8.371 (QIS II): ENT Lecture 1

Entanglement as a Resource

Angus Lowe*

April 22, 2026

1 Introduction

It's worth starting from the beginning, more or less. Let A and B be two systems (AKA parties, registers, etc) and $\mathcal{H}_A$ and $\mathcal{H}_B$ be two Hilbert spaces of dimensions d_A and d_B, respectively, corresponding to these two systems. We say that a quantum state $|\psi\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B$ for the joint system AB is a *product* state if there exists a pair of states $|\varphi_1\rangle_A$ and $|\varphi_2\rangle_B$ such that $|\psi\rangle_{AB} = |\varphi_1\rangle_A \otimes |\varphi_2\rangle_B$. Otherwise, we say that $|\psi\rangle$ is *entangled*. Some fundamental properties of entanglement are worth stating immediately, to convince you that the concept is interesting and important. Entanglement is:

1. *Not just correlation.* Entanglement may resemble classical correlation at first glance, but it is *not*. (E.g., monogamy, nonlocality, etc).

2. *Generic mathematically.* Most quantum states in the Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$ are *highly* (volume law) entangled [HLW06].

3. *Restricted physically.* The states we write down to model realistic many-body systems are usually *not* highly entangled in the above sense. Thermal states of local Hamiltonians are separable at a sufficiently high temperature (see, e.g., [BLMT25]), and even pure ground states, e.g., those of geometrically-local and gapped systems are suspected to have limited (area law) entanglement[1].

4. *Needed for exponentially faster quantum algorithms.* Entanglement is necessary (but not sufficient) for quantum computation which cannot be simulated efficiently by classical means.

5. *Helpful for QI proofs.* "Purifying" noisy systems as entangled systems on a larger Hilbert space is a powerful tool for reasoning about information theory and security (Shannon

*alowe7@mit.edu

[1] This has only been rigorously shown in 1D [Has07], and in two or higher dimensions remains an open problem except in special cases [AAG22].

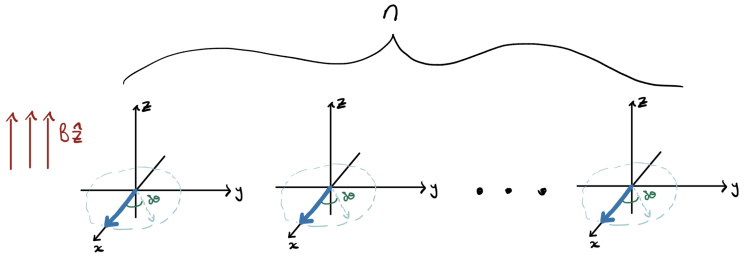


Figure 1: Measuring the angle θ from Larmor precession, perhaps to infer the strength of the magnetic field B.

theory, analyzing QKD, compressed oracles, etc). We will see an example of this idea shortly.

6. *Useful for science.* Entanglement can enable precise measurements with Heisenberg-limited scaling [TA14]. Bell inequalities were originally proposed to test the validity of quantum theory, thus securing the Nobel Prize in Physics for Aspect, Clauser, and Zellinger in 2022. And the list goes on.

We will not be able to cover all these points in this lecture, but let us briefly touch on one example pertaining to the last point.

1.1 Quantum-enhanced measurement

Recall from quantum mechanics that a spin-$\frac{1}{2}$ particle in a uniform magnetic field, originally pointing perpendicular to the field, precesses (see Fig. 1). The frequency of this precession is called the Larmor frequency and depends in particular on the strength of the magnetic field and the magnetic moment of the particle. Hence, measuring this frequency can reveal important information about the particles and their environment, which has practical applications; for example, in NMR and MRI technologies. In quantum information terms, we have a qubit A in some initial state $|\psi_0\rangle \in \mathcal{H}_A \cong \mathbb{C}^2$ which undergoes time evolution generated by the Hamiltonian $H_1 := \left(\frac{\omega}{2}\right) Z$. This results in a unitary $U_\theta := e^{-i\theta Z/2}$ being applied to $|\psi_0\rangle$ where θ is ω times the total evolution time. Our goal is to measure the parameter θ as precisely as possible after n uses of this unitary.

One simple strategy is to repeatedly let a single spin precess; that is, prepare the state $|\psi_0\rangle = |+\rangle$, evolve by U_θ, and measure in the $\{|+\rangle, |-\rangle\}$ basis. After some algebra, we find the probability of seeing "$+$" is given by

$$p_+ = \frac{1}{2} + \frac{1}{2}\cos(\theta). \tag{1}$$

If we tried to estimate the probability of this event, we would be computing the mean of a Bernoulli random variable with standard deviation $\sqrt{p_+(1-p_+)} = \frac{1}{2}\sin(\theta)$. For n independent applications of this procedure, the empirical mean of the estimate of p_+

has standard deviation $\frac{1}{2\sqrt{n}}\sin(\theta)$, and therefore, by propagation of errors[2], our error in the estimate of θ will scale like $\delta\theta \sim \frac{1}{\sqrt{n}}$. This scaling behaviour is known as the *standard quantum limit*.

We can do better by exploiting an entangled initial state. Suppose that, instead of performing the experiments separately, we prepare n qubits in the initial state $|\psi_0\rangle \in (\mathcal{H}_A)^{\otimes n}$. If we put them all in the same magnetic field, the relevant Hamiltonian to consider is $H = \frac{\omega}{2}\sum_{j=1}^{n} Z_j$, which generates $U_\theta^{\otimes n}$. If $|\psi_0\rangle = |+\rangle^{\otimes n}$ then we have gained nothing from shifting to this perspective, compared to the previous discussion. However, if we pick the entangled state

$$|\psi_0\rangle = |\mathrm{GHZ}\rangle := \frac{1}{\sqrt{2}}\left(|0^n\rangle + |1^n\rangle\right) \tag{2}$$

it is not hard to see that there is a measurement in a certain basis (left as an exercise, perhaps in the language of circuits) which gives us the outcome "+" with probability

$$p_+ = \frac{1}{2} + \frac{1}{2}\cos(n\theta) \tag{3}$$

and gives "−" otherwise. Now, the phases have added coherently. By propagation of errors once again, the uncertainty in estimating θ scales like $\frac{1}{n}$, a quadratic improvement in precision over the unentangled case. This is known as *Heisenberg-limited scaling*.

1.2 Teleportation & superdense coding

Let us briefly turn our attention to another usage of entangled states with which you may already be familiar. Suppose Alice has a qubit A' in some quantum state $|\psi\rangle$ and she wishes for Bob's system B to be prepared in the same state. They do not have the ability to send qubits to each other; however, they can send classical bits, and they share an entangled EPR pair

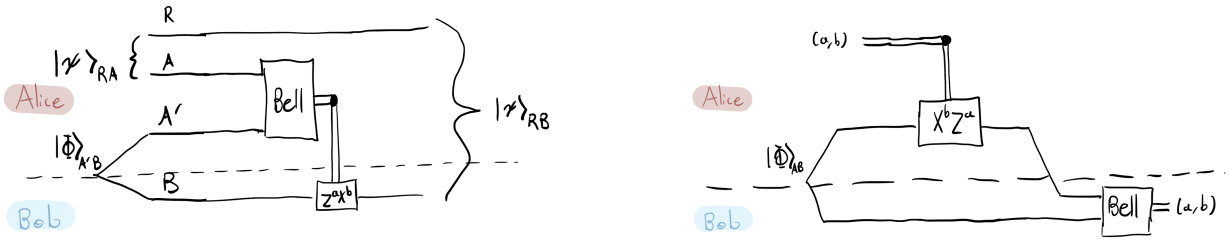
$$|\Phi\rangle_{AB} = \frac{1}{\sqrt{2}}\left(|0\rangle_A \otimes |0\rangle_B + |1\rangle_A \otimes |1\rangle_B\right). \tag{4}$$

Then it is possible to accomplish this task using a protocol that sends two classical bits to Bob, as depicted in Fig. 2a. The end result is a recipe for sending a qubit using two classical bits and one "entangled bit". This means that a single qubit of communication can be simulated by 2 classical bits of communication together with 1 EPR pair, which can be written:

$$2\ \text{cbits} + 1\ \text{ebit} \geqslant 1\ \text{qbit}. \tag{5}$$

Note that we have yet to define what the unit "ebit" means: we will see how to do this in Section 4, and that it is the right quantum analogue of a shared random bit. This is *quantum state teleportation* [BBC+93]. It is also used in modular quantum computation and fault-tolerance.

[2] $\delta f(x) = \delta x / |\frac{df}{dx}|$.



(a) Quantum teleportation.

(b) Superdense coding.

Figure 2

Similarly, the *superdense coding protocol* [BW92] (also depicted in Fig. 2b) gives a way of going in the other direction:

$$1 \text{ qbit} + 1 \text{ ebit} \geqslant 2 \text{ cbits}. \tag{6}$$

The reader ought to work this out carefully for themselves. The above discussion can be made more rigorous (see, e.g., [DHW04]), and provides a clear example of how entanglement can supplement other resources at our disposal – classical and quantum communication, in this case – enabling tasks which would otherwise be impossible.

Remainder of this lecture. The previous two sections, along with prior content in this course, lends credence to the viewpoint that *entanglement is a resource* in its own right, with diverse applications. In the remainder of this lecture, our goal will be to *(i) quantify entanglement as a resource*, and *(ii) develop mathematical tools from information theory* which are powerful enough to be applied in settings beyond the main topics in this lecture.

2 Mathematical background

2.1 Bipartite states

Vectorization. There is a linear, bijective map $\mathrm{vec}: \mathbb{C}^{d_A \times d_B} \to \mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}$, called *vectorization*, with the following action on standard basis states:

$$\mathrm{vec} : |i\rangle_A \langle j|_B \mapsto |i\rangle_A \otimes |j\rangle_B . \tag{7}$$

In particular, it is not hard to show that product states of the form $|\psi\rangle |\varphi\rangle$ get mapped to $|\psi\rangle\langle\overline{\varphi}|$ when you apply vec^{-1}, where the bar means entry-wise complex conjugate of the vector.

Singular value decomposition (SVD). For any matrix $X \in \mathbb{C}^{d_A \times d_B}$ of rank r, we can write

$$X = \sum_{\alpha=1}^{r} \sqrt{p_\alpha} |u_\alpha\rangle\langle\overline{v_\alpha}| \tag{8}$$

for some nonnegative *singular values* $\sqrt{p_\alpha} \in \mathbb{R}$ as well as left and right *singular vectors* $|u_\alpha\rangle$, and $|\overline{v_\alpha}\rangle$, respectively. Moreover, these singular vectors are orthonormal within their respective Hilbert spaces. It will become apparent momentarily why we choose to define the right singular vectors with the bar, and square-root the singular values.

Schmidt decomposition. By applying the vectorization map to both sides of Eq. (8), we see the following holds for an arbitrary bipartite quantum state $|\psi\rangle \in \mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}$:

$$\boxed{|\psi\rangle_{AB} = \sum_{\alpha=1}^{r} \sqrt{p_\alpha}\, |u_\alpha\rangle_A \otimes |v_\alpha\rangle_B} \tag{9}$$

where, once again, the $\sqrt{p_\alpha} \in \mathbb{R}$ are nonnegative, and $|u_\alpha\rangle$, and $|v_\alpha\rangle$ are orthonormal. In this context, they are called *Schmidt values* and *Schmidt vectors*, respectively, and r is the *Schmidt rank*. Also, by normalization, we must have $\sum_\alpha p_\alpha = 1$, justifying our notation.

The vectorization perspective means there is almost nothing to prove here. It also makes statements like

$$|\Phi\rangle \equiv \frac{1}{\sqrt{2}} (|0\rangle\,|0\rangle + |1\rangle\,|1\rangle) = \frac{1}{\sqrt{2}} (|+\rangle\,|+\rangle + |-\rangle\,|-\rangle) \tag{10}$$

nearly trivial: up to the $\frac{1}{\sqrt{2}}$ factor, the EPR state is the vectorization of the 2×2 identity matrix, which is diagonal in every basis.

Partial trace. Suppose we have a bipartite state ρ_{AB}. How do we describe the state of our system on A, ignorant of B? Namely, we want a description such that all observables give the same result as if we had performed the measurement locally on A and ignored B. This is the quantum analogue of marginalizing a probability vector.

The right answer here is to *partial trace*[3]. Like vectorization, we define this on basis states and extend it by linearity:

$$\begin{aligned} \mathrm{tr}_B \colon |i\rangle\langle j|_A \otimes |k\rangle\langle \ell|_B &\mapsto |i\rangle\langle j|_A \otimes \mathrm{tr}(|k\rangle\langle \ell|_B) && (11)\\ &= \langle k|\ell\rangle\, |i\rangle\langle j|_A && (12)\\ &= \delta_{k\ell}\, |i\rangle\langle j|_A\,. && (13) \end{aligned}$$

If $\rho = |\psi\rangle\langle\psi| =: \psi$ is a pure state where $|\psi\rangle$ has a Schmidt decomposition as in Eq. (9), one may compute:

$$\rho_A \coloneqq \mathrm{tr}_B(\psi) = \sum_{\alpha=1}^{r} p_\alpha\, |u_\alpha\rangle\langle u_\alpha| = XX^\dagger, \qquad \rho_B \coloneqq \mathrm{tr}_A(\psi) = \sum_{\alpha=1}^{r} p_\alpha\, |v_\alpha\rangle\langle v_\alpha| = X^\dagger X \tag{14}$$

where X is as in Eq. (8). These are called the *reduced states* of ρ. In particular, ρ_A and ρ_B have the same eigenvalues if ρ is pure.

[3]The mathematical reason for this is that it is the adjoint to the operation "tensor with the identity matrix".

2.2 Information theory toolkit

Classical sources of information. We define a *source of information* as anything that prepares a register X of dimension $d := d_X$ according to some probability vector $p\colon \{1,\dots,d\} \to [0,1]$ in $\mathbb{R}^d$, potentially many times. That probability vector is the state of the register. Typically, we are interested in messages generated by that source which are composed of letters drawn iid from the *alphabet* $[d] \equiv \{1,\dots,d\}$. We formally describe this as a joint register $(X_1,\dots,X_n)$ which is in the state $p^{\otimes n} : [d]^n \to [0,1]$, a product distribution.

For example, there may be a source of information that generates a bitstring (so $d = 2$ in this case) where $p(0) = 1-\gamma$ and $p(1) = \gamma$ for some $0 \leqslant \gamma \leqslant 1$. This leads to

$$p^{\otimes n}(x) = \prod_{j=1}^{n} p(x_j) = \gamma^{|x|}(1-\gamma)^{n-|x|} \tag{15}$$

for every $x \in \{0,1\}^n$, where $|x|$ is the Hamming weight of x. This is equivalent to n iid Bernoulli random variables with parameter γ. It may be tempting to posit human beings as sources of information, perhaps moreso with the advent of LLMs, though the iid assumption seems difficult to justify. A monkey at a typewriter would presumably give $p(\ell) = \frac{1}{26}$ for every letter ℓ in the English alphabet.

Entropy. Entropy measures our uncertainty about a source of information. Another interpretation: The entropy of a register is the average amount of "surprise" upon observing its value. Suppose I want to claim I am "three times as surprised" at flipping three heads in a row as I am when flipping a heads once. Then my measure of the surprise better be additive, which suggests logarithms. Thus, the surprise of $x \in [d]$ is defined as $-\log p(x)$, and the *(Shannon) entropy* of the register X is

$$\boxed{H(X)_p \equiv H(p) := -\mathop{\mathbb{E}}_{x\sim p} \log p(x) = -\sum_{x\in[d]} p(x)\log p(x).} \tag{16}$$

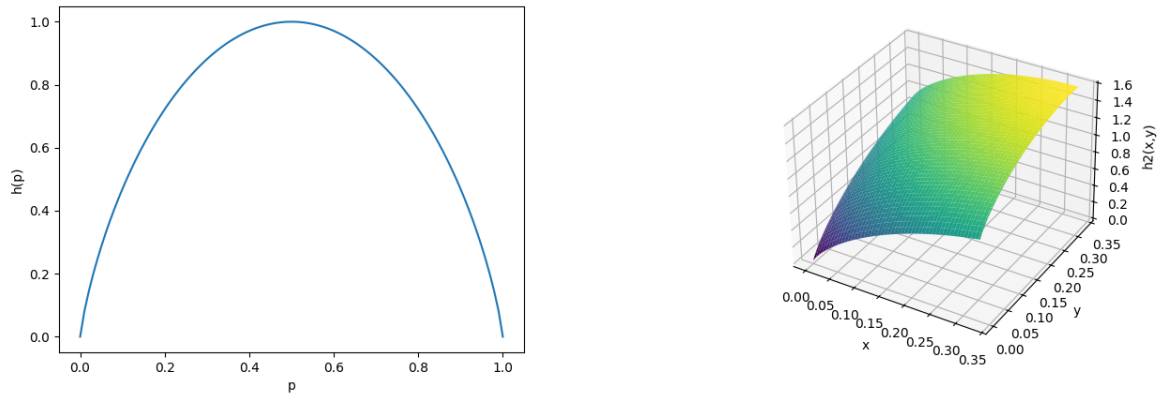
Here and throughout, $0\log 0 \equiv 0$.

Some useful special cases are $H(\mathrm{Unif}) = \log d$ and $H((1,0,0,\dots)) = 0$, which represent the two extreme cases.

If we sample according to the distribution p a total of n times independently, we will see an outcome $x \in [d]^n$ that satisfies

$$-\delta \leqslant -\frac{1}{n}\log p^{\otimes n}(x) - H(p) \leqslant \delta \tag{17}$$

for some small constant δ, with arbitrarily high probability as $n \to \infty$. To see this, use



(a) (b)

Figure 3: (a) Binary entropy $h(p) \equiv H((1-p,p))$, which attains its maximum at $1/2$. (b) Entropy of probability vector $(1-x-y,x,y)$ on three letters, which attains its maximum at $(1/3,1/3,1/3)$. Both cases demonstrate the continuity of the entropy function.

the law of large numbers (LLN) and note

$$\mathbb{E}\log p^{\otimes n}(\mathbf{x}) = \mathbb{E}\log\prod_{j=1}^{n} p(\mathbf{x}_j) \tag{18}$$

$$= \sum_{j=1}^{n} \mathbb{E}\log p(\mathbf{x}_j) \tag{19}$$

$$= -nH(p). \tag{20}$$

So Eq. (17) follows from the convergence of the empirical mean to the true mean asymptotically as $n \to \infty$, i.e., the LLN. And if you do not wish to take that for granted, the method described in the next subsection gives a way to prove LLN from more basic principles.

We also mention the definition of the *relative entropy* (AKA KL divergence) between two distributions p and q over an alphabet $[d]$:

$$\boxed{D(p \parallel q) \coloneqq \sum_{x\in[d]} p(x)\log\frac{p(x)}{q(x)} = -H(p) - \sum_{x\in[d]} p(x)\log q(x)} \tag{21}$$

This is well-defined so long as the support of p is contained in that of q, and gives us a way to compare how "close" two distributions are to each other. It is often the right way to measure closeness in information theory.

The method of types. We now introduce the *method of types*. This is a tool developed in information theory with broad application to statistics and quantum information. To

estimate the probabilities $p(1), \ldots, p(d)$ which describe a source of information, one may construct the empirical histogram coming from n samples. As we increase the number of samples, we expect the histogram to match the true distribution more closely. The method of types makes this precise.

The histogram is a mapping of the outcomes $x_1, \ldots, x_n \in [d]$ to the vector $t_x \in \mathbb{Q}^d$, called the *type* of x. (So empirical histogram and type mean the same thing.) For example, if $d = 4$ and $n = 6$

$$x = 141224, \qquad t_x = \frac{1}{6}(2, 2, 0, 2). \tag{22}$$

This is clearly a many-to-one mapping, and we can use it to partition the set of all strings $x \in [d]^n$ into different *type classes* labelled by their corresponding type vector $t \in \mathbb{Q}^d$:

$$C(t) \coloneqq \{x \in [d]^n : t_x = t\}. \tag{23}$$

The set of all type classes for strings of length n on d letters is denoted by $\mathcal{C}_n^d$. Let us count the number of different type classes: there is one for each vector $t \in \mathbb{Q}^d$, and these are all nonnegative multiples of $1/n$ summing to one. Hence, there are

$$|\mathcal{C}_n^d| \leqslant (n+1)^d \tag{24}$$

such classes, which is only polynomially large in n. As a consequence, there are exponentially many strings in at least some of these classes. In fact, as we will show momentarily, the largest of these turns out to contain a number of strings which is comparable to that of the entire set, d^n.

For example, consider a source which emits n random bits, so $d = 2$ and we set $p(0) = 1-\gamma$ and $p(1) = \gamma$ for some $0 \leqslant \gamma \leqslant 1$. Then the possible types are

$$(1,0),\ (\tfrac{n-1}{n}, \tfrac{1}{n}),\ \ldots,\ (\tfrac{n-k}{n}, \tfrac{k}{n}),\ \ldots,\ (0,1). \tag{25}$$

which is uniquely specified by the Hamming weight k. Observe that $C((1,0))$ is just the all-zeros string, $C((\frac{n-1}{n}, \frac{1}{n}))$ is the set of all strings with a single one, and so on. We can easily count the size of these different type classes: if t corresponds to Hamming weight k then

$$\frac{1}{(n+1)^2} 2^{nH(t)} \leqslant |C(t)| = \binom{n}{k} \leqslant 2^{nH(t)} \tag{26}$$

where the upper bound is because

$$1 = (1-\gamma+\gamma)^n \qquad (\gamma \equiv \tfrac{k}{n}) \tag{27}$$

$$= \sum_{j=0}^{n} \binom{n}{j} \gamma^j (1-\gamma)^{n-j} \tag{28}$$

$$\geqslant \binom{n}{k} \gamma^j (1-\gamma)^{n-k} \tag{29}$$

$$= \binom{n}{k} 2^{-nH(t)}. \tag{30}$$

The lower bound follows from Stirling's-approximation-type bounds[4].

A minor leap of faith from the $d = 2$ case (or see [CT05, Chapter 11]) takes us to tight bounds for the size of a type class for general d:

$$\boxed{\frac{1}{(n+1)^d} 2^{nH(t)} \leqslant |C(t)| \leqslant 2^{nH(t)}.} \tag{31}$$

Now that we have characterized the size of the type classes, let us characterize their probabilities. By making use of the expression in Eq. (21) and a bit of algebra we can write

$$\boxed{p^{\otimes n}(x) = 2^{-n(H(t_x) + D(t_x \,\|\, p))}.} \tag{32}$$

In particular, the probability of observing x only depends on its type, and strings with types that are "far" from the true distribution p are exponentially suppressed. For this reason they are called *atypical*, in contrast to the set of *typical* strings which is defined through:

$$\mathcal{T}^n_{p,\delta} \coloneqq \{x \in [d]^n : D(t_x \,\|\, p) \leqslant \delta\}. \tag{33}$$

(The parameter $\delta > 0$ is something to be specified in the discussion surrounding use of the set.)

As a demonstration of the utility of these facts, let us prov a form of the *LLN*, which is: *in the asymptotic limit $n \to \infty$, a string is almost surely typical*. We compute

$$\begin{aligned}
\Pr_{x \sim p^{\otimes n}}[x \notin \mathcal{T}^n_{p,\delta}] &= \sum_{t : D(t \,\|\, p) > \delta} \sum_{x \in C(t)} p^{\otimes n}(x) && (34)\\
&= \sum_{t : D(t \,\|\, p) > \delta} |C(t)| \cdot 2^{-n(H(t) + D(t \,\|\, p))} && (35)\\
&\leqslant \sum_{t : D(t \,\|\, p) > \delta} 2^{-n\delta} && (36)\\
&\leqslant |\mathcal{C}^d_n| \cdot 2^{-n\delta} && (37)\\
&\leqslant (n+1)^d 2^{-n\delta} && (38)\\
&\leqslant 2^{-n\left(\delta - \frac{d \log n}{n}\right)}. && (39)
\end{aligned}$$

From the last line, it is apparent that this goes to zero as $n \to \infty$. This statement implies the standard form of the (weak) LLN, which is that empirical means converge to their true means. In particular, this proves Eq. (18). We depict it schematically in Fig. 4.

Source coding. Shannon gave a nearly complete formulation of classical information theory in 1948, a few years after finishing his Master's thesis here at MIT[5]. We start

[4] Use the exact bounds $e(n/e)^n \leqslant n! \leqslant en(n/e)^n$. A stronger lower bound is possible: see Example 11.1.3 in [CT05].

[5] The curious reader can start here for more historical details: https://en.wikipedia.org/wiki/History_of_entropy#Information_theory

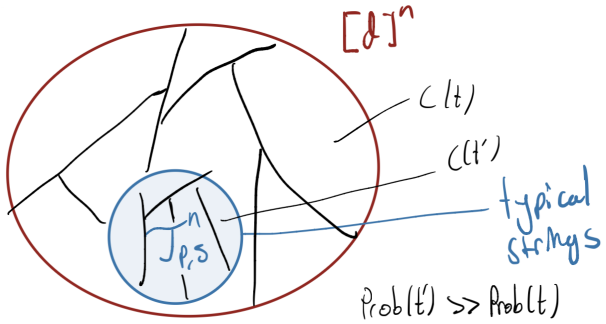


Figure 4: Type classes $C(t), C(t')$, etc within the set of all n-letter strings $[d]^n$. The type class $C(t')$ contains only typical strings, and is much more likely to be observed than $C(t)$. For large n, a sampled string $x \sim p^{\otimes n}$ is overwhelmingly likely to be within the blue region $\mathcal{T}^n_{p,\delta}$, the typical set.

with compression of a classical source of information, also known as *source coding*. It begins with the following problem: Given a source of information with distribution $p : [d] \to [0,1]$, what is the optimal *rate* at which I can compress the source? Here, we say that we can compress the source at rate R if there is a family of protocols (one protocol for each n) with the following behaviour: For any $\delta > 0$ we apply the operation `Compress` to the register $X = (X_1, \ldots, X_n)$, mapping it to a register Y of dimension at most $2^{\lceil n(R+\delta)\rceil}$, so that the original message $x \in [d]^n$ can later be recovered with probability of error tending to zero as $n \to \infty$ (using an operation `Decompress` applied to Y). In this case we say that the rate R is *achievable*. In other words, R represents the number of bits per letter that we need to represent our message. By optimal rate we mean the infimum over all achievable rates, call it $R^\star_{\text{compr}}(p)$. *Shannon's source coding theorem* states:

$$\boxed{R^\star_{\text{compr}}(p) = H(p).} \tag{40}$$

So the answer is we can compress all the way down to $H(p)$ bits per letter.

The achievability proof is straightforward using the method of types: By Eq. (39), $x \sim p^{\otimes n}$ is typical with probability at least

$$1 - 2^{-n\left(\delta - \frac{d \log n}{n}\right)}. \tag{41}$$

The size of the typical set is

$$
\begin{aligned}
|\mathcal{T}^n_{p,\delta}| &= \sum_{t:D(t\,\|\,p)\leqslant\delta} |C(t)| && (42)\\
&\leqslant \sum_{t:D(t\,\|\,p)\leqslant\delta} 2^{nH(t)} && (43)\\
&\leqslant \sum_{t:D(t\,\|\,p)\leqslant\delta} 2^{n(H(p)+\delta')} && (44)\\
&\leqslant |\mathcal{C}^d_n| \cdot 2^{n(H(p)+\delta')} && (45)\\
&\leqslant (n+1)^d 2^{n(H(p)+\delta')} && (46)\\
&= 2^{n\left(H(p)+\delta'-d\frac{\log(n+1)}{n}\right)} && (47)
\end{aligned}
$$

In the second inequality, we used the fact that closeness in relative entropy implies closeness in *total variation distance* (see [Wik26]), which in turn means closeness of $H(p)$ to $H(t)$ by continuity of the Shannon entropy[6]. So $\delta' > 0$ is some small constant depending on δ. It is evident that the size of this set can be made arbitrarily close to $2^{nH(p)}$ by choosing parameters appropriately. Hence, the compression protocol is simple: *Let the typical strings $x \in \mathcal{T}^n_{p,\delta}$ be labelled $\{1, 2, \ldots, |\mathcal{T}^n_{p,\delta}|\}$. If x is typical, send the label. Otherwise, declare failure.*

An equivalent description of the compression task is given by Fig. 5. Namely, Alice has a source of information in her register A and is trying to send a message generated by that source to Bob, using as few letters as possible. She records her message in a separate register R before sending it to Bob so she can check for correctness at the end of the protocol. This results in an initial state $p_{\text{init}} : [d]^n \times [d]^n$ which is perfectly correlated:

$$
p_{\text{init}}(x, y) = \begin{cases} p^{\otimes n}(x), & x = y \\ 0, & \text{o/w.} \end{cases} \tag{48}
$$

She applies a compression map to A, sends the message to Bob using $k \leqslant n \log d$ bits, and then Bob applies a decoding map. A successful procedure is one where Bob has preserved correlation with the correct answer stored in R. Hence, we want the resulting classical state $\widetilde{p}_{\text{init}}$ at the end of the protocol to be close to the original perfectly correlated state. Though somewhat non-standard, this second viewpoint is the one which is most easily generalized to the quantum case.

Shared randomness distillation. Let us now turn to another task in classical information, called *shared randomness distillation* (AKA randomness compression), which will be important in our study of entanglement. It begins with Alice and Bob who share a perfectly correlated state p_{init}, depending on p in the same way as above (except this time

[6] We have not shown this here but the proof is straightforward. See, e.g., Box 11.2 in [NC10] on *Fannes' Inequality*.

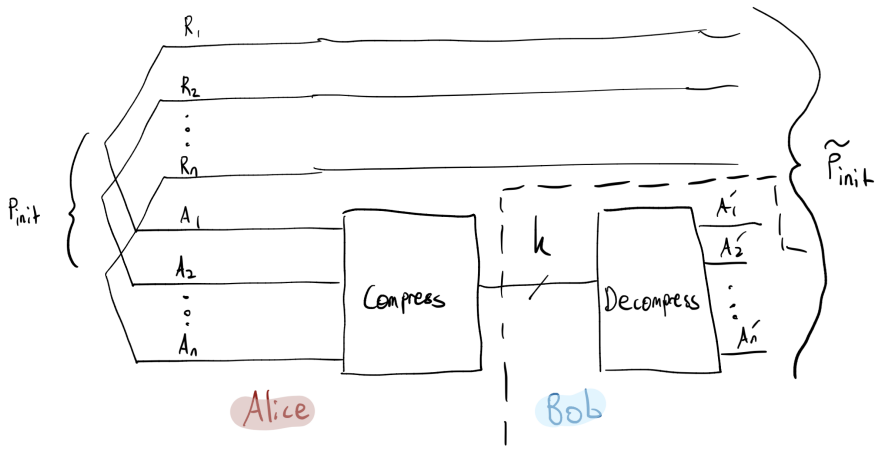


Figure 5: Source coding task (non-standard, but equivalent description).

shared with Bob instead of the reference register R) and their goal is to share $k \leqslant n \log d$ *uniformly* random bits by the end of their procedure. We grant them the ability to partially measure their registers and condition on the outcomes observed. We are interested in computing the optimal rate of distillation $R^{\star}_{\text{dis.}}(p)$, defined as the supremum over achievable k/n.

We prove achievability using the method of types once again. Let Alice and Bob take identical actions, so it suffices to describe what Alice does. She applies the map $x \mapsto (t_x, x)$ to her registers and looks only at the type $t = t_x$. If it's atypical she declares failure. Otherwise, she proceeds. since the probability of a string depends only on its type, all strings compatible with the measurement outcome t have the same probability. Hence, the distribution conditioned on the outcome t is uniform over $C(t)$. Since there are

$$|C(t)| \geqslant \frac{1}{(n+1)^d} 2^{nH(t)} \tag{49}$$

$$\geqslant \frac{1}{(n+1)^d} 2^{n(H(p)-\delta')} \tag{50}$$

$$= 2^{n(H(p)-\delta'-d\frac{\log(n+1)}{n})} \tag{51}$$

such strings, where $\delta' > 0$ is an arbitrarily small constant (once again by continuity), we see that $H(p)$ shared random bits per letter is achievable for this task as well. So,

$$\boxed{R^{\star}_{\text{dis.}}(p) = H(p).} \tag{52}$$

3 Non-asymptotic conversion of entanglement

Let us now return to the question of how we can quantify entanglement as a resource. A natural desired criterion for such a quantity is that it will allow us to say that $|\psi\rangle$ is

"more entangled" than $|\varphi\rangle$ if and only if $|\psi\rangle$ can be transformed into $|\varphi\rangle$ "for free", in a sense which we are about to explain in a bit more detail.

LOCC. We first define *local operations and classical communication (LOCC)* as the set of quantum operations on AB which can be implemented by arbitrarily many rounds of (i) local quantum operations followed by (ii) classical communication between A and B. We omit a rigorous mathematical definition as it would take us too far afield, but hopefully the definition makes intuitive sense. The relevant property about this set of operations is: *entanglement cannot be created by LOCC*. Thus, LOCC serves as a reasonable choice of operations that can be considered "free", and we shall adopt this viewpoint for the remainder of this lecture.

Entanglement transformation & majorization. It turns out that we can exactly characterize whether one entangled state can be transformed into another as follows. Let $|\psi\rangle$ have Schmidt coefficients $\sqrt{p_1}, \ldots, \sqrt{p_d}$ and $|\varphi\rangle$ have Schmidt coefficients $\sqrt{q_1}, \ldots, \sqrt{q_d}$. Then: $|\psi\rangle$ *can be converted to* $|\varphi\rangle$ *if and only if* $p \prec q$. Here $p \prec q$ means that q *majorizes* p, which means that

$$\sum_{i=1}^{k} p_i^{\downarrow} \leqslant \sum_{i=1}^{k} q_i^{\downarrow} \text{ for every } k = 1, \ldots, d, \tag{53}$$

where the downarrow $\downarrow$ means we are sorting the Schmidt coefficients in non-increasing order first. This is known as *Nielsen's theorem* [Nie99].

Although Nielsen's theorem gives an exact criterion for which entangled states can be transformed into each other, we are left with the drawback that certain states can be *incomparable*. (Mathematically, we have a partial ordering.) For instance, consider the case where $p = (\frac{1}{2}, \frac{1}{2}, 0)$ and $q = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$. Then it is easy to check that $p \not\prec q$ and $q \not\prec p$. Therefore, there are examples for which one cannot say that one state is "more entangled" than another using this characterization.

To resolve this issue, we must move to the asymptotic setting in which one considers many identical copies n of the states being compared, and admits small errors in the desired transformation which go to zero as $n \to \infty$. This has the appealing feature of giving us a single number $E(\cdot)$ that consistently quantifies the amount of entanglement present in a given state, allowing us to say, roughly speaking, that $E(|\psi\rangle) \geqslant E(|\varphi\rangle)$ if and only if the transformation $|\psi\rangle^{\otimes n} \to |\varphi\rangle^{\otimes n(1-\delta)}$ is possible using just LOCC in the asymptotic setting of large n. We will discuss this setting more precisely below. But first, let us see what $E(\cdot)$ is.

4 Entanglement entropy with examples

We will soon see that a good choice for a measure of entanglement is the entanglement entropy. We define it below, and give some properties. Firstly, the *(von Neumann) entropy*

of a density matrix ρ describing a system A with eigenvalues $p_1, \ldots, p_d$ is given by

$$S(A)_\rho \equiv S(\rho) \coloneqq -\operatorname{tr}(\rho \log \rho) = -\sum_x p_x \log p_x. \tag{54}$$

The *entanglement entropy* of a bipartite state $|\psi\rangle_{AB}$ with Schmidt coefficients $\sqrt{p_1}, \ldots, \sqrt{p_d}$ and reduced density matrices $\rho_A = \operatorname{tr}_B(|\psi\rangle\langle\psi|)$ and $\rho_B = \operatorname{tr}_A(|\psi\rangle\langle\psi|)$ is given by:

$$\boxed{E(|\psi\rangle) \coloneqq S(A)_\psi \equiv S(\rho_A) = S(\rho_B) = -\sum_x p_x \log p_x.} \tag{55}$$

If $|\psi\rangle$ is a product state, then the reduced density matrix on A or B is pure, and we get $E = 0$. On the other hand,

$$E(|\Phi\rangle) = S(I/2) = 1, \tag{56}$$

which is the largest the entanglement entropy can be for qubits. This is the sense in which the maximally entangled EPR pair $|\Phi\rangle$ is one "ebit" of entanglement.

5 Entanglement distillation & dilution

We will now show that entanglement is a "fungible" resource in the asymptotic setting. That is, any two entangled states can be converted into each other using just LOCC, at an exchange rate which is determined by the entanglement entropy. For example, if $E(|\psi\rangle) = 2E(|\varphi\rangle)$ then for large enough n we can convert n copies of $|\psi\rangle$ into $\lfloor (2-\delta)n \rfloor$ copies of $|\varphi\rangle$ for arbitrarily small δ. Furthermore, we can also go in the other direction, converting n copies of $|\varphi\rangle$ to $\lfloor (1/2-\delta)n \rfloor$ copies of $|\psi\rangle$.

To prove this, we will show that EPR pairs represent a sort of "gold standard" to and from which entanglement in other forms can be converted. The process of converting EPR pairs to other entangled states is known as *entanglement dilution*, while the conversion back to EPR pairs is known as *entanglement distillation*. We will let $R^\star_{\text{dil.}}(|\psi\rangle_{AB})$ denote the infimum over all achievable rates of dilution, where a rate $R_{\text{dil.}}$ is deemed achievable if the transformation

$$\texttt{dilute} : |\Phi\rangle^{\otimes \lceil n(R_{\text{dil.}}+\delta)\rceil} \mapsto \approx |\psi\rangle^{\otimes n} \tag{57}$$

can be implemented using only LOCC, for arbitrarily small δ, with fidelity going to one as $n \to \infty$. Likewise, $R^\star_{\text{dis.}}(|\psi\rangle_{AB})$ denotes the supremum over all achievable rates of distillation, where a rate $R_{\text{dis.}}$ is deemed achievable if the transformation

$$\texttt{distill} : |\psi\rangle^{\otimes n} \mapsto \approx |\Phi\rangle^{\otimes \lfloor n(R_{\text{dis.}}-\delta)\rfloor} \tag{58}$$

can be implemented using only LOCC, for arbitrarily small δ, with fidelity going to one as $n \to \infty$.

Using this terminology, our goal in the rest of this section will be to show that entanglement distillation and dilution have optimal rates that are both equal to the entanglement entropy:

$$\boxed{R^{\star}_{\text{dis.}}(|\psi\rangle_{AB}) = R^{\star}_{\text{dil.}}(|\psi\rangle_{AB}) = E(|\psi\rangle_{AB}) \equiv S(A)_{\psi}.} \tag{59}$$

Note: we already know something about these two rates by physical considerations. Namely, it holds that

$$R^{\star}_{\text{dis.}}(|\psi\rangle_{AB}) \leqslant R^{\star}_{\text{dil.}}(|\psi\rangle_{AB}). \tag{60}$$

Informally, we cannot squeeze more entanglement out of $|\psi\rangle$ than it takes to make it in the first place. This is because we can always construct a protocol that begins with $nR^{\star}_{\text{dil.}}$ EPR pairs, converts to n copies of $|\psi\rangle$, and then converts back to $nR^{\star}_{\text{dis.}}$ copies of $|\Phi\rangle$. If we end up with more than we started with, we have created entanglement using LOCC, which is impossible. This argument is reminiscent of the proofs by contradiction which one often encounters in classical thermodynamics. Indeed, analogies between the two fields, and attempts to solidify them, have served as the motivation for recent research [Lam25, HY25].

Also note that, because of the inequality in Eq. (60), our task has been reduced to showing $E(|\psi\rangle)$ is an achievable rate for both distillation and dilution. By contrast, in classical information theory we often need a *converse* direction to the achievability proofs in order to show that the achievable rates are *optimal*. These were annoying enough that we skipped them in the previous sections. Remarkably, the proofs in the quantum case look as if they could be simpler, equipped with this knowledge. Our plan in the remainder of this section is to graciously accept this gift and forge ahead with showing achievability.

Schumacher compression. *Schumacher compression* [Sch95] is the quantum analogue of source coding. We begin with the same picture as in source coding, except we replace the correlated initial state with the n copies of the entangled state $|\psi\rangle_{AR}$, which we assume has a Schmidt decomposition of the form[7]

$$|\psi\rangle_{AR} = \sum_{x_1=1}^{d} \sqrt{p(x_1)}\, |x_1\rangle \otimes |x_1\rangle\,. \tag{61}$$

so that

$$|\psi\rangle_{AR}^{\otimes n} = \sum_{x \in [d]^n} \sqrt{p^{\otimes n}(x)}\, |x\rangle \otimes |x\rangle\,. \tag{62}$$

A successful protocol for Schumacher compression with rate R performs a quantum operation on the n copies of A, mapping them to a system A' of $\lceil n(R+\delta) \rceil$ qubits for

[7] This is without loss of generality, since what we choose to call the Schmidt vectors is irrelevant to the task.

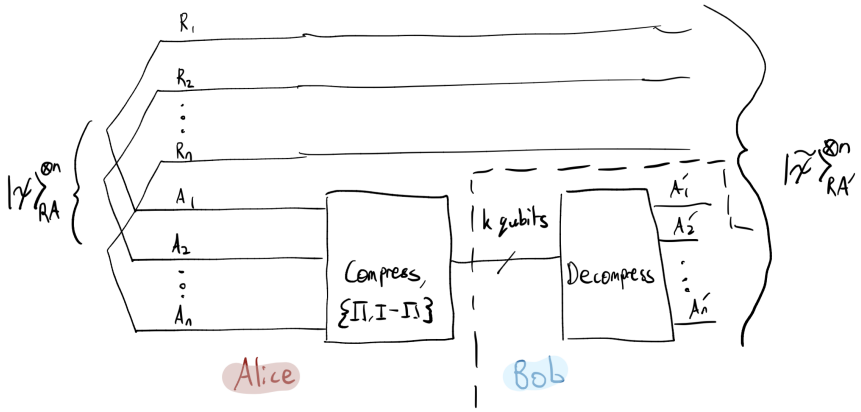


Figure 6: Schumacher compression.

some arbitrarily small δ, such that the following holds: there exists an operation acting only on the system A' that takes the compressed state back to $|\psi\rangle_{AR}^{\otimes n}$ with fidelity going to one as $n \to \infty$. (See Fig. 6.)

The reason for considering compression on half of our system, leaving a reference system untouched, is similar to that in the classical case: to capture what we *mean* by compression, we ought to preserve correlations; or, in this case, entanglement.

The protocol is simple to describe using our knowledge of classical compression: perform the measurement, acting only on the copies of A, onto the *typical subspace*. That is, define

$$\Pi_{T^n_{p,\delta}} \coloneqq \sum_{x \in T^n_{p,\delta}} |x\rangle\langle x| \tag{63}$$

and perform the binary measurement with the "Yes" outcome corresponding to $I_R \otimes (\Pi_{T^n_{p,\delta}})_A$ in order to compress. This measurement returns "Yes" with probability equal to

$$(\langle\psi|)^{\otimes n}(I \otimes \Pi_{T^n_{p,\delta}})(|\psi\rangle)^{\otimes n} = 1 - \Pr_{x \sim p^{\otimes n}}[x \notin T^n_{p,\delta}] \tag{64}$$

$$\geqslant 1 - 2^{-n\left(\delta - \frac{d \log n}{n}\right)} \tag{65}$$

so it is overwhelmingly likely to succeed for large enough n. This implies that the post-measurement state (a mixed state) has high fidelity with the original state, by the Gentle Measurement Lemma described in Section A.

On the other hand, the dimension of the output space is

$$|T^n_{p,\delta}| \leqslant 2^{n\left(H(p) + \delta' - d\frac{\log(n+1)}{n}\right)} \tag{66}$$

by Eq. (47). Using Eq. (61) together with the definition of entanglement entropy, we may conclude that there is a unitary operation which maps the output register to

$$\left\lceil n\left(E(|\psi\rangle) + \delta' - d\frac{\log(n+1)}{n}\right)\right\rceil \tag{67}$$

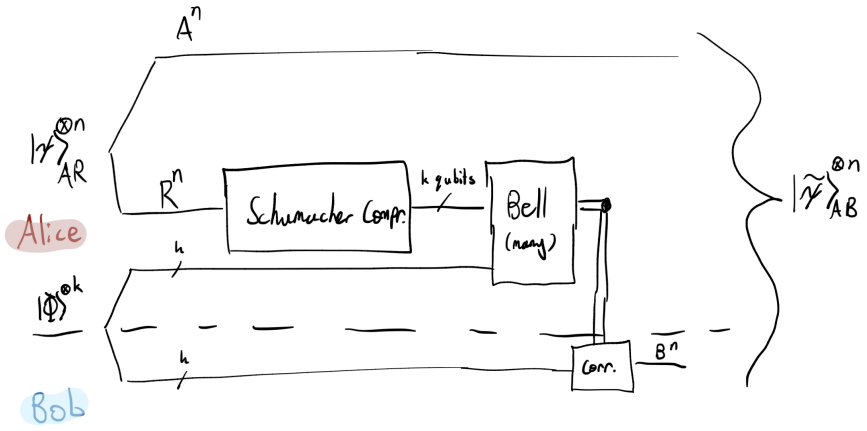


Figure 7: Entanglement dilution.

qubits, and an isometry (ancilla + unitary) operation that takes us back to a state having high fidelity with the original. So $E(|\psi\rangle)$ *is an achievable rate for the task of Schumacher compression*.

Entanglement dilution. We have all the ingredients we need to describe a procedure for entanglement dilution that succeeds with high probability. The trick is to combine Schumacher compression with quantum teleportation, as shown in Fig. 7.

Entanglement distillation. Distillation is also straightforward to describe, given our previous discussion on shared *randomness* distillation. Define the *type class* measurement

$$\Pi_{C(t)} \coloneqq \sum_{x \in C(t)} |x\rangle\langle x| \,. \tag{68}$$

This means, for example,

$$\Pi_{\mathcal{T}^n_{p,\delta}} = \sum_{t : D(t \,\|\, p) \leqslant \delta} \Pi_{C(t)}. \tag{69}$$

Then, beginning with the state

$$|\psi\rangle_{AB}^{\otimes n} = \sum_{x \in [d]^n} \sqrt{p^{\otimes n}(x)}\, |x\rangle \otimes |x\rangle \tag{70}$$

we first measure the n copies of A using the typical subspace measurement $\{\Pi_{\mathcal{T}^n_{p,\delta}}, I - \Pi_{\mathcal{T}^n_{p,\delta}}\}$ as in Schumacher compression. By the same reasoning as in Schumacher compression (using the Gentle Measurement Lemma), the post-measurement mixed state is arbitrarily close to one in fidelity with $|\psi\rangle^{\otimes n}$, so we can carry out the rest of the analysis *as if we are still acting on* n *copies of* $|\psi\rangle$*, but with the additional knowledge that any* t *we observe*

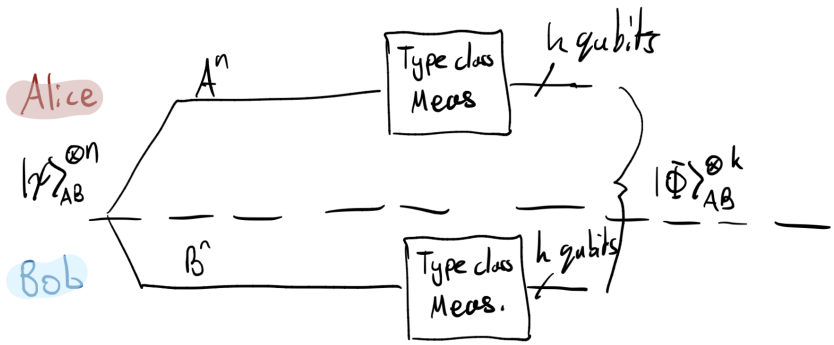


Figure 8: Entanglement distillation.

is typical. Now, we apply the type class measurement $\{\Pi_{C(t)}\}_t$ on the n copies of A. The post-measurement state is proportional to

$$(\Pi_{C(t)} \otimes I)\,|\psi\rangle^{\otimes n} \propto \sum_{x \in C(t)} \sqrt{p^{\otimes n}(x)}\,|x\rangle \otimes |x\rangle \tag{71}$$

$$\propto \sum_{x \in C(t)} |x\rangle \otimes |x\rangle \tag{72}$$

where the second line follows from Eq. (32); specifically, the statement that every string in a type class has the same probability. This gives us a maximally entangled state across two systems of dimension

$$|C(t)| \geqslant 2^{n(H(p)-\delta'-d\frac{\log(n+1)}{n})} \tag{73}$$

where the inequality comes from Eq. (51). Combining with the observation $H(p) = E(|\psi\rangle)$, we can produce a state having arbitrarily high fidelity with

$$\left\lfloor n(E(|\psi\rangle) - \delta' - d\tfrac{\log(n+1)}{n}) \right\rfloor \tag{74}$$

EPR pairs, so $E(|\psi\rangle)$ *is an achievable rate for entanglement distillation*, as claimed. This is depicted in Fig. 8.

6 Conclusion

We have seen that entanglement is a resource, and that the entanglement entropy can be used to quantify this resource for bipartite pure states. Moreover, this quantity has the appealing feature of corresponding to the optimal rate of conversion to and from EPR pairs, i.e., rates for the tasks of entanglement distillation and dilution, in the asymptotic setting.

Now, let me leave you with some questions to ponder if you have made it this far:

- Is shared randomness a resource too? If so, can it help with quantum tasks?
- Why does the distillation protocol we described look so different from the dilution protocol? Can we make it look more similar, using teleportation once again? To do this, we probably need to "go the other way" and use error correction in place of Schumacher compression. Does this work for distilling entanglement from mixed states too? (See [BDSW96].)
- In general, how do we measure mixed state entanglement? This is important for quantifying the monogamy of entanglement phenomenon mentioned in the introduction, for example.

Typography

For text I am using the Pazo Math font via the `mathpazo` package with small caps and linespread set to 1.025, and for math I am using Euler fonts via the `eulervm` package.

A Gentle Measurement Lemma

Here is a version of Winter's Gentle Measurement Lemma [Win99] for projective measurements[8].

Lemma A.1. *Let ρ be a quantum state and $0 \preceq \Pi \preceq I$ be an orthogonal projection operator such that $\mathrm{tr}(\Pi\rho) \geqslant 1-\delta$. Let ρ' be the post-measurement state, i.e.,*

$$\rho' = \Pi\rho\Pi + (I-\Pi)\rho(I-\Pi). \tag{75}$$

It holds that

$$F(\rho,\rho') \geqslant 1-2\sqrt{\delta}. \tag{76}$$

Proof. The Fuchs-van de Graaf (FvDG) inequalities are

$$1-F \leqslant T \leqslant \sqrt{1-F^2} \tag{77}$$

where T is the trace distance. By convexity of the trace distance, we have

$$T(\rho,\rho') \leqslant \delta T\left(\rho, \frac{(I-\Pi)\rho(I-\Pi)}{\mathrm{tr}((I-\Pi)\rho)}\right) + T\left(\rho, \frac{\Pi\rho\Pi}{\mathrm{tr}(\Pi\rho)}\right) \tag{78}$$

$$\leqslant \delta + \sqrt{1-F(\rho,\sigma)^2} \tag{79}$$

[8] It seems Winter called them "tender" measurements, originally.

where in the second line we set $\sigma = \Pi\rho\Pi/\operatorname{tr}(\Pi\rho)$. Now,

$$F(\rho,\sigma) = \frac{1}{\sqrt{\operatorname{tr}(\Pi\rho)}}\left\|\sqrt{\rho}\sqrt{\Pi\rho\Pi}\right\|_1 \tag{80}$$

$$= \frac{1}{\sqrt{\operatorname{tr}(\Pi\rho)}}\operatorname{tr}\left(\sqrt{\sqrt{\rho}\Pi\rho\Pi\sqrt{\rho}}\right) \tag{81}$$

$$= \frac{1}{\sqrt{\operatorname{tr}(\Pi\rho)}}\operatorname{tr}(\Pi\rho) \tag{82}$$

$$\geqslant \sqrt{1-\delta}. \tag{83}$$

Therefore,

$$T(\rho,\rho') \leqslant \delta + \sqrt{\delta} \tag{84}$$

$$\leqslant 2\sqrt{\delta} \tag{85}$$

which completes the proof by (FvDG). □

References

[AAG22] Anurag Anshu, Itai Arad, and David Gosset. An area law for 2d frustration-free spin systems. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC '22, page 12–18. ACM, June 2022. URL: http://dx.doi.org/10.1145/3519935.3519962, doi:10.1145/3519935.3519962.

[BBC+93] Charles H. Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K. Wootters. Teleporting an unknown quantum state via dual classical and einstein-podolsky-rosen channels. *Phys. Rev. Lett.*, 70:1895–1899, Mar 1993. URL: https://link.aps.org/doi/10.1103/PhysRevLett.70.1895, doi:10.1103/PhysRevLett.70.1895.

[BDSW96] Charles H. Bennett, David P. DiVincenzo, John A. Smolin, and William K. Wootters. Mixed-state entanglement and quantum error correction. *Physical Review A*, 54(5):3824–3851, November 1996. URL: http://dx.doi.org/10.1103/PhysRevA.54.3824, doi:10.1103/physreva.54.3824.

[BLMT25] Ainesh Bakshi, Allen Liu, Ankur Moitra, and Ewin Tang. High-temperature gibbs states are unentangled and efficiently preparable, 2025. URL: https://arxiv.org/abs/2403.16850, arXiv:2403.16850.

[BW92] Charles H. Bennett and Stephen J. Wiesner. Communication via one- and two-particle operators on einstein-podolsky-rosen states. *Phys. Rev. Lett.*, 69:2881–2884, Nov 1992. URL: https://link.aps.org/doi/10.1103/PhysRevLett.69.2881, doi:10.1103/PhysRevLett.69.2881.

[CT05] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, April 2005. URL: http://dx.doi.org/10.1002/047174882X, doi:10.1002/047174882x.

[DHW04] Igor Devetak, Aram W. Harrow, and Andreas Winter. A family of quantum protocols. *Phys. Rev. Lett.*, 93:230504, Dec 2004. URL: https://link.aps.org/doi/10.1103/PhysRevLett.93.230504, doi:10.1103/PhysRevLett.93.230504.

[Has07] M B Hastings. An area law for one-dimensional quantum systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(08):P08024–P08024, August 2007. URL: http://dx.doi.org/10.1088/1742-5468/2007/08/P08024, doi:10.1088/1742-5468/2007/08/p08024.

[HLW06] Patrick Hayden, Debbie W. Leung, and Andreas Winter. Aspects of generic entanglement. *Communications in Mathematical Physics*, 265(1):95–117, March 2006. URL: http://dx.doi.org/10.1007/s00220-006-1535-6, doi:10.1007/s00220-006-1535-6.

[HY25] Masahito Hayashi and Hayata Yamasaki. The generalized quantum stein's lemma and the second law of quantum resource theories. *Nature Physics*, 21(12):1988–1993, October 2025. URL: http://dx.doi.org/10.1038/s41567-025-03047-9, doi:10.1038/s41567-025-03047-9.

[Lam25] Ludovico Lami. A solution of the generalized quantum stein's lemma. *IEEE Transactions on Information Theory*, 71(6):4454–4484, 2025. URL: http://dx.doi.org/10.1109/TIT.2025.3543610, doi:10.1109/tit.2025.3543610.

[NC10] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010.

[Nie99] M. A. Nielsen. Conditions for a class of entanglement transformations. *Phys. Rev. Lett.*, 83:436–439, Jul 1999. URL: https://link.aps.org/doi/10.1103/PhysRevLett.83.436, doi:10.1103/PhysRevLett.83.436.

[Sch95] Benjamin Schumacher. Quantum coding. *Phys. Rev. A*, 51:2738–2747, Apr 1995. URL: https://link.aps.org/doi/10.1103/PhysRevA.51.2738, doi:10.1103/PhysRevA.51.2738.

[TA14] Géza Tóth and Iagoba Apellaniz. Quantum metrology from a quantum information science perspective. *Journal of Physics A: Mathematical and Theoretical*, 47(42):424006, October 2014. URL: http://dx.doi.org/10.1088/1751-8113/47/42/424006, doi:10.1088/1751-8113/47/42/424006.

[Wik26] Wikipedia contributors. Pinsker's inequality, 2026. URL: https://en.wikipedia.org/wiki/Pinsker%27s_inequality.

[Win99] A. Winter. Coding theorem and strong converse for quantum channels. *IEEE Transactions on Information Theory*, 45(7):2481–2485, 1999. doi:10.1109/18.796385.