One useful property is the unit probability property [13], which holds for both strong and weak typical sequences.

Unit Probability Theorem. Given $\epsilon > 0$. For any $\delta > 0$, when N is sufficiently large,

$$P\big(i^N \in \mathcal{T}_\epsilon^{(N)}(I)\big) \geq 1 - \delta, \tag{5}$$

and

$$P\big(i^N \in \mathcal{U}_\epsilon^{(N)}(I)\big) \geq 1 - \delta. \tag{6}$$

This means that as N approaches infinity, the probability that a given sequence i^N is typical approaches one.

Now we move on to the quantum case. The definition of a quantum information source [8] we use here is based on the idea that entanglement is what we are trying to compress and decompress. Formally, an identical, independently distributed (i.i.d) quantum source is described by a Hilbert space H and a density matrix ρ on that Hilbert space, represented by (ρ, H). We can view the state ρ as part of a larger system which is in a pure state, and the mixed nature of ρ is due to the entanglement between H and the remainder of the system. At each use, a quantum source emits a quantum state that is on average ρ. After

3

N consecutive uses, the average output is $\rho^{\otimes N}$. We now develop a quantum version of the strong typicality.

Suppose the density matrix ρ can be decomposed as

$$\rho = \sum_{i \in \mathcal{I}} P(i)|i\rangle\langle i|, \tag{7}$$

where the $|i\rangle$'s form an orthonormal set, and $P(i)$'s are eigenvalues of ρ, which obey the same rules as a probability distribution. An ϵ-strong-typical product state is a state $|i_1\rangle|i_2\rangle \cdots |i_N\rangle$ where $i^N = i_1, i_2, ..., i_N$ forms a (classical) ϵ-strong-typical sequence.

We define the ϵ-strong-typical subspace $T(N, \rho, \epsilon)$ as the subspace spanned by all ϵ-strong-typical product states. These product states form a basis of $T(N, \rho, \epsilon)$. The projector $Q(N, \rho, \epsilon)$ onto the subspace $T(N, \rho, \epsilon)$ is

$$Q(N, \rho, \epsilon) = \sum_{i^N \text{ is } \epsilon\text{-strong-typical}} |i_1\rangle\langle i_1| \otimes |i_2\rangle\langle i_2| \otimes \cdots \otimes |i_N\rangle\langle i_N|. \tag{8}$$

By generalizing the properties of strong-typical sequences to the quantum form, we have the strong-typical subspace theorems:

Unit Probability Theorem. Given $\epsilon > 0$. For any $\delta > 0$, when N is sufficiently large,

$$\text{tr}\left(Q(N, \rho, \epsilon)\rho^{\otimes N}\right) \geq 1 - \delta. \tag{9}$$

Proof.

$$\begin{aligned}
\text{tr}\left(Q(N, \rho, \epsilon)\rho^{\otimes N}\right) &= \sum_{i^N \text{ is } \epsilon\text{-strong-typical}} P(i_1)P(i_2)\cdots P(i_N) \\
&= \sum_{i^N \text{ is } \epsilon\text{-strong-typical}} P(i^N).
\end{aligned} \tag{10}$$

When i^N is ϵ-strong-typical, it is also ϵ-weak-typical, and the result follows from the unit probability theorem of weak typicality.

3 Quantum data compression with wrong source

Now we present a lossless quantum data compression protocol that works even when our knowledge of the information source is wrong. We show that in this non-ideal scenario, the compression rate is the quantum cross entropy, and the fidelity approaches one as N approaches infinity.

Suppose we develop our compression and decompression protocol with the belief that our quantum source is described by density matrix σ_0 and Hilbert space H, despite that in reality, the quantum source is (ρ_0, H). Usually, this misinformation includes mismatches on both eigenvalues and eigenbasis:

$$\sigma_0 = \sum_{i=1}^{D} q_i |a_i\rangle\langle a_i|, \ \rho_0 = \sum_{i=1}^{D} p_i |b_i\rangle\langle b_i|, \ \{q_i\} \neq \{p_i\}, \ \{|a_i\rangle\} \neq \{|b_i\rangle\}. \tag{11}$$

Unlike the case when we have perfect knowledge of the source [7, 8], direct projection to the typical subspace of σ_0 doesn't work, because the overlap between typical subspaces $T(N, \rho_0, \epsilon)$ and $T(N, \sigma_0, \epsilon)$ becomes empty when N becomes large and ϵ stays small. Also, the typical subspace $T(N, \sigma_0, \epsilon)$ has dimension $2^{NS(\sigma_0)}$, which suggests a compression rate of $S(\sigma_0)$. When $S(\rho_0) > S(\sigma_0)$, this implies a compression rate below the optimal lossless compression rate for the true state ρ_0, meaning a failure in preserving fidelity.

3.1 Revisit the classical case

We turn to the classical case to find inspirations. Suppose our classical information source emits the i-th letter with probability p_i, instead of the wrong, perceived q_i. One simple way of source coding is to assign the i-th letter with a codeword of length $l_i = \log \frac{1}{q_i}$. The expectation length $\langle l \rangle$ of a single codeword is

$$\langle l \rangle = \sum_{i=1}^{D} p_i l_i = \sum_{i=1}^{D} p_i \log \frac{1}{q_i} = H(p, q). \tag{12}$$

Hence, the compression rate is the classical cross entropy $H(p, q)$.

In practice, we can only assign integer length of codewords. One simple way is to let $l_i = \lceil \log \frac{1}{q_i} \rceil$. By the properties of ceiling functions,

$$H(p, q) \leq \langle l \rangle < H(p, q) + 1. \tag{13}$$

This idea is called the variable-length coding, which means that we assign shorter codewords to letters with higher probability of occurrence.

3.2 A simple quantum protocol

The quantum generalization can be tricky, since we have to deal with superposition of basis states, making the lengths of codes indeterminate [9, 14].

Our protocol starts from some preparation work. We first treat q_i's as some probability distribution which represents a classical source, and assign codeword C_i with length $l_i = \log \frac{1}{q_i}$ to the i-th letter. In practice, we can only deal with integer number of qubits, so the precise version is $l_i = \lceil \log \frac{1}{q_i} \rceil$. For simplicity, we will carry out our discussion without worrying about this subtlety, and take it into account in the end. As we believe that the quantum source is (σ_0, H), we construct each unit of computational basis $|i\rangle$ by assigning the first l_i available qubits to $|C_i\rangle$. To keep track of a codeword's length, we define the length observable $L = \sum_{i=1}^{D} l_i |i\rangle\langle i|$. When dealing with N copies of the source state, the computational basis we use is $\{|i_1\rangle|i_2\rangle \cdots |i_N\rangle\}$.

We then do a unitary evolution $U = \sum_{i=1}^{D} |i\rangle\langle a_i|$ to map the true source state ρ_0 to the computational basis $|i\rangle$. In the new basis, we have

$$\rho = U\rho_0 U^\dagger = \sum_{j,k} \langle a_j|\rho_0|a_k\rangle |j\rangle\langle k|. \tag{14}$$

For simplicity, let $r_{jk} = \langle a_j|\rho_0|a_k\rangle$, $r_j = r_{jj}$.

The expectation length of a single codeword is

$$\langle l \rangle = \text{tr}(\rho L) = \sum_i \langle a_i|\rho_0|a_i\rangle l_i = \sum_i \langle a_i|\rho_0|a_i\rangle \log \frac{1}{q_i} = -\sum_i r_i \log q_i. \tag{15}$$

By the definition of the quantum cross entropy, we have

$$S(\rho_0, \sigma_0) = -\text{tr}(\rho_0 \log \sigma_0) = -\sum_i \langle a_i|\rho_0|a_i\rangle \log\langle a_i|\sigma_0|a_i\rangle = -\sum_i r_i \log q_i = \langle l \rangle. \tag{16}$$

From (16), we can see that r_i's give the "true" probability distribution in the "wrong" basis, which relates quantum cross entropy and classical cross entropy:

$$S(\rho_0, \sigma_0) = H(r, q), \quad r_i = \langle a_i|\rho_0|a_i\rangle, \quad q_i = \langle a_i|\sigma_0|a_i\rangle. \tag{17}$$

Here, r and q are probability distributions viewed in the orthonormal basis of σ_0.

When the quantum source emits N copies, the state we need to compress is

$$\rho^{\otimes N} = \left(\sum_{j_1, k_1} r_{j_1 k_1} |j_1\rangle\langle k_1| \right) \otimes \left(\sum_{j_2, k_2} r_{j_2 k_2} |j_2\rangle\langle k_2| \right) \otimes \cdots \otimes \left(\sum_{j_N, k_N} r_{j_N k_N} |j_N\rangle\langle k_N| \right). \tag{18}$$

The total length l_{total} of the codewords is just an addition of each codeword, and the total length observable Λ can be defined as $\Lambda = L_1 + L_2 + \cdots + L_N$. For a basis state

$|i^N\rangle = |i_1\rangle|i_2\rangle \cdots |i_N\rangle$, $l_{total} = \sum_{n=1}^{N} \log \frac{1}{q_{i_n}}$. The expectation length of N codewords is $\langle l_{total} \rangle = N\langle l \rangle = NS(\rho_0, \sigma_0)$. We now show that the first $NS(\rho_0, \sigma_0)$ qubits contains all the information of $\rho^{\otimes N}$ as N goes to infinity.

Fix $\epsilon > 0$. We define a projector Π:

$$\Pi = \sum_{\text{length condition}} |i_1\rangle\langle i_1| \otimes |i_2\rangle\langle i_2| \otimes \cdots \otimes |i_N\rangle\langle i_N|, \tag{19}$$

where the length condition for $i^N = i_1, i_2, ..., i_N$ is

$$\left| \frac{1}{N} \sum_{n=1}^{N} \log \frac{1}{q_{i_n}} - S(\rho_0, \sigma_0) \right| \leq \epsilon. \tag{20}$$

When i^N is ϵ-strongly typical, define i.i.d random variables $I_1, I_2, ..., I_N$ such that $I_n = \log \frac{1}{q_{i_n}}$. The expectation value is $E(I) = \sum_i r_i \log \frac{1}{q_i}$. For any $\delta > 0$, when N is sufficiently large, by the law of large numbers,

$$P\left(\left| \frac{1}{N} \sum_{n=1}^{N} I_n - E(I) \right| \leq \epsilon \right) = P\left(\left| \frac{1}{N} \sum_{n=1}^{N} \log \frac{1}{q_{i_n}} - S(\rho_0, \sigma_0) \right| \leq \epsilon \right) \geq 1 - \delta, \tag{21}$$

which fulfills the length condition. Hence, $Q(N, \rho, \epsilon) \leq \Pi$, and

$$\text{tr}(\Pi \rho^{\otimes N}) \geq \text{tr}\left(Q(N, \rho, \epsilon) \rho^{\otimes N} \right) \geq 1 - \delta. \tag{22}$$

We apply Π to project $\rho^{\otimes N}$ onto the subspace where the total codeword length $l_{total} \in [NS(\rho_0, \sigma_0) - \epsilon, NS(\rho_0, \sigma_0) + \epsilon]$:

$$\gamma = \frac{\Pi \rho^{\otimes N} \Pi}{\text{tr}(\Pi \rho^{\otimes N})}. \tag{23}$$

We calculate the quantum fidelity to show that our data compression is indeed faithful:

$$F(\rho^{\otimes N}, \gamma) = \left(\text{tr} \sqrt{\sqrt{\rho^{\otimes N}} \gamma \sqrt{\rho^{\otimes N}}} \right)^2 = \text{tr}(\Pi \rho^{\otimes N}) \geq 1 - \delta. \tag{24}$$

When i^N doesn't satisfy the length condition, it doesn't satisfy strong typicality. When we project onto fewer qubits than $NS(\rho_0, \sigma_0)$, we miss out all the typical states, which composes the majority of all possible quantum states, and the data compression fails. Hence, $S(\rho_0, \sigma_0)$ is also the optimal compression rate under this protocol.

7

For the last step, we apply unitary operator $U^\dagger$ to turn the state back into its original basis, and complete our data compression protocol.

If we take into account the fact that the number of qubits has to be integer, then we have $\langle l \rangle \in [S(\rho_0, \sigma_0), S(\rho_0, \sigma_0) + 1)$, and we need no more than $NS(\rho_0, \sigma_0) + N$ qubits for a successful data compression even when our knowledge of the quantum source is wrong. Of course, if our perceived source state σ_0 is far from the true source state ρ_0, $S(\rho_0, \sigma_0)$ is huge and we will be better off by just sending all the information-bearing qubits, which gives a compression rate of $\lceil \log D \rceil$ qubits.

4 Conclusion

In this work, we present a simple quantum protocol that does lossless quantum source coding, and show that the corresponding compression rate is the quantum cross entropy. Since data compression can be viewed as a machine learning task, quantum cross entropy also acts as a loss function, whose minimum is the von Neumann entropy. This is consistent with the result that von Neumann entropy is the optimal compression rate. It will be interesting to evaluate time and query complexity [15, 16, 17, 18] required in carrying out this protocol on quantum computers. With a quantitative measure like complexity, it enables us to compare this protocol with other quantum compression protocols and find improvements. It will also be desirable if any connections between the quantum cross entropy and the holographic codes can be drawn [19, 20]. Ultimately, a broader and deeper understanding of the quantum cross entropy can guide us in designing efficient quantum machine learning algorithms, which leads to solutions to challenging problems like modeling our universe on a quantum computer.

Acknowledgement

References

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

[2] 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.

[3] Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019.

[4] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1):1–6, 2018.

[5] Kerstin Beer, Dmytro Bondarenko, Terry Farrelly, Tobias J Osborne, Robert Salzmann, Daniel Scheiermann, and Ramona Wolf. Training deep quantum neural networks. *Nature communications*, 11(1):1–6, 2020.

[6] Zhou Shangnan and Yixu Wang. Quantum cross entropy and maximum likelihood principle. *arXiv preprint arXiv:2102.11887*, 2021.

[7] Benjamin Schumacher. Quantum coding. *Phys. Rev. A*, 51:2738–2747, Apr 1995.

[8] Michael A Nielsen and Isaac L Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2011.

[9] Benjamin Schumacher and Michael D Westmoreland. Indeterminate-length quantum coding. *Physical Review A*, 64(4):042304, 2001.

[10] Richard Jozsa and Stuart Presnell. Universal quantum information compression and degrees of prior knowledge. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 459(2040):3061–3077, 2003.

[11] Dénes Petz and Milán Mosonyi. Stationary quantum source coding. *Journal of Mathematical Physics*, 42(10):4857–4864, 2001.

[12] Charles H Bennett, Aram W Harrow, and Seth Lloyd. Universal quantum data compression via nondestructive tomography. *Physical Review A*, 73(3):032336, 2006.

[13] Abbas El Gamal and Young-Han Kim. *Network information theory*. Cambridge university press, 2011.

[14] S.L. Braunstein, C.A. Fuchs, D. Gottesman, and Hoi-Kwong Lo. A quantum analog of huffman coding. *IEEE Transactions on Information Theory*, 46(4):1644–1649, 2000.

[15] Giulio Chiribella, Giacomo Mauro D'Ariano, Paolo Perinotti, and Benoit Valiron. Quantum computations without definite causal structure. *Phys. Rev. A*, 88:022318, Aug 2013.

[16] Stefano Facchini and Simon Perdrix. Quantum circuits for the unitary permutation problem. In *International Conference on Theory and Applications of Models of Computation*, pages 324–331. Springer, 2015.

[17] Zhou Shangnan. Complexity, entropy, and markov chains. *arXiv preprint arXiv:1902.10538*, 2019.

[18] Andris Ambainis. Understanding quantum algorithms via query complexity. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 3265–3285. World Scientific, 2018.

[19] Zhao Yang, Patrick Hayden, and Xiao-Liang Qi. Bidirectional holographic codes and sub-ads locality. *Journal of High Energy Physics*, 2016(1):175, 2016.

[20] Fernando Pastawski and John Preskill. Code properties from holographic geometries. *Physical Review X*, 7(2):021022, 2017.