Sevag Gharibian

Lecture 6: Entropy and Entanglement Distillation

> *"Sea water is rendered potable by evaporation; wine and other liquids can be submitted to the same process, for, after having been converted into vapours, they can be condensed back into liquids."*
> — Aristotle (writing about distillation)

The theme of the last two lectures has been of a quantum information theoretic nature — we have studied cloning (or rather, lack thereof), entanglement, and non-local correlations. Before progressing to our next main theme of quantum algorithms, we now give a brief taste of more advanced ideas in quantum information. In the process, we will continue getting used to working with quantum states in both the state vector and density operator formalisms.

The main questions we ask in this lecture are the following:

- *How can we quantify the "amount" of entanglement in a composite quantum system?*
- *Under what conditions can "less entangled" states be "converted" to "more entangled" states?*

The first question will require the foundational concept of *entropy*, introduced in the context of classical information theory by Claude Shannon in 1948. The entropy is worthy of a lecture in itself, being an extremely important quantity. The second question above is tied to the discovery of *entanglement distillation*, in which, similar to the age-old process of distilling potable water from salt water (or more fittingly for our analogy, "pure" water from "dirty" water), one can "distill" pure entanglement from noisy entanglement.

1 Entropy

One of the most influential scientific contributions of the 20th century was the 1948 paper of Claude Shannon, "A Mathematical Theory of Communication", which single-handed founded the field of information theory. Roughly, the aim of information theory is to study information transmission and compression. For this, Shannon introduced the notion of *entropy*, which intuitively quantifies "how random" a data source is, or the "average information content" of the source. It turns out that a *quantum* generalization of entropy will be vital to quantifying entanglement; as such, we begin by defining and motivating the classical *Shannon entropy*.

1.1 Shannon entropy

Let X be a discrete random variable taking values from set $\{x_1, \ldots, x_n\}$, where $\Pr(x_i) := \Pr(X = x_i)$ denotes the probability that X takes value x_i. Then, the Shannon entropy $H(X)$ is defined as

$$H(X) = \sum_{i=1}^{n} -\Pr(x_i) \log \left(\Pr(x_i)\right). \tag{1}$$

Here, the logarithm is taken base 2, and we define $0 \cdot \log 0 = 0$.

Motivation. Before getting our hands dirty understanding $H(x)$, let us step back and motivate it via data compression. Suppose we have a data source whose output we wish to transmit from (say) Germany to Canada. Naturally, we may wish to first *compress* the data, so that we need to transmit as few bits as possible between the two countries. Furthermore, a compression scheme is useless unless we are later able to *recover* or *uncompress* the data in Canada. This raises the natural question: *How few bits can one transmit,*

so as to ensure recovery of the data on the receiving end? Remarkably, Shannon's noiseless coding theorem says that this quantity is given by the entropy! Roughly, the theorem says that in order to reliably transmit N i.i.d. (independently and identically distributed) random variables X_i from a random source X, it is necessary and sufficient to instead send $H(X)$ bits of communication.

Getting our hands dirty. We now explore the sense in which $H(X)$ indeed quantifies the "randomness" or "uncertainty" of X by considering two boundary cases. In the first boundary case, X models a fair coin flip, i.e. X takes value HEADS or TAILS with probability 1/2 each. Then,

$$H(X) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = \frac{1}{2} + \frac{1}{2} = 1. \tag{2}$$

Therefore, we interpret a fair coin as encoding, on average, *one* bit of information. Alternatively, in the information transmission setting, we would need to transmit a single bit to convey the answer of the coin flip from Germany to Canada. This intuitively makes sense — since the outcome of the coin flip is completely random, there is no way to *guess* its outcome in Canada with success probability greater than 1/2 (i.e. a random guess).

The second boundary case is treated in the exercise below.

Exercise. Suppose random variable Y models a biased coin, e.g. takes value HEADS and TAILS with probability 1 and 0, respectively. What is $H(Y)$?

In this example, there is no "uncertainty"; we know the outcome will be HEADS. Thus, in the communication setting, one can interpret this as saying *zero* bits of communication are required to transmit the outcome of the coin flip from Germany to Canada (assuming both Germany and Canada know the probabilities of obtaining HEADS and TAILS beforehand). Indeed, the answer to the exercise above is $H(Y) = 0$.

Exercise. Let random variable Z take values in set $\{0, 1, 2\}$ with probabilities $\{1/4, 1/4, 1/2\}$, respectively. What is $H(Z)$?

Detour: Deriving the formula for entropy. The entropy formula is odd-looking; to understand how it arises, the key observation is the intuition behind the coin flip examples, which says that "when an event is *less* likely to happen, it reveals *more* information". To capture this intuition, Shannon started with a formula for *information content* $I(x_i)$, which for any possible event x_i for random variable X, is given by

$$I(x_i) = \log\left(\frac{1}{\Pr(x_i)}\right) = -\log\left(\Pr(x_i)\right). \tag{3}$$

Since the log function is strictly monotonically increasing (i.e. $I(x) > I(y)$ if $x > y$ for $x, y \in (0, \infty)$), it holds that $I(x_i)$ captures the idea that "rare events yield more information". But $I(x)$ also has three other important properties we expect of an "information measure"; here are the first two:

1. (Information is non-negative) $I(x) \geq 0$, and
2. (If an event occurs with certainty, said occurrence conveys no information) if $\Pr(x) = 1$, then $I(x) = 0$.

For the third important property, we ask — why did we use the log function? Why not any other monotonically increasing function satisfying properties (1) and (2) above? Recall that, by definition, two random variables X and Y are *independent* if

$$\Pr(X = x \text{ and } Y = y) = \Pr(X = x)\Pr(Y = y). \tag{4}$$

Moreover, if X and Y are independent, then intuitively one expects the information conveyed by the joint event $z := (X = x \text{ and } Y = y)$ to be *additive*, i.e. $I(z) = I(x) + I(y)$. But this is precisely what the information content $I(x)$ satisfies, due to its use of the log function, as you will now verify.

Exercise. Let X and Y be independent random variables. Then, for $z := (X = x \text{ and } Y = y)$, express $I(z)$ in terms of $I(x)$ and $I(y)$.

We can now use the information content to derive the formula for entropy — $H(X)$ is simply the *expected* information content over all possible events $\{x_1, \ldots, x_n\}$. (Recall here that for random variable X taking values $x \in \{x_i\}$, its expectation $E[x]$ is given by $E[x] = \sum_i \Pr(x_i) \cdot x_i$.) This is precisely why at the start of this section, we referred to $H(x)$ as the *average* information content of a data source.

1.2 Von Neumann Entropy

Recall the first aim of this lecture was to use entropy to measure entanglement. For this, we shall require a quantum generalization of the Shannon entropy $H(X)$, denoted the *von Neumann* entropy $S(\rho)$, for density operator ρ. To motivate this definition, let us recall the "hierarchy of matrix classes" we introduced in discussing measurements:

- Hermitian operators, $\text{Herm}\left(\mathbb{C}^d\right)$, which generalize the real numbers.
- Positive semidefinite operators, $\text{Pos}\left(\mathbb{C}^d\right)$, which generalize the non-negative real numbers.
- Orthogonal projection operators, $\Pi(\mathbb{C}^d)$, which generalize the set $\{0, 1\}$.

Note that $\Pi(\mathbb{C}^d) \subseteq \text{Pos}\left(\mathbb{C}^d\right) \subseteq \text{Herm}\left(\mathbb{C}^d\right)$, and that the notion of "generalize" above means the *eigenvalues* of the operators fall into the respective class the operators generalize. (For example, matrices in $\text{Pos}\left(\mathbb{C}^d\right)$ have non-negative eigenvalues.) Applying this same interpretation to the set of *density operators* acting on $\mathbb{C}^d$, $\text{D}\left(\mathbb{C}^d\right)$, we thus have that density operators generalize the notion of a *probability distribution*. Indeed, any probability distribution can be embedded into a diagonal density matrix, as you will now show.

Exercise. Let $\{p_i\}_{i=1}^d$ denote a probability distribution. Define diagonal matrix $\rho \in \mathcal{L}(\mathbb{C}^d)$ such that $\rho(i, i) = p_i$. Show that ρ is a density matrix.

Since the eigenvalues $\lambda_i(\rho)$ of a density operator $\rho \in \text{D}\left(\mathbb{C}^d\right)$ form a probability distribution, the natural approach for defining a *quantum* entropy formula is to apply the classical Shannon entropy to the spectrum of ρ:

$$S(\rho) := H\left(\{\lambda_i(\rho)\}_{i=1}^d\right) = \sum_{i=1}^{d} -\lambda_i(\rho) \log\left(\lambda_i(\rho)\right). \tag{5}$$

Operator functions. It is important to pause now and take stock of what we have done in defining $S(\rho)$ in Equation (5): In order to apply a function $f : \mathbb{R} \mapsto \mathbb{R}$ to a Hermitian matrix $H \in \text{Herm}\left(\mathbb{C}^d\right)$, we instead applied f to the *eigenvalues* of H. Why does this "work"? Let us look at the Taylor series expansion of f, which for e.g. $f = e^x$ is (the series converges for all x)

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots. \tag{6}$$

This suggests an idea — to define e^H, perhaps we can substitute H in the right hand side of the Taylor series expansion of e^x:

$$e^H := I + H + \frac{H^2}{2!} + \frac{H^3}{3!} + \cdots. \tag{7}$$

Indeed, this leads to our desired definition; that to generalize the function $f(x) = e^x$ to Hermitian matrices, we apply f to the eigenvalues of H, as you will now show.

Exercise. Let H have spectral decomposition $H = \sum_i \lambda_i |\lambda_i\rangle\langle\lambda_i|$. Show that in Equation (7),

$$e^H = \sum_i e^{\lambda_i} |\lambda_i\rangle\langle\lambda_i|.$$

This idea of applying functions $f : \mathbb{R} \mapsto \mathbb{R}$ to the eigenvalues of Hermitian operators is used so frequently in quantum information that we give such "generalized f" a name — *operator functions*. In the case of $S(\rho)$, by setting $f(x) = \log x$, we can rewrite Equation (5) as

$$S(\rho) = -\mathrm{Tr}(\rho \log(\rho)). \tag{8}$$

Exercise. Verify that Equations (5) and (8) are equal.

Exercise. Let $f(x) = x^2$. What is $f(X)$, for X the Pauli X operator? Why does this yield the same results as multiplying X by itself via matrix multiplication?

Exercise. Let $f(x) = \sqrt{x}$. For any pure state $|\psi\rangle \in \mathbb{C}^d$, define rank one density operator $\rho = |\psi\rangle\langle\psi|$. What is $\sqrt{\rho}$?

Exercise. What is $\sqrt{Z}$ for Z the Pauli Z operator? Is it uniquely defined?

1.2.1 Properties of the von Neumann entropy

Let us now see how properties of $H(X)$ carry over to $S(\rho)$. These will prove crucial in our understanding of quantifying entanglement shortly.

1. *When does a quantum state have no entropy?*

   Recall in our biased coin flip example that if an outcome occurs with probability 1 in our distribution, then $H(X) = 0$. Quantumly, the analogue of this statement is that $S(\rho) = 0$ if and only if ρ is a *pure* state, i.e. $\rho = |\psi\rangle\langle\psi|$ for some $|\psi\rangle \in \mathbb{C}^d$. This is because a recall a pure state is the special case of a mixed state in which one of the states in the preparation procedure is picked with certainty.

   Exercise. Prove that for any pure state $|\psi\rangle$, $S(|\psi\rangle\langle\psi|) = 0$.

2. *When does a quantum state have maximum entropy?*

   We saw that when X represents a fair coin flip, $H(X) = 1$. This is, in fact, the *unique* distribution maximizing H. Applying this directly to the definition of $S(\rho)$, we find that $S(\rho)$ is maximized over all $\rho \in \mathrm{D}\left(\mathbb{C}^2\right)$ if and only if both eigenvalues of ρ are $1/2$. This implies that $\rho = I/2$. Moreover, this statement generalizes to any dimension $d \geq 2$ — for $\rho \in \mathrm{D}\left(\mathbb{C}^d\right)$, $S(\rho)$ is maximized if and only if $\rho = I/d$.

   Exercise. For $\rho = I/d$, what is $S(\rho)$?

3. *Quantum information is non-negative.*

   Since $H(X) \geq 0$, it immediately follows by definition that $S(\rho) \geq 0$.

4. *What is the quantum analogue of independent probability distributions X and Y?*

   Recall that in defining information content, the log function was chosen so as to ensure information is additive when two random variables X and Y are independently distributed. The quantum analogue of this has a natural expression: Let $\rho, \sigma \in \mathrm{D}\left(\mathbb{C}^d\right)$ be density matrices. Then, ρ and σ are independent if their joint state is $\rho \otimes \sigma$. Below, you will prove that this indeed preserves our desired additivity property of information for independent quantum states.

Exercise. Prove that $S(\rho \otimes \sigma) = S(\rho) + S(\sigma)$.

There are other important properties of $S(\rho)$ which we do not wish to focus on at present; for completeness, however, let us briefly mention two more: (1) For arbitrary, possibly entangled, bipartite mixed states ρ_{AB}, $S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B)$ (subadditivity), and (2) $S(\sum_{i=1}^n p_i \rho_i) \geq \sum_{i=1}^p p_i S(\rho_i)$ for $\{p_i\}_{i=1}^n$ a distribution (concavity). Here, and henceforth in this course, we use the shorthand

$$\rho_A := \mathrm{Tr}_B(\rho_{AB}). \tag{9}$$

2 Quantifying entanglement in composite quantum systems

With the notion of entropy in hand, we return to the following fundamental question. Let $\rho_{AB} \in \mathrm{D}\left(\mathbb{C}^d \otimes \mathbb{C}^d\right)$ be a bipartite quantum state. Can one efficiently determine if ρ_{AB} is entangled? (Recall this means that ρ_{AB} cannot be written $\rho_{AB} = \sum_i p_i \rho_{A,i} \otimes \rho_{B,i}$ as a convex combination of product states.) Roughly, if one uses d to encode the size of the input (i.e. the input is the entire $d^2 \times d^2$ matrix representing ρ_{AB}), then deciding this question turns out to be NP-hard. This directly implies that quantifying "how much" entanglement is in ρ_{AB} is also NP-hard. However, there is a special case in which we *can* do both tasks efficiently — the case of bipartite *pure* states $|\psi_{AB}\rangle \in \mathbb{C}^d \otimes \mathbb{C}^d$. It is here in which the von Neumann entropy plays a role.

In fact, a previous lecture already discussed an efficient test for entanglement for $|\psi_{AB}\rangle$ — the latter is entangled if and only if

$$\rho_A := \mathrm{Tr}_B(|\psi_{AB}\rangle\langle\psi_{AB}|) \tag{10}$$

has rank at least 2. This, in turn, followed because it immediately implies the Schmidt rank of $|\psi_{AB}\rangle$ is at least 2. However, we can say more. Suppose we have Schmidt decomposition

$$|\psi_{AB}\rangle = \sum_{i=1}^d s_i |a_i\rangle|b_i\rangle. \tag{11}$$

Then, intuitively, as with the example of a Bell pair, $|\psi_{AB}\rangle$ is "highly entangled" if all the Schmidt coefficients s_i are approximately equal in magnitude, and $|\psi_{AB}\rangle$ is "weakly entangled" if there exists a single s_i whose magnitude is approximately 1. Do we know of a function which quantifies precisely this sort of behavior on the set $\{s_i\}$? Indeed, the entropy function! This notion of s_i being "spread out" versus "concentrated" is highly reminiscent of our fair versus biased coin flip example for the Shannon entropy. We can therefore use the von Neumann entropy to define an entanglement measure $E(|\psi_{AB}\rangle)$ as

$$E(|\psi_{AB}\rangle) := S(\rho_A). \tag{12}$$

Exercise. What is $E(|\Phi^+_{AB}\rangle)$ for $|\Phi^+_{AB}\rangle = \frac{1}{2}(|00\rangle + |11\rangle)$ a Bell state?

Exercise. What is $E(|+\rangle_A|-\rangle_B)$?

Exercise. Unlike the example of the fair coin, the Schmidt coefficients s_i of $|\psi_{AB}\rangle$ are not probabilities, but amplitudes (i.e. we do not have $\sum_i s_i = 1$, but rather $\sum_i s_i^2 = 1$). Show, however, that the notion of probabilities is recovered in the formula $S(\rho_A)$, i.e. show that the eigenvalues of ρ_A are the precisely set $\left\{s_i^2\right\}_{i=1}^d$, which *do* form a distribution.

Finally, let us close this section with a natural question — does $E(|\psi_{AB}\rangle)$ still measure entanglement when its input is allowed to be a mixed state ρ_{AB} (as opposed to a pure state $|\psi_{AB}\rangle$)? The answer is given in the following exercise.

Exercise. Define $E(\rho_{AB}) := S(\mathrm{Tr}_B(\rho_{AB})) = S(\rho_A)$. Recall that the maximally mixed state on two qubits is a product state, i.e. $I/4 = I/2 \otimes I/2$. Show that $E(I/4) = 1$. Why does this imply when cannot use E as an entanglement measure for bipartite mixed states?

3 Entanglement distillation

Now that we have a notion of how to quantify entanglement in pure states, we can become greedy — under what circumstances is it possible to "increase" the amount of entanglement in a composite quantum system? This is a highly non-trivial question, as fundamental communication tasks such as teleportation require highly entangled Bell pairs as a resource. Unfortunately, experimentally producing such pure states is generally a difficult task due to noise from the environment. In other words, in a lab one is typically able to produce *mixed* states, as opposed to pure states. Moreover, *even if* Alice could produce perfect Bell pairs in a lab on Earth, when she sends half of a Bell pair to Bob on Mars, the transmitted qubit will again typically be subject to noise, yielding a shared mixed state ρ_{AB} between Alice and Bob. Do Alice and Bob have any hope of running the teleportation protocol given ρ_{AB}?

Local Operations and Classical Communication (LOCC). To answer the question, it is important to first define the rules of the game. Since Alice and Bob are spatially separated, they are not able to apply joint quantum operations to both systems A and B, e.g. they cannot apply a non-factorizable unitary $U_{AB} \in \mathrm{U}\left(\mathbb{C}^d \otimes \mathbb{C}^d\right)$ to ρ_{AB}. However, they *can* apply local unitaries and measurements, e.g. factorizable unitaries of the form $U_A \otimes U_B$ for $U_A, U_B \in \mathrm{U}\left(\mathbb{C}^d\right)$ (i.e. Alice locally applies U_A, Bob locally applies U_B). They can also pick up the phone and call one another to transmit classical information. Taken together, this set of allowed operations is given a name: Local Operations and Classical Communication (LOCC). The question is thus: *Given a shared mixed state ρ_{AB}, can Alice and Bob use LOCC to "purify" or "distill" Bell states out of ρ_{AB}?* The answer is sometimes *yes*, and protocols accomplishing this are called *distillation protocols*, as they recover "pure" entanglement from "noisy" or mixed state entanglement.

A simple distillation protocol. We shall discuss a simple distillation protocol, known as the *recurrence protocol*. Given as input a mixed two-qubit state $\rho_{AB} \in \mathrm{D}\left(\mathbb{C}^2 \otimes \mathbb{C}^2\right)$, our aim is to distill the Bell state known as the *singlet*, $|\Psi^-\rangle = (|01\rangle - |10\rangle/\sqrt{2}$; note $I \otimes Y|\Psi^-\rangle = i|\Phi^+\rangle$, making it easy to convert one Bell state into the other for the teleportation scheme. There is a required precondition for the protocol to work — the input state ρ_{AB} must have sufficient initial overlap with $|\Psi^-\rangle$, i.e.

$$F(\rho_{AB}) := \langle\Psi^-|\rho_{AB}|\Psi^-\rangle > \frac{1}{2}. \tag{13}$$

In other words, in transmitting half of $|\Psi^-\rangle$ from Alice to Bob, the resulting mixed state ρ_{AB} should not have deviated "too far" from $|\Psi^-\rangle$. Henceforth, we shall use shorthand F to denote $F(\rho_{AB})$ for brevity.

Suppose Alice and Bob share two copies of ρ_{AB}; let us label them $\rho_{A_1B_1}$ and $\rho_{A_2B_2}$, where Alice holds systems A_1, A_2, and Bob holds B_1, B_2. Each round of the distillation protocol proceeds as follows.

1. (Twirling operation) Alice picks a Pauli operator U from set $\{I, X, Y, Z\}$ uniformly at random, and communicates this choice to Bob. They each locally apply operator $\sqrt{U}$ to $\rho_{A_iB_i}$ for $i \in \{1, 2\}$ (note that $\sqrt{U}$ is defined using the notion of operator functions, introduced in Section 1.2), obtaining

   $$\sigma_{A_iB_i} := (\sqrt{U_A} \otimes \sqrt{U_B})\rho_{A_iB_i}(\sqrt{U_A}^\dagger \otimes \sqrt{U_B}^\dagger).$$

   This random choice of Pauli and its subsequent local application is together called the *twirling* map $\Phi : \mathrm{D}\left(\mathbb{C}^2 \otimes \mathbb{C}^2\right) \mapsto \mathrm{D}\left(\mathbb{C}^2 \otimes \mathbb{C}^2\right)$, and is not a unitary map (due to the random choice over Pauli operators); nevertheless, it can clearly be implemented given the ability to flip random coins and apply arbitrary single qubit gates. (The formal framework for studying such operations is via *Trace Preserving Completely Positive Maps*, and is beyond the scope of this course.) The nice thing about the twirling operation is that, for any input ρ_{AB}, $\Phi(\rho_{AB})$ can be diagonalized in the Bell basis, i.e. can be written

   $$\Phi(\rho_{AB}) = F|\Psi^-\rangle\langle\Psi^-| + \frac{1-F}{3}|\Psi^+\rangle\langle\Psi^+| + \frac{1-F}{3}|\Phi^+\rangle\langle\Phi^+| + \frac{1-F}{3}|\Phi^-\rangle\langle\Phi^-| \tag{14}$$

   for Bell basis $\{|\Phi^+\rangle, |\Phi^-\rangle, |\Psi^+\rangle, |\Psi^-\rangle\}$, where F is the same from Equation (13).

2. (Convert from $|\Psi^-\rangle$ to $|\Phi^+\rangle$) Alice applies Pauli Y to her half of each state to obtain states:

$$\sigma_{A_iB_i} = \frac{1-F}{3}|\Psi^-\rangle\langle\Psi^-| + \frac{1-F}{3}|\Psi^+\rangle\langle\Psi^+| + F|\Phi^+\rangle\langle\Phi^+| + \frac{1-F}{3}|\Phi^-\rangle\langle\Phi^-| \tag{15}$$

   This shifts most of the weight in $\Phi(\rho_{A_iB_i})$ from $|\Psi^-\rangle$ to $|\Phi^+\rangle$, since $F > 1/2$ by Equation (13).

3. (Application of CNOT gates) Alice applies a CNOT gate with qubit A_1 as the control and A_2 as the target. Bob does the same for B_1 and B_2.

4. (Local measurements) Alice and Bob each locally measure A_2 and B_2 in the standard basis, obtaining outcomes a and b in $\{0, 1\}$, respectively. They pick up the phone to compare their measurement results a and b. If $a = b$, they keep the remaining composite system on AB, denoted $\sigma'_{A_1B_1}$. Otherwise if $a \neq b$, they throw out all systems and start again.

5. (Convert from $|\Phi^+\rangle$ to $|\Psi^-\rangle$) Alice applies Pauli Y to her half of $\sigma'_{A_1B_1}$ to convert its $|\Phi^+\rangle$ component back to $|\Psi^-\rangle$.

To get a better sense of this protocol in action, let us apply it to a concrete example. Suppose Alice sends half of the singlet to Bob, and along the way, the state $|\Psi^-\rangle$ is injected with some completely random noise, denoted by the identity matrix:

$$\rho_{AB} = \frac{1}{2}|\Psi^-\rangle\langle\Psi^-| + \frac{1}{8}I = \frac{3}{4}|\Psi^-\rangle\langle\Psi^-| + \frac{1}{12}|\Psi^+\rangle\langle\Psi^+| + \frac{1}{12}|\Phi^+\rangle\langle\Phi^+| + \frac{1}{12}|\Phi^-\rangle\langle\Phi^-|, \tag{16}$$

where the second equality follows by recalling the identity matrix diagonalizes in any basis, including the Bell basis. (The noise-inducing channel above is formally known as the *depolarizing channel* in quantum information theory.)

Exercise. Show that $\sqrt{Z} \otimes \sqrt{Z}$ maps $|\Phi^+\rangle$ to $|\Phi^-\rangle$ and vice versa. Using the additional identities that $\sqrt{X} \otimes \sqrt{X}$ maps $|\Phi^+\rangle$ to $|\Psi^+\rangle$ and vice versa, and $\sqrt{Y} \otimes \sqrt{Y}$ maps $|\Phi^-\rangle$ to $|\Psi^+\rangle$ and vice versa, show that the twirling map leaves ρ_{AB} in Equation (16) invariant.

Exercise. Show that applying $Y_A \otimes I$ to ρ_{AB} yields in Step 2 that

$$\sigma_{AB} = \frac{1}{12}|\Psi^-\rangle\langle\Psi^-| + \frac{1}{12}|\Psi^+\rangle\langle\Psi^+| + \frac{3}{4}|\Phi^+\rangle\langle\Phi^+| + \frac{1}{12}|\Phi^-\rangle\langle\Phi^-|. \tag{17}$$

Finally, Steps 3 and 4 are a bit messier. The intuition here is that the CNOT entangles $\sigma_{A_1B_1}$ with $\sigma_{A_2B_2}$, and the measurement forces the bits in the second system (formerly in state $\sigma_{A_2B_2}$) to match; via the entanglement just created, this increases the weight in Equation (17) on the terms where bits match, i.e. $|\Phi^+\rangle\langle\Phi^+|$ and $|\Phi^-\rangle\langle\Phi^-|$. Thus, the final Step 5 will yield a state with higher overlap on the desired singlet state $|\Psi^-\rangle$.

To run through the full technical analysis for Steps 3 and 4 would be tedious, so we analyze one term of the computation for brevity. Before Step 3 is run, Alice and Bob share state

$$\sigma_{A_1B_1} \otimes \sigma_{A_2B_2} \in \mathrm{D}\left(\mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \mathbb{C}^2\right), \tag{18}$$

where recall Alice holds qubits A_1, A_2, and Bob holds B_1, B_2. Since each $\sigma_{A_iB_i}$ has 4 terms in its mixture, the tensor product in Equation (18) has 16 terms. By linearity, Step 3 then applies gates $\mathrm{CNOT}_{A_1A_2}$ and $\mathrm{CNOT}_{B_1B_2}$ to each of these 16 terms, where our notational convention is that CNOT_{12} has qubit 1 as the control and qubit 2 as the target. Let us analyze one of these 16 terms: $|\Phi^+\rangle\langle\Phi^+|_{A_1B_1} \otimes |\Phi^+\rangle\langle\Phi^+|_{A_2B_2}$.

Exercise. Show that

$$(\mathrm{CNOT}_{A_1A_2} \otimes \mathrm{CNOT}_{B_1B_2})|\Phi^+\rangle_{A_1B_1} \otimes |\Phi^+\rangle_{A_2B_2} = \frac{1}{2}\left(|0000\rangle + |0011\rangle + |1100\rangle + |1111\rangle\right)_{A_1B_1A_2B_2}.$$

If Alice and Bob run Step 4 on this state and obtain matching outcomes $a = b$, what does the state on qubits A_1B_1 collapse to?

Finally, let us briefly state what this protocol buys us. A careful but tedious analysis yields that with probability at least 1/4, this protocol maps the input state $\rho_{A_1B_1} \otimes \rho_{A_2B_2}$ to an output state $\sigma'_{A_1B_1}$ such that (for $F_\rho := F(\rho_{A_1B_1})$)

$$F(\sigma'_{A_1B_1}) = \frac{F_\rho^2 + \frac{1}{9}(1-F_\rho)^2}{F_\rho^2 + \frac{2}{3}F_\rho(1-F_\rho) + \frac{5}{9}(1-F_\rho)^2}. \tag{19}$$

So long as $F_\rho > 1/2$, one can show $F(\sigma'_{A_1B_1}) > F_\rho$; thus, recursively applying this protocol (using many pairs of input states ρ_{AB}) improves our overlap with our desired target state of $|\Psi^-\rangle$.