LECTURE NOTES ON MATHEMATICAL ASPECTS OF QUANTUM INFORMATION THEORY

DARIO TREVISAN

Contents

Date: August 29, 2023.

1. Introduction

Classical information theory studies the laws of storage and communication of information. As a scientific field it can be collocated at the intersection of probability theory, statistics, computer science, statistical mechanics, information engineering, and electrical engineering, although its birth as an independent field is traditionally set in the 1940s with C. Shannon's work [Sha48]. Understanding the limitations but also possibilities provided by the quantum mechanical aspects of nature is the subject of *quantum information theory*, which has been as an independent research area since the 1990s. The theory naturally makes use of further mathematical tools, in particular functional analysis, as it is strongly based on the postulates of quantum mechanics.

Aim of this course is to give a self-contained introduction to the main mathematical aspects of the theory, focusing on the problem of quantifying how information deteriorates when transmitted through a noisy communication channel, taking into account its quantum mechanical description. This is the content of the quantum coding theorem (Theorem 7.2), which is a counterpart (actually, an extension) of the classical Shannon's fundamental limit established in [Sha48].

We structured the course so that no prior knowledge in classical information theory, nor in quantum mechanics, should be required. Our target audience consists of mathematicians with a background in probability, analysis or mathematical physics. Most of the exposition is borrowed from the various excellent monographs available, in particular [NC02; Hol19; Wil11; BŻ17]. These focus mostly on the *elementary* setting of quantum systems represented by finite-dimensional Hilbert spaces, where operators become, after taking coordinates, just matrices and topological and measure-theoretic considerations can be avoided (to be precise, [Hol19] is however an excellent reference on Gaussian systems). To differentiate a bit, we hint at how some concepts can be also given in the *abstract* setting of C^*-algebras and briefly mention some infinite-dimensional examples, such as quantum spin chains but also quantum Gaussian systems. This is not strictly necessary for understanding the main aspects of quantum information theory, but it may help stimulating further connections with other areas of non-commutative mathematics. The exposition of these concepts is mostly taken from [AF01; Naa13; Mey95]. We try to avoid the formalism of infinite dimensional Hilbert spaces and the use of unbounded operators, which we believe it would require an entire course to be addressed properly: we recommend instead the monograph [Mor19] for a detailed description of its mathematical aspects connected to quantum mechanics.

These notes will follow the exposition given in the course, and each section below should roughly correspond to a lecture. In Section 2, we describe the core mathematical objects of quantum mechanics, i.e., states and measurements (or observables), as operator counterparts of probability distributions and random variables. In Section 3, we introduce the quantum analogues of classical noisy communication channels, which in Shannon's theory are modelled by Markov kernels. We devote Section 4 to discuss some operator inequalities which play a fundamental role in the subsequent analysis, in particular when applied to distance measures between quantum states (Section 5) and quantum (von Neumann) entropy (Section 6). Finally, in Section 7 we prove the quantum coding theorem.

2. Postulates of Quantum Mechanics

Classical physics (think e.g., to the classical mechanics of point particles) describes a system and its time-evolution in terms of three fundamental mathematical objects:

(1) A set Ω (called *phase space*) representing via its elements $\omega \in \Omega$ all the possible *states* of the system. For example, a single particle is described by its position $q \in \mathbb{R}^3$ and momentum $p \in \mathbb{R}^3$, so one defines $\omega = (q, p) \in \Omega = \mathbb{R}^3 \times \mathbb{R}^3$

(2) A family of *observables*, i.e., functions $X : \Omega \to \mathcal{X}$, defined on Ω taking values in a set of possible outcomes $\mathcal{X}$, representing quantities of interest, i.e., (at least in theory) accessible via physical measurements. In the example, position $p = p(\omega)$ and momentum $q(\omega)$ are naturally observables.
(3) A family of transformations of Ω into itself – usually implicitly given through differential equations – that represent the evolution of the system with respect to time. In the example, the classical equations are Newton's laws, here formulated via Hamilton equations.

A similar scheme turns out to be quite useful for modelling purposes, hence it naturally appears in other settings. In *elementary probability theory* – we avoid throughout Kolmogorov axioms for the sake of simplicity – analogues of such three components can be similarly found:

(1) The (finite) set Ω is called *sample space*, and its elements $\omega \in \Omega$ describe all the possible outcomes of a random experiment. In the standard example of throwing a die with 6 faces, one lets $\Omega = \{1, 2, 3, 4, 5, 6\}$. The main difference with the classical physics scheme is that the state[1] can be any *probability distribution* $\rho : \Omega \to [0, 1]$, such that $\sum_{\omega \in \Omega} \rho(\omega) = 1$. Full confidence (almost sure certainty) about a specific $\bar{\omega}$ translates to a Dirac distribution $\rho = \delta_{\bar{\omega}}$, which can be then identified with $\bar{\omega}$ itself.
(2) *Random variables* X on Ω, taking values in a set $\mathcal{X}$, play the role of observables. Usually, one deals first with *events* $V \subseteq \Omega$, which model logical statements (i.e., either true or false) on the outcomes and are naturally associated with *indicator random variables* 1_V, taking values in $\{0, 1\}$.
(3) The theory of stochastic processes provides a way to introduce a family of transformations of Ω into itself, which are usually also of stochastic nature. In order to accommodate such additional randomness, one usually enlarges the sample space to the path space over Ω, but in common situations (for example, in the case of Markov chains) only the description of the marginal probability distributions can be sufficient for applications.

The two theories fruitful meet in (classical) statistical physics: when the phase space Ω is too large (e.g. because of large number of particles) so that the state $\omega \in \Omega$ cannot be known with full certainty, one relaxes the notion of deterministic state by allowing for probability distributions on Ω. Observables (e.g., pressure, volume, temperature etc.) then actually coincide with random variables.

Quantum mechanics is a theory (circa 100 years old) supported by a vast experimental evidence, which provides the most accurate description/prediction of physical phenomena at very small scales (atoms, molecules, light). The theory enjoys a very different status from *classical* ones (such as classical mechanics, but also general relativity theory) because of its inherently probabilistic features: it only provides the odds that some event will occur.

We introduce the postulates of quantum mechanics following the same three-objects scheme outlined above. In this section, we only describe the first two ones (states and observables) leaving the third to the next section, where we address the quantum analogue of Markov kernels. As stated in the introduction, we also limit ourselves to the *elementary* setting of finite dimensional systems – which corresponds to the case of finite sample space Ω – avoiding all the functional-analytic difficulties – which roughly corresponds to the application of measure theory in Kolmogorov axiomatization of probability. We end this section by discussing the abstract approach based on C^*-algebras, which is a quite elegant formalism and allows also to deal with certain infinite dimensional settings.

[1] here representing the state of knowledge of an observer if we agree upon a subjective intepretation of probability theory

2.1. **(Elementary) quantum systems and their states.** An *elementary quantum system* is described by a finite-dimensional complex Hilbert space $(H, \langle\cdot|\cdot\rangle)$.

Following a consolidated convention, we let the scalar product be linear in the second variable and anti-linear in the first variable. The induced norm is written as $\|\cdot\|$. We use throughout Dirac's (*ket*) notation $|\psi\rangle \in H$, so that (*bra*) vectors $\langle\varphi| \in H^*$ denote the linear functionals

$$\langle\varphi| : H \to \mathbb{C}, \qquad |\psi\rangle \mapsto \langle\varphi|\psi\rangle.$$

The (Riesz) correspondence $|\psi\rangle \mapsto \langle\psi|$ provides an anti-linear isomorphism between H and H^*. The advantage of Dirac's notation is that families of vectors over an index set I can be conveniently written simply as $(|i\rangle)_{i\in I}$ (if it does not generate ambiguities).

Example 2.1. If $H = \mathbb{C}^d$, then the standard basis can be conveniently written as $(|i\rangle)_{i=1}^d$. Since in information theory it is customary to count starting from 0, we may also write $(|k\rangle)_{k=0,\dots(d-1)}$ and call it the *computational basis*. The case $d = 2$ provides an example of a two-level (i.e., two dimensional) quantum system, also called a single-*qubit* system, in analogy with the classical bit sample space $\{0, 1\}$.

One is tempted to think of H as the analogue of the sample space Ω, however a more precise description would be in terms of the complex projective space, i.e., by taking equivalence classes of elements $|\psi\rangle$ with respect to multiplication with non-null complex numbers. Still, it is more convenient to keep the linear structure of H and define as *state vector* any vector $|\psi\rangle \in H$ with unit norm, $\langle\psi|\psi\rangle = \|\psi\|^2 = 1$. From a physics perspective, however, $|\psi\rangle$ will be indistinguishable from any multiple $e^{i\theta}|\psi\rangle$ with $\theta \in \mathbb{R}$ (called a *phase*). State vectors $|\psi\rangle$ are often called *wave functions* (or slight improperly *pure states*).

Notice that, even if H is finite dimensional in such elementary setting, the set of state vectors is infinite. In fact, given an orthonormal basis $(|i\rangle)_{i\in I} \subseteq H$ (with I finite) any state vector $|\psi\rangle$ can be represented as a linear combination

$$|\psi\rangle = \sum_{i\in I} \alpha_i |i\rangle$$

where $\alpha_i = \langle i|\psi\rangle \in \mathbb{C}$ are often called amplitudes and satisfy

$$\sum_{i\in I} |\alpha_i|^2 = 1.$$

Such a representation is often referred to as a quantum superposition of the state vectors $(|i\rangle)_{i\in I}$. The quantities $|\alpha_i|^2 = |\langle i|\psi\rangle|^2$ can be interpreted as probabilities (see (2.1)) but one should *not* (at least not only) think of $|\psi\rangle$ as a classical probability distribution over the $|i\rangle$'s with probabilities $|\alpha_i|^2$. This is because while global phases are irrelevant, changing a single phase in an amplitude may yield a completely different state vector.

Instead, we define the quantum analogue of probability distributions over the set of state vectors as certain class of linear operators $\rho : H \to H$, called *density operators*, whose extreme points, called *pure states*, correspond to *state vectors*.

Precisely, given a state vector $|\psi\rangle$, we associate to it the orthogonal projection operator $P_{|\psi\rangle} : H \to H$ on the subspace spanned by $|\psi\rangle$. Dirac's notation is quite useful here, since one can write

$$P_{|\psi\rangle} = |\psi\rangle\langle\psi|.$$

Such operator plays the role of a Dirac δ probability distribution concentrated on $|\psi\rangle$. Notice that, if two state vectors differ by a phase $|\varphi\rangle = e^{i\theta}|\psi\rangle$, the density operators coincide. General *density operators* are then defined as convex combinations of pure states:

$$\rho = \sum_{i\in I} p_i |\psi_i\rangle\langle\psi_i|,$$

where $(|\psi_i\rangle)_{i\in I} \subseteq H$ is any (finite) family of state vectors and $(p_i)_{i\in I} \subseteq [0,1]$, $\sum_{i\in I} p_i = 1$ is a classical probability distribution over the set I.

We can characterize density operators $\rho : H \to H$ as self-adjoint (i.e., Hermitian), positive operators with unit trace[2]. The set of density operators is denoted with $\mathcal{S}(H)$. Using the spectral theorem for self-adjoint operators (in finite dimensions), it is a simple exercise to show that indeed any $\rho \in \mathcal{S}(H)$ can be written as

$$\rho = \sum_{i\in I} p_i \, |i\rangle \langle i| \,,$$

with a classical probability distribution $(p_i)_{i\in I}$ and an orthonormal basis $(|i\rangle)_{i\in I} \subseteq H$. A density operator $\rho \in \mathcal{S}(H)$ is pure (or is a pure state of the system H) if $\rho = |\psi\rangle\langle\psi|$ for some state vector $|\psi\rangle \in H$.

Any fixed orthonormal basis $(|i\rangle)_{i\in I}$ of H (with $|I| = d = \dim(H)$ elements) allows to represent any operator $A : H \to H$ as a matrix $(A_{ij})_{i,j\in I} \in \mathbb{C}^{d\times d}$ with complex entries $A_{ij} = \langle i|Aj\rangle$. For a density operator $\rho \in \mathcal{S}(H)$, we thus obtain a *density matrix* $(\rho_{ij})_{i,j\in I}$ which is hermitian, non-negative and with unit trace. In particular, its diagonal elements $(\rho_{ii})_{i\in I}$ can be used to define a classical probability distribution over I. Thus, we may think of density matrices as non-commutative extensions of classical probability distributions over a set I, which in turn can be identified with diagonal matrices. Notice however that such identification depends on the chosen basis: when $H = \mathbb{C}^d$, this is usually understood with respect to the standard (computational) basis.

2.2. **Measurements and observables.** We next introduce the quantum analogue of functions over a classical phase space, or of random variables on a sample space. Again, in the elementary setting of a finite dimensional H, we assume that the "set of possible values" $\mathcal{X}$ is finite.

In most expositions, observables $A \in \mathcal{O}(H)$ on a quantum system H are straightforwardly defined as self-adjoint operators $A : H \to H$. The spectrum, i.e. the set of eigenvalues $\sigma(A) \subseteq \mathbb{R}$ plays the role of the "set of possible values' of the observable A, which are those physically measured through a (often ideal) device interacting with the quantum system. Although these are by far the most commonly encountered, such observables would correspond only to real-valued random variables. Moreover, it is not clear at first sight why such operators should be the correct analogue.

Following instead a path similar to that of elementary probability theory, we first describe the mathematical objects play the role of events, i.e., of logical propositions about an elementary quantum system H: these are given by the *subspaces* $V < H$. The 0-dimensional subspace $\{0\}$ corresponds to a *false* proposition about the system, the whole $V = H$ corresponds instead to a *true* one. One-dimensional subspaces spanned by a state vector $|\psi\rangle$ can be interpreted as the proposition "*the quantum system is in the state associated to* $|\psi\rangle$". To any $V < H$, we associate its *indicator observable* $\mathbb{1}_V : H \to H$ which is defined as the orthogonal projection operator on V. In particular, it is self-adjoint $\mathbb{1}_V = \mathbb{1}_V^*$ and $\mathbb{1}_V^2 = \mathbb{1}_V$, so that its spectrum is $\sigma(\mathbb{1}_V) = \{0,1\}$ (except in the trivial cases $V = \{0\}$ where $\mathbb{1}_V = 0$ is the null operator, and $V = H$ where $\mathbb{1}_H$ is the identity operator). We thus recover the idea introduced above that the spectrum of an observable plays the role of the "set of possible values".

[2]Let us recall some standard definitions, see Section 3.1 for further notation. We write $\mathcal{L}(H)$ for the set of linear operators $A : H \to H$. The adjoint of an operator $A \in \mathcal{L}(H)$ is the (unique) operator $A^* \in \mathcal{L}(H)$ such that $\langle A\varphi|\psi\rangle = \langle\varphi|A^*\psi\rangle$ for every $|\varphi\rangle, |\psi\rangle \in H$. A is self-adjoint (or Hermitian) if $A = A^*$, and we write $A \in \mathcal{O}(H)$. An operator $A \in \mathcal{L}(H)$ is positive, and we write $A \geq 0$ or $A \in \mathcal{O}_{\geq}(H)$, if $A \in \mathcal{O}(H)$ and $\langle\psi|A\psi\rangle \geq 0$ for every $|\psi\rangle \in H$. If $A \geq 0$ is also invertible, we write $A > 0$ or $A \in \mathcal{O}_{>}(H)$. We write $A \geq B$ or $B \leq A$, if $A - B \geq 0$ and A, $B \in \mathcal{O}(H)$. The trace of an operator $A \in \mathcal{L}(H)$ is defined as $\mathrm{tr}[A] = \sum_{i\in I} \langle i|Ai\rangle$, where $(|i\rangle)_{i\in I}$ denotes any orthonormal basis of H. The trace is linear and cyclic, i.e., $\mathrm{tr}[AB] = \mathrm{tr}[BA]$, and $A \geq 0$ implies $\mathrm{tr}[A] \geq 0$.

We should think of the observable $\mathbb{1}_V$ as associated to a physical device that, when applied to the system, yields outcomes 1 if V holds or 0 if V does not hold, with some probability, according to the state of the system. Precisely, if the state is described by the density operator $\rho \in \mathcal{S}(H)$, we postulate that by *measuring* $\mathbb{1}_V$, the probability of observing that V holds (i.e., we measure 1) is given by *Born's rule*:

$$\mathbb{P}_\rho(V) = \mathbb{P}(\mathbb{1}_V = 1) := \mathrm{tr}[\mathbb{1}_V \rho].$$

To see that this is indeed a probability, i.e. $\mathbb{P}_\rho(V) \in [0,1]$, it is sufficient to argue in the case that $\rho = |\psi\rangle\langle\psi|$ is a pure state associated to a state vector $|\psi\rangle$. Then,

$$\mathbb{P}_{|\psi\rangle}(V) = \langle\psi|\mathbb{1}_V\psi\rangle = \langle\mathbb{1}_V\psi|\mathbb{1}_V\psi\rangle = \|\mathbb{1}_V\psi\|^2 \leq \|\psi\|^2 = 1.$$

having used that $\mathbb{1}_V = \mathbb{1}_V^* = \mathbb{1}_V^2$ is an orthogonal projection. As a further postulate, we require that, *after having measured* $\mathbb{1}_V$ *and observed that* V *holds*, the state of the system H is updated from ρ to the density operator given by the so-called *collapse of wave function*:

$$\rho_V = \frac{\mathbb{1}_V \rho \mathbb{1}_V}{\mathbb{P}_\rho(V)}.$$

This expression should be compared with the rule for conditional probability, i.e.,

$$\mathbb{P}(\cdot|V) = \frac{\mathbb{P}(\cdot \text{ and } V)}{\mathbb{P}(V)}.$$

Remark 2.2. The interpretation of Born's rule and the collapse of the state is subject to many debates, even more than the frequentist vs Bayesian dispute in probability and statistics. The underlying issue is whether probabilities in quantum mechanics represent states of knowledge of a subject about a system or have a deeper, possibly objective, meaning – something one would hope from a physical theory. A restricted but somehow safer interpretation is that quantum states and the derived probabilities describe the relative frequencies in the ideal infinite limit of a repeated sequence of independent experiments in a prepared situation (so that the measurements are classical i.i.d. sequences). Of course such a frequentist-like interpretation seems to be very limited and has the same drawbacks as the interpretation of classical probability as frequency: how can we ensure that the experiments are independent and identical? but it still may guide the intuition.

If V is the linear subspace generated by a state vector $|\varphi\rangle$, then $\mathbb{1}_V = |\varphi\rangle\langle\varphi| \in \mathcal{O}(H)$ is the same projection operator $P_{|\varphi\rangle}$ as the pure state associated to $|\varphi\rangle$ from the previous section. This coincidence is due to our elementary, finite-dimensional, setting: also in elementary probability over a finite sample space Ω, the indicator function $1_{\{\bar\omega\}}$ apparently coincides with the Dirac probability distribution $\delta_{\bar\omega}$. However, recalling a bit of measure theory, the former should be thought as a function, while the latter as a measure (hence belonging to a dual space).

Measuring the indicator observable $\mathbb{1}_V = |\varphi\rangle\langle\varphi| \in \mathcal{O}(H)$ yields therefore outcome 1 if "the quantum system is in the state associated to $|\psi\rangle$", with probability

$$\mathbb{P}_\rho(V) = \mathrm{tr}[\rho\,|\varphi\rangle\langle\varphi|] = \langle\varphi|\rho\varphi\rangle.$$

If $\rho = |\psi\rangle\langle\psi| \in \mathcal{S}(H)$ is the pure state associated to the state vector $|\psi\rangle$, then

$$\mathbb{P}_{|\psi\rangle}(V) = |\langle\varphi|\psi\rangle|^2 \in [0,1]. \tag{2.1}$$

Clearly, such a probability could be any value between 0 and 1, differently from the analogue result in elementary probability, where in could be instead only in $\{0,1\}$ (since pure states are Dirac δ distributions). Finally, if the outcome of the measurement of $\mathbb{1}_V$ is 1, then the state of the system updates to the pure state $|\varphi\rangle\langle\varphi|$ associated to $|\varphi\rangle$ (so even if before measuring and observing 1 it was not precisely $|\varphi\rangle$, after it is exactly so).

What happens to the system if, after measuring $\mathbb{1}_V$ the observed outcome is 0, i.e. V does not hold? Writing $\mathbb{1}_V = \mathbb{1}_H - \mathbb{1}_{V^\perp}$, where $V^\perp$ is the orthogonal subspace to

V, we postulate that this is equivalent to measuring $\mathbb{1}_{V^\perp}$ and observing that $V^\perp$ holds (interpreted as the negation of the proposition associated to V). This happens with probability

$$\mathbb{P}_\rho(V^\perp) = \mathbb{P}_\rho(\mathbb{1}_V = 0) = \mathrm{tr}\left[\mathbb{1}_{V^\perp}\rho\right] = 1 - \mathrm{tr}\left[\mathbb{1}_V\rho\right] = 1 - \mathbb{P}_\rho(V),$$

which is consistent with additivity of elementary probability (and the fact that we should only observe either 0 or 1). Moreover, by the collapse of the state, the density operator updates in this case to

$$\rho_{V^\perp} = \frac{\mathbb{1}_{V^\perp}\rho\mathbb{1}_{V^\perp}}{\mathbb{P}_\rho(V^\perp)}.$$

We can also ask the following question: can we describe the state of the system after $\mathbb{1}_V$ has been measured (remember that we think of $\mathbb{1}_V$ as a physical device interacting with the system), but no value has been observed? This is also postulated to be the convex combination

$$\rho_V\mathbb{P}_\rho(V) + \rho_{V^\perp}\mathbb{P}_\rho(V^\perp) = \mathbb{1}_V\rho\mathbb{1}_V + \mathbb{1}_{V^\perp}\rho\mathbb{1}_{V^\perp}. \tag{2.2}$$

Such an expression looks very similar to the right hand side of the law of total probability

$$\mathbb{P}(\cdot) = \mathbb{P}(\cdot|V)\mathbb{P}(V) + \mathbb{P}(\cdot|V^c)\mathbb{P}(V^c),$$

but in the quantum setting we cannot conclude (except in special situations) that (2.2) actually equals ρ. Physically, we interpret this discrepancy with the classical laws by stating that simply by allowing the observable $\mathbb{1}_V$ interact with the system, it generates a perturbation in its state (a dynamical explanation of this is usually given in terms of the so-called *de-coherence* phenomenon).

Another fundamental difference between the classical and quantum settings is that, due to possible non-commutativity of operators, differences arises when computing the probability that two "events" V, W hold. Precisely, let $V, W < H$ be subspaces and consider the associated indicator observables $\mathbb{1}_V$, $\mathbb{1}_W \in \mathcal{O}(H)$. We say that V, W are *compatible*[3] if the operators commute:

$$[\mathbb{1}_V, \mathbb{1}_W] = 0, \quad \text{i.e.,} \quad \mathbb{1}_V\mathbb{1}_W = \mathbb{1}_W\mathbb{1}_V \ (= \mathbb{1}_{V\cap W}).$$

If V, W are compatible, measuring first $\mathbb{1}_V$ and then $\mathbb{1}_W$ yields joint outcomes in $\{0,1\}^2$ with the same probability distribution as measuring in the opposite order. Moreover, according to the valued observed, the state is updated to the same density operator. For example, the probability of measuring first $\mathbb{1}_V$, then $\mathbb{1}_W$ and obtaining that both V and W hold is, according to the above postulates and the product rule of elementary probability[4]:

$$\begin{aligned}\mathbb{P}_\rho(\text{first } \mathbb{1}_V = 1, \text{ then } \mathbb{1}_W = 1) = \mathbb{P}_\rho(V)\mathbb{P}_{\rho_V}(W) &= \mathrm{tr}[\rho\mathbb{1}_V]\cdot\frac{\mathrm{tr}[\mathbb{1}_V\rho\mathbb{1}_V\mathbb{1}_W]}{\mathrm{tr}[\rho\mathbb{1}_V]} \\ &= \mathrm{tr}[\rho\mathbb{1}_V I_W\mathbb{1}_V] = \mathrm{tr}[\rho\mathbb{1}_{V\cap W}],\end{aligned} \tag{2.3}$$

yields that the updated density operator

$$\rho_{V,W} = \frac{\mathbb{1}_W\rho_V\mathbb{1}_W}{\mathbb{P}_{\rho_V}(\mathbb{1}_W = 1)} = \frac{\mathbb{1}_W\mathbb{1}_V\rho\mathbb{1}_V\mathbb{1}_W}{\mathbb{P}_\rho(V,W)},$$

which are in both cases symmetric expressions in V, W. Notice that if V, W are orthogonal, i.e., $\mathbb{1}_V\mathbb{1}_W = 0$, then they are compatible.

In the *incompatible* case of non-commuting operators, the probabilities and the updated states may depend on the order in which the measurements are performed. Consider the simplest case of one-dimensional V, W, i.e.,

$$\mathbb{1}_V = |\varphi_0\rangle\langle\varphi_0|, \quad \mathbb{1}_W = |\varphi_1\rangle\langle\varphi_1|,$$

[3]Here one should be careful not to confuse with the definition of compatible and incompatible events in classical probability theory

[4]i.e., $\mathbb{P}(A,B) = \mathbb{P}(A)\mathbb{P}(B|A)$

and let the system be in the pure state corresponding to the state vector $|\psi\rangle$. Then, measuring first $\mathbb{1}_V$ and then $\mathbb{1}_W$ yields as observed outcomes that V and W hold with probability (repeating the computation in (2.3)) given by

$$\mathbb{P}_{|\psi\rangle}(\text{first } V, \text{ then } W) = \mathrm{tr}[\rho \mathbb{1}_V \mathbb{1}_W \mathbb{1}_V] = |\langle\varphi_1|\varphi_0\rangle\langle\varphi_0|\psi\rangle|^2 .$$

Measuring first $\mathbb{1}_W$ and then $\mathbb{1}_V$ instead gives as observed outcomes that both W and V hold with probability

$$\mathbb{P}_{|\psi\rangle}(\text{first } W, \text{ then } V) = |\langle\varphi_1|\varphi_0\rangle\langle\varphi_0|\psi\rangle|^2 ,$$

which is different e.g. if $\langle\varphi_0|\varphi_1\rangle \neq 0$, $\langle\varphi_0|\psi\rangle \neq 0$ but $\langle\varphi_1|\psi\rangle = 0$.

Let us extend the above description to measurements with values in a general (but finite) set $\mathcal{X}$. Here it is convenient to recall that a random variable X with values in $\mathcal{X}$ can be identified, in elementary probability theory, with its induced system of alternatives $(\{X = x\})_{x\in\mathcal{X}}$, that is a family of events, such that exactly one among them is always satisfied. By considering the associated indicator variables, this amounts to require that

$$1_{\{X=x\}} 1_{\{X=y\}} = 0 \quad \text{for every } x \neq y \in \mathcal{X}, \text{ and} \qquad \sum_{x\in\mathcal{X}} 1_{\{X=x\}} = 1_\Omega.$$

This suggests to define an (elementary) measurement X on a quantum system H as a collection of closed subspaces $X = (V_x)_{x\in\mathcal{X}}$ – or equivalently the corresponding indicator observables $X = (\mathbb{1}_{V_x})_{x\in\mathcal{X}}$ – such that the following two conditions hold:

$$\mathbb{1}_{V_x} \mathbb{1}_{V_y} = 0 \quad \text{for every } x \neq y \in \mathcal{X}, \text{ and} \qquad \sum_{x\in\mathcal{X}} \mathbb{1}_{V_x} = \mathbb{1}_H.$$

Such a family of operators is an elementary instance of a so-called *projection-valued measure* (PVM). The first condition yields that all the V_x's are *compatible*, hence they can be measured in any order yielding the outcomes with well-defined probabilities. We refer to this operation as *measuring* X. By the above postulates, if the quantum system is in the state ρ and X is measured, the probability that V_x holds – we simply write $X = x$ in such a case – is

$$\mathbb{P}_\rho(X = x) = \mathbb{P}_\rho(V_x) = \mathrm{tr}(\rho \mathbb{1}_{V_x}).$$

The family $(\mathbb{P}_\rho(X = x))_{x\in\mathcal{X}}$ is a classical probability distribution, which can be thought as the law of X (if the system is the state ρ). If V_x is observed to hold, then we simply say that x is observed and the collapse of the state yields that the density operator updates to

$$\rho_{V_x} = \rho_{|X=x} = \frac{\mathbb{1}_{V_x} \rho \mathbb{1}_{V_x}}{\mathbb{P}_\rho(X = x)}.$$

By the first condition and (2.3), and the second condition and additivity of elementary probability, we have that, when measuring X, exactly one among the V_x's must be observed to hold. Furthermore, if X is measured but the outcome is not observed, the density operator ρ still updates to the convex combination

$$\sum_{x\in\mathcal{X}} \rho_{|X=x} \mathbb{P}_\rho(X = x),$$

thus extending (2.2).

We say that two measurements $X = (V_x)_{x\in\mathcal{X}}$, $Y = (W_y)_{y\in\mathcal{Y}}$ are *compatible* if $\mathbb{1}_{V_x} \mathbb{1}_{W_x} = \mathbb{1}_{W_x} \mathbb{1}_{V_x}$ for every $x \in \mathcal{X}$, $y \in \mathcal{Y}$. In such a case, measuring X and Y yields observed values x, y with a probability $\mathbb{P}_\rho(X = x, Y = y)$ which does not depend on the order of the measurements, and also a well-defined updated state $\rho_{|X=x,Y=y}$.

Remark 2.3. Given a measurement X with values in $\mathcal{X}$ and subset $A \subseteq \mathcal{X}$, we may either define $\mathbb{P}_\rho(X \in A)$ as the sum of the probabilities $\mathbb{P}_\rho(X = x)$ for $x \in A$, or as the probability that, letting V_A be the subspace spanned by the union of the subspaces V_x for $x \in \mathcal{X}$, by measuring $\mathbb{1}_{V_A}$ one observes that V_A holds, i.e., $\mathbb{P}_\rho(V_A)$. It is straightforward to

check that these two probabilities coincide, but one should be careful with the application of Born's rule: after measuring X and observing that the outcome x belongs to A, one should update the state ρ to

$$\frac{\sum_{x\in A}\rho_{|X=x}\mathbb{P}_\rho(X=x)}{\mathbb{P}_\rho(V_A)},$$

while by simply measuring $\mathbb{1}_{V_A}$ and observing that V_A holds, Born's rule prescribes to update the state to

$$\rho_{V_A}=\frac{\mathbb{1}_{V_A}\rho\mathbb{1}_{V_A}}{\mathbb{P}_\rho(V_A)},$$

which is different (in general).

Let us go back to the case of (elementary) quantum observables, that we can now define as measurements X with values in $\mathcal{X}\subseteq\mathbb{R}$. More precisely, given a measurement $X=(V_x)_{x\in\mathcal{X}}$ with $\mathcal{X}\subseteq\mathbb{R}$, we define the associated observable as the self-adjoint operator

$$A_X=\sum_{x\in\mathcal{X}}x\mathbb{1}_{V_x}\in\mathcal{O}(H),$$

clearly reminiscent of the representation of a simple random variable $X=\sum_{x\in\mathcal{X}}xI_{\{X=x\}}$. The observable A_X has spectrum $\sigma(A_X)=\mathcal{X}$: indeed its spectral decomposition is straightforwardly read from the representation. Viceversa, the spectral theorem (in finite dimensions) ensures that any self-adjoint $A\in\mathcal{O}(H)$ can be represented as

$$A=\sum_{\lambda\in\sigma(A)}\lambda\mathbb{1}_{\{A=\lambda\}},$$

where we write $\{A=\lambda\}$ for the eigenspace associated to the eigenvalue $\lambda\in\sigma(A)$. Thus A naturally corresponds to the measurement $X_A=(\mathbb{1}_{A=\lambda})_{\lambda\in\sigma(A)}$. The probabilities

$$\mathbb{P}_\rho(A=\lambda)=\mathrm{tr}[\rho\mathbb{1}_{A=\lambda}],$$

which play the role of the distribution of A (if the system is in the state ρ), allow then to compute mean, variances and other classical quantities via the usual definitions for random variables. For example, the mean of $A\in\mathcal{O}(H)$ is defined by

$$(A)_\rho=\sum_{\lambda\in\sigma(A)}\lambda\mathbb{P}_\rho(A=\lambda)=\mathrm{tr}[\rho A].$$

More generally, given any function $f:\sigma(A)\to\mathbb{R}$, we define (via so-called *functional calculus*)

$$f(A)=\sum_{\lambda\in\sigma(A)}f(\lambda)\mathbb{1}_{\{A=\lambda\}},$$

so that

$$(f(A))_\rho=\sum_{\lambda\in\sigma(A)}f(\lambda)\mathbb{P}_\rho(A=\lambda)=\mathrm{tr}[\rho f(A)].$$

In particular, the variance of A is given by

$$\sigma^2_\rho(A)=((A-(A)_\rho\mathbb{1})^2)_\rho=(A^2)_\rho-(A)^2_\rho=\sum_{\lambda\in\sigma(A)}(\lambda-(A)_\rho)^2\mathbb{P}_\rho(A=\lambda).$$

Finally, let us remark that, given two observables A, $B\in\mathcal{O}(H)$, if the associated measurements are compatible, then clearly they commute, $[A,B]=0$. Conversely, it is a well-known fact that if they commute, then the associated measurements are compatible. This can be seen by representing them as commuting Hermitian matrices with respect to a chosen basis, hence they can be simultaneously put into diagonal form by conjugation with the the same unitary U.

2.3. **C^*-algebras approach.** The above definitions of states, measurements and observables are particularly simple because we are restricted to the elementary case of finite dimensional quantum systems H. Quite interestingly, the earliest historical formulation of quantum mechanics (Heisenberg's matrix mechanics) already dealt with infinite dimensional systems. Essentially (in Schrödinger's representation and setting for simplicity the reduce Planck constant $\hbar$ equal to 1) one considers $H = L^2(\mathbb{R}, dx)$ and the "observables" describing position Q and momentum P of a particle on the real line given by the operators

$$(Q\psi)(x) = x\psi(x), \quad (P\psi)(x) = -i\partial_x\psi(x).$$

Clearly, these operators are not well defined for every $\psi \in L^2(\mathbb{R}, dx)$. The crucial property they satisfy is the so-called *canonical commutation relation* (CCR):

$$[Q, P] = QP - PQ = i\mathbb{1}_H \tag{2.4}$$

at least when tested on smooth compactly supported functions. One can actually argue that any pair of operators on a Hilbert space satisfying (2.4) must be unbounded (Exercise 2.11). This motivates the need for a more sophisticated spectral theory ensuring that they enjoy a similar decomposition as in the finite dimensional case (in order to define the associated notion of measurement).

A different approach, that we briefly describe here, consists of searching for a family of bounded operators that still contain all the useful information about the observables P and Q and, by duality, about the possible states of the quantum system. In the classical case, this amounts to define probability distributions via Riesz theorem, as certain linear functionals over continuous functions on a compact topological space. It turns out that the correct structure of such an abstract family is that of a C^*-algebra $\mathcal{A}$, defined as follows:

i) $\mathcal{A}$ is a complex Banach space,
ii) with an additional product operation $(a, b) \mapsto ab$ that yields a structure of Banach algebra, i.e., it is associative and distributive with respect to the addition operation, there exists an identity element[5] $\mathbb{1}$, and the norm satisfies $\|ab\| \leq \|a\| \, \|b\|$ for every $a, b \in \mathcal{A}$,
iii) and with an additional anti-linear map $* : \mathcal{A} \to \mathcal{A}$, $a \mapsto a^*$, that is an involution $(a^*)^* = a$, satisfying $(ab)^* = b^*a^*$, for $a, b \in \mathcal{A}$ and the C^*-identity holds:

$$\|a^*a\| = \|a\|^2 . \tag{2.5}$$

Remark 2.4. If $\mathcal{A}$ is a Banach algebra and $*$ is an enjoys all the properties but (2.5), to conclude that $\mathcal{A}$ is a C^*-algebra it is enough to argue that

$$\|a^*a\| \geq \|a\|^2 .$$

Indeed, the inequality $\|a\|^2 \leq \|a^*a\| \leq \|a^*\| \, \|a\|$ implies then $\|a\| = \|a^*\|$, hence (2.5).

A $*$-homomorphism between C^*-algebras $\mathcal{A}$, $\mathcal{B}$ is a map $\pi : a \mapsto \pi(a)$ which is well-behaved with respect to all the operations, i.e. it is a ring homomorphism and $\pi(a^*) = \pi(a)^*$. Two C^*-algebras as *isomorphic* if there exists an invertible $*$-homomorphism between them.

This notion may seem complicated at first, but one easily checks that

(1) The space $\mathcal{A} = C(K; \mathbb{C})$ of continuous complex-valued functions on a compact Hausdorff topological space K is indeed a C^*-algebra (with the natural sum and product operations, endowed with the uniform norm $\|f\| = \sup_{x \in K} |f(x)|$, letting $\mathbb{1}(x) = 1$ for $x \in K$, and $f^*(x) = \overline{f(x)}$). In fact, any C^*-algebra $\mathcal{A}$ whose product is commutative, i.e., $ab = ba$ for every $a, b \in \mathcal{A}$, is isomorphic to this case (Gelfand theorem).

[5] C^*-algebras may be defined without the identity element $\mathbb{1}$, but for simplicity we restrict to such a case.

(2) The space of $d \times d$ complex matrices $\mathcal{A} = \mathbb{C}^{d\times d}$, endowed with the natural matrix sum and product operations, the matrix norm

$$\|A\|^2 = \sup_{v\in\mathbb{C}^d\setminus\{0\}} \frac{\|Av\|^2}{\|v\|^2} = \sup_{v\in\mathbb{C}^d\setminus\{0\}} \frac{\langle v|A^*Av\rangle}{\langle v|v\rangle}$$

and A^* being the conjugate transpose of A, is a C^*-algebra. To see that the C^*-identity holds, we simply use Cauchy-Schwarz inequality to argue that $\|A\|^2 \leq \|A^*A\|$.

(3) Generalizing the above example, the space of linear operators $\mathcal{A} = \mathcal{L}(H)$ on a complex finite dimensional Hilbert space H endowed with the operator norm and the adjoint $A \mapsto A^*$ is also a C^*-algebra. When H is infinite dimensional, one should restrict $\mathcal{A} = \mathcal{B}(H)$ to the space of linear bounded operators.

Motivated by the analogy with operator algebras, given a C^*-algebra and $a \in \mathcal{A}$, we say that a is *self-adjoint* if $a = a^*$, a is *positive* if there exists $b \in \mathcal{A}$ such that $a = b^*b$, *unitary* if $aa^* = a^*a = \mathbb{1}$, and define the *spectrum* $\sigma(a) \subseteq \mathbb{C}$ as the set of $\lambda \in \mathbb{C}$ such that $a - \lambda\mathbb{1}$ is not invertible (with respect to the product operation in $\mathcal{A}$). Notice that if a is self-adjoint, then $\sigma(a) \subseteq \mathbb{R}$. Although it is not immediate to prove, it holds that self-adjoint elements $a \in \mathcal{A}$ such that $\sigma(a) \subseteq [0, \infty)$ are exactly the positive elements, i.e., one can represent $a = b^*b$ for some $b \in \mathcal{A}$.

Back to quantum mechanics, the C^*-algebra approach consists of reversing the order of the postulates and begin by defining the notion of *observables* for a quantum system H as the self-adjoint elements in a C^*-algebra $\mathcal{A}$. The *states* are then successively defined as continuous linear functionals

$$\eta : \mathcal{A} \to \mathbb{C}$$

that are positive, i.e. $\eta(a) \geq 0$ for every $a = b^*b$ positive, and such that $\eta(\mathbb{1}) = 1$. The case of elementary quantum systems H is recovered by letting, $\mathcal{A} = \mathcal{L}(H)$ the C^*-algebra of bounded operators and letting $\eta(A) = \mathrm{tr}[A\rho]$ for a density operator $\rho \in \mathcal{S}(H)$ (using duality in finite dimensions, one proves that $\rho \mapsto \eta$ is bijective). In fact, any C^*-algebra is isomorphic to a sub-algebra of $\mathcal{B}(H)$ for some Hilbert space, as shown by the Gelfand-Naimark-Segal construction (Theorem 3.5).

Back to the CCR framework, i.e., on $L^2(\mathbb{R}, dx)$, instead of working with unbounded operators Q and P, one defines the so-called family of *Weyl operators* $(W(r,s))_{(r,s)\in\mathbb{R}^2}$ that are formally given by the imaginary exponentials

$$W(r,s) = e^{i(sQ-rP)}.$$

A rigorous definition as bounded operators on $H = L^2(\mathbb{R}, dx)$ is

$$W(r,s)\psi(x) = e^{is(x-r/2)}\psi(x-r).$$

This definition is motivated (non-rigorously) from the validity of (2.4). Clearly, we must have

$$W(r,0)\psi(x) = e^{-ir(-i\partial_x)}\psi(x) = e^{-r\partial_x} = \psi(x-r),$$

and

$$W(0,s)\psi(x) = e^{isx}\psi(x).$$

Using (2.4) and the Baker-Campbell-Hausdorff formula (truncated an the first commutator, since it is a multiple of the identity, hence all the subsequent terms vanish) we obtain

$$W(r,0)W(0,s) = e^{-irP}e^{isQ} = e^{i(sQ-rP)+\frac{1}{2}[isQ,-irP]} = W(r,s)e^{irs/2}.$$

One can check straightforwardly from the definition that

$$W(r,s)^* = W(-r,-s)$$

and

$$W(r_1,s_1)W(r_2,s_2) = e^{-i(r_1s_2-r_2s_1)/2}W(r_1+r_2, s_1+s_2), \tag{2.6}$$

which encodes the CCR (2.4). By definition, the *Weyl algebra* is the C^*-algebra generated by the Weyl operators (as a closed sub-algebra of $\mathcal{B}(L^2(\mathbb{R}, dx))$).

The Weyl algebra has a rich structure and its study would require an entire course on its own. Let us however conclude with the following definition: given a state η on the Weyl algebra, its *characteristic function* is defined as

$$\mathbb{R}^2 \ni (r, s) \mapsto \eta(W(r, s)) \in \mathbb{C}.$$

In complete analogy with the characteristic function (i.e., Fourier transform) of classical probability measures on $\mathbb{R}^2$, we say that a state η is a quantum (bosonic) *Gaussian* state if its characteristic function is an exponential of a quadratic polynomial in the variables r, s (with complex coefficients).

Finally, let us notice, the CCR can be extended from $\mathbb{R}$ to any $\mathbb{R}^d$, by using multiplication by coordinates $Q_j = x_j$ and partial derivations $P_j = -i\partial_{x_j}$, for $j \in \{1, \dots, d\}$. Similarly, one defines an associated Weyl algebra generated by operators $(W(r, s))_{r,s\in\mathbb{R}^d}$.

2.4. **Exercises.**

Exercise 2.1 (Hilbert-Schmidt scalar product)**.** Let H be an elementary quantum system and $A, B \in \mathcal{L}(H)$. Prove that

$$(A, B) \mapsto \mathrm{tr}[A^*B]$$

defines a scalar product on $\mathcal{L}(H)$ (called Hilbert-Schmidt scalar product). By choosing an orthonormal basis $(|i\rangle)_{i\in I}$, write explicitly its expression in terms of the matrices representing A and B.

Exercise 2.2. Let H be an elementary quantum system and $A, B \in \mathcal{L}(H)$. Discuss the validity of the following statements.

(1) If $A, B \in \mathcal{O}(H)$, then $\mathrm{tr}[AB] \in \mathbb{R}$.
(2) If $\mathrm{tr}[AB] \in \mathbb{R}$ for every $B \in \mathcal{O}(H)$, then necessarily $A \in \mathcal{O}(H)$.
(3) If $A, B \in \mathcal{O}_{\geq}(H)$, then $\mathrm{tr}[AB] \geq 0$.
(4) If $A \in \mathcal{O}(H)$ and $\mathrm{tr}[AB] \geq 0$ for every $B \in \mathcal{O}_{\geq}(H)$, then necessarily $A \geq 0$.

Exercise 2.3 (A quantum Jensen inequality)**.** On an elementary quantum system H, consider an observable $A \in \mathcal{O}(H)$. Let $f : \sigma(A) \to \mathbb{R}$ be convex, i.e.

$$f\left(\sum_{x\in\sigma(A)} xp_x\right) \leq \sum_{x\in\sigma(A)} f(x)p_x$$

for every probability distribution $(p_x)_{x\in\sigma(A)}$. For every density operator $\rho \in \mathcal{S}(H)$, prove the following inequality:

$$f((A)_\rho) \leq (f(A))_\rho.$$

Exercise 2.4 (Purity of a density operator)**.** Given a density operator $\rho \in \mathcal{S}(H)$ on an elementary quantum system H, define its *purity* as $\mathrm{tr}[\rho^2]$.

(1) Prove that the purity always belongs to the interval $[\dim(H)^{-1}, 1] \subseteq (0, 1]$. *Hint: for the lower bound, write* $\mathrm{tr}[\rho^2] = \dim(H)(\rho^2)_\sigma$, *where* $\sigma = \mathbb{1}_H/\dim(H)$ *and use the previous exercise, or use Cauchy-Schwarz inequality with respect to the Hilbert-Schmidt scalar product)*
(2) Prove that ρ is a pure state if and only if its purity equals 1.

Exercise 2.5 (Pauli operators)**.** On a single-qubit system $\mathbb{C}^2$ define the *Pauli operators* σ_x σ_y, σ_z, represented in the computational basis $|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $|1\rangle \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ as the matrices

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_x = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

(1) Show that the Pauli operators are observables and unitary (hence involutions). Determine their spectra.
(2) Show the commutation relations, for any permutation (j,k,ℓ) of the triple (x,y,z)

$$[\sigma_j,\sigma_k]=2i\epsilon_{jk\ell}\sigma_\ell,$$

(where $\epsilon_{jk\ell}$ denotes the Levi-Civita symbol, i.e., $+1$ if the permutation is even, -1 if it is odd).
(3) Deduce that $\mathrm{tr}[\sigma_j^2]=1$ and $\mathrm{tr}[\sigma_j\sigma_k]=0$, i.e., they are orthonormal with respect to the Hilbert-Schmidt scalar product on $\mathcal{L}(\mathbb{C}^2)$. Can you describe the linear space generated by them?
(4) Prove that $\{\mathbb{1}_{\mathbb{C}^2},\sigma_x,\sigma_y,\sigma_z\}\subseteq\mathcal{L}(\mathbb{C}^2)$ are an orthonormal basis.

Exercise 2.6. In the same setting as the previous exercise, given a vector $b=(b_x,b_y,b_z)\in\mathbb{C}^3$, define the operator

$$b\cdot\sigma=b_x\sigma_x+b_y\sigma_y+b_z\sigma_z.$$

Prove that the following properties hold:

(1) $b\cdot\sigma=b'\cdot\sigma$ for $b,b'\in\mathbb{C}^3$ if and only if $b=b'$,
(2) $(b\cdot\sigma)^*=\bar{b}\cdot\sigma$, where $\bar{b}=(\bar{b}_x,\bar{b}_y,\bar{b}_z)$.
(3) one has

$$(b\cdot\sigma)^*(b'\cdot\sigma)=(b\cdot b')\mathbb{1}_{\mathbb{C}^2}+i(b\times b')\cdot\sigma,$$

where $\cdot$ denote respectively the scalar and cross products on $\mathbb{C}^3$ (use the same formulas as in $\mathbb{R}^3$ but take the conjugate of the components of the first vector).

Exercise 2.7 (Bloch sphere)**.** In the same setting as the two previous exercises, given a density operator $\rho\in\mathcal{S}(\mathbb{C}^2)$,

(1) prove that it can be represented as

$$\rho=\frac{1}{2}\left(\mathbb{1}_{\mathbb{C}^2}+b\cdot\sigma\right),\tag{2.7}$$

for a unique $b=b(\rho)\in\mathbb{R}^3$ with $|b|^2=b_x^2+b_y^2+b_z^2\leq 1$ (often called the Bloch ball),
(2) compute the purity of ρ (see Exercise 2.4) in terms of b.
(3) show that ρ is a pure state if and only of $|b|^2=1$ (i.e., it belongs to the so-called Bloch sphere).

Exercise 2.8 (Creation and annihilation operators on a two-level system)**.** On the two-level quantum system $\mathbb{C}^2$, consider the following *annihilation* (or lowering) operator $a\in\mathcal{L}(\mathbb{C}^2)$, acting on the computational basis as follows[6]:

$$a\,|0\rangle=0,\quad a\,|1\rangle=|0\rangle\,.$$

(1) Represent a, as well as its adjoint a^* (called the creation or raising) as a matrices in the computational basis.
(2) Prove that a, a^* satisfy the following *canonical anti-commutation relation* (CAR):

$$\{a,a^*\}=aa^*+a^*a=\mathbb{1}_{\mathbb{C}^2}$$

(3) Represent also the the *number operator* $N=a^*a$ as a matrix, and show that $\sigma(N)=\{0,1\}$. We think of this observable as describing if a quantum particle is located or not at a certain site (if $N=0$, the site is empty, if $N=1$ the site is occupied by a particle).

[6] it is customary to use lower case letters for this operator.

(4) Given a state vector $|\psi\rangle = \alpha_0 |0\rangle + \alpha_1 |1\rangle \in \mathbb{C}^2$, compute the probabilities
$$\mathbb{P}_{|\psi\rangle}(N = 0), \quad \mathbb{P}_{|\psi\rangle}(N = 1)$$
in terms of the amplitudes α_0, α_1. What is the state of the system after measuring N? and after observing $N = 1$?
(5) Answer the same questions in the previous point if the state is described by a density operator $\rho \in \mathcal{L}(\mathbb{C}^2)$ represented by b in the Bloch ball (i.e., as in (2.7)).
(6) Give an explicit representation of a, a^* and N in terms of Pauli operators.

Exercise 2.9 (Creation and annihilation operators on a d-level system)**.** For $d \geq 2$, consider the d-level quantum system $\mathbb{C}^d$ and define the following *annihilation* (or lowering) operator $a \in \mathcal{L}(\mathbb{C}^d)$, generalizing the case $d = 2$ from the previous exercise, in terms of the computational basis $(|k\rangle)_{k=0}^{d-1}$:
$$a |0\rangle = 0, \quad a |k\rangle = \sqrt{k} |k-1\rangle \text{ for every } k = 1, \dots, d-1.$$
as well as its adjoint (the creation operator or raising operator) a^*, and the number operator $N = a^* a$.

(1) What is the spectrum of N?
(2) Assuming that the system is described by a state $\rho \in \mathcal{S}(\mathbb{C}^d)$ and N is measured, what are the probabilities $\mathbb{P}_\rho(N = k)$ (e.g. in terms of the density matrix of ρ with respect to the computational basis)
(3) After measuring N, one observes $N = 0$. How is the state updated according to Born's rule?

Exercise 2.10 (Commutators act as derivations)**.** On an elementary quantum system H, consider operators $A, B, C \in \mathcal{L}(H)$. Show the identity
$$[A, (BC)] = [A, B]C + B[A, C],$$
i.e., the operator $[A, \cdot]$ satisfies a Leibniz-type rule for the product. Deduce by induction that, for every $n \geq 1$,
$$[A, B^n] = \sum_{k=0}^{n-1} B^k [A, B] B^{n-1-k}.$$

Exercise 2.11 (CCR cannot be realized by bounded operators)**.** Prove that one cannot define two operators Q, $P \in \mathcal{L}(H)$ satisfying (2.4) on a finite dimensional Hilbert space H – or ever as bounded operators, Q, $P \in \mathcal{B}(H)$ on a general Hilbert space H. *(Hint: compute* $[Q, P^n]$ *and consider its operator norm as* $n \to \infty$*)*

Exercise 2.12 (Pure states on the Weyl algebra)**.** Let $\mathcal{A} \subseteq \mathcal{L}^2(\mathbb{R}, dx)$ denote the Weyl algebra. Show that, for every $|\psi\rangle \in L^2(\mathbb{R}, dx)$, the functional
$$W(r, s) \mapsto |\psi\rangle\, W(r, s)\psi = \int_{\mathbb{R}} \bar{\psi}(x) W(r, s)\psi(x) dx$$
defines a state $\eta = \eta_{|\psi\rangle}$. For which (complex valued) $|\psi\rangle$ the state η turns out to be Gaussian?

Exercise 2.13 (Translations on the Weyl algebra)**.** Given a state η on the Weyl algebra $\mathcal{A} \subseteq L^2(\mathbb{R}, dx)$, show that, for every given $(r, s) \in \mathbb{R}^2$, the functional
$$a \in \mathcal{A} \mapsto \eta(W(-r, -s) a W(r, s))$$
also defines a state on $\mathcal{A}$, and write its characteristic function in terms of the characteristic function of η.

Exercise 2.14 (A quantum Bochner theorem)**.** Recall that a function $f : \mathbb{R}^d \to \mathbb{C}$ is called *positive semidefinite* if, for every finite family $(x_i)_{i=1}^d$, the matrix $(f(x_i - x_j))_{i,j=1}^d \in \mathbb{C}^{d\times d}$ is hermitian and positive semidefinite. A theorem by Bochner proves that any continuous positive semidefinite function f with $f(0) = 1$ is the characteristic function of a Borel probability measure μ on $\mathbb{R}^d$, i.e., one has the representation

$$f(x) = \int_{\mathbb{R}^d} e^{ix\cdot\xi} d\mu(\xi), \quad \text{for every } x \in \mathbb{R}^d.$$

(1) Show the (easy) implication: the characteristic function of Borel probability measure μ is indeed a continuous positive semidefinite function f with $f(0) = 1$.

A quantum analogue of Bochner theorem (see e.g. [Gos21]) describes exactly the characteristic functions of states η on the Weyl algebra. Let us give the following definition. Given a bilinear form $\beta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$, we say that a function $f : \mathbb{R}^d \to \mathbb{C}$ is *β-positive semidefinite* if, for every finite family $(x_i)_{i=1}^d$, the matrix $(f(x_i - x_j)e^{\beta(x_j,x_i)})_{i,j=1}^d \in \mathbb{C}^{d\times d}$ is positive semidefinite.

Define $\omega : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{C}$,

$$\omega((r_1, s_1), (r_2, s_2)) = (r_1, s_1) \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} r_2 \\ s_2 \end{pmatrix} = \det \begin{pmatrix} r_1 & r_2 \\ r_1 & s_1 \end{pmatrix} = r_1 s_2 - r_2 s_1 \quad (2.8)$$

(which actually defines a non-degenerate *symplectic form*). Then, the quantum analogue of Bochner theorem states that any continuous $\frac{i}{2}\omega$-positive semidefinite function $f : \mathbb{R}^2 \to \mathbb{C}$ with $f(0) = 1$ is the characteristic function of a state η on the Weyl algebra:

$$f(r, s) = \eta(W(r, s)).$$

(2) Show the implication: the characteristic function of a state η on the Weyl algebra is a continuous $\frac{i}{2}\omega$-positive semidefinite[7] function f with $f(0) = 1$. *(Hint: use* (2.6)*)*

3. Quantum Channels

In this section we introduce transformations of quantum systems, generalizing the probabilistic notion of Markov kernels and completing the program of describing the three fundamental objects (states, observables, transformations) in the (elementary) quantum mechanical setting.

3.1. **Tensor products.** Let us recall some basic facts an notations for operators, and in particular tensor products. For simplicity, we only deal with finite dimensional spaces H, K. Write $\mathcal{L}(H; K)$ for the space of linear operators $A : H \to K$. The adjoint operator $A^* : K \to H$ is defined as usual by requiring that

$$\langle\varphi|A^*\psi\rangle = \langle A\varphi|\psi\rangle\,, \quad \text{for every } |\psi\rangle \in K,\ \varphi \in H.$$

When $H = K$, we write $\mathcal{L}(H) = \mathcal{L}(H; H)$ and always endow it with the Hilbert-Schmidt scalar product

$$\langle A|B\rangle = \mathrm{tr}[A^*B], \quad \text{for } |A\rangle\,, |B\rangle \in \mathcal{L}(H),$$

see also Exercise 2.1. An isometry $U : H \to K$ is a linear map preserving the norms (or equivalently the scalar products)

$$\langle U\varphi|U\psi\rangle = \langle\varphi|\psi\rangle\,, \quad \text{for every } |\varphi\rangle\,, |\psi\rangle \in H,$$

or equivalently, such that $U^*U = \mathbb{1}_H$.

The *tensor product* $H \otimes K$ between H and K can be abstractly defined as the linear space generated by formal expressions (*elementary tensors*) $|\varphi\rangle \otimes |\psi\rangle$ quotiented so that the expressions become bi-linear, i.e.,

$$(|\varphi_0\rangle + |\varphi_1\rangle) \otimes |\psi\rangle = |\varphi_0\rangle \otimes |\psi\rangle + |\varphi_1\rangle \otimes |\psi\rangle\,,$$

[7] also called KLM condition

$$|\varphi\rangle \otimes (|\psi_0\rangle + |\psi_1\rangle) = |\varphi\rangle \otimes |\psi_0\rangle + |\varphi\rangle \otimes |\psi_1\rangle\,,$$

(as well as similar relations for multiplication by scalars). It is common in Dirac's notation to abbreviate $|\varphi\rangle \otimes |\psi\rangle = |\varphi, \psi\rangle$. We endow the tensor product $H \otimes K$ with the scalar product defined on elementary tensors as

$$\langle \varphi_0 \otimes \psi_0 | \varphi_0 \otimes \psi_0 \rangle = \langle \varphi_0 | \varphi_1 \rangle \langle \psi_0 | \psi_1 \rangle\,.$$

and extended by linearity (on should check that this is indeed a well-defined scalar product). The dimension of $H \otimes K$ is $\dim(H)\dim(K)$ and an orthonormal basis is given by $(|i, j\rangle)_{i\in I, j\in J} = (|i\rangle \otimes |j\rangle)_{i\in I, j\in J}$ for orthonormal bases $(|i\rangle)_{i\in I} \subseteq H$, $(|j\rangle)_{j\in J} \subseteq K$. Elements in the tensor product are therefore represented as $I \times J$ complex matrices, hence they can be thought as operators. A more natural correspondence between tensors and operators is the isomorphism between $H \otimes K^*$ and $\mathcal{L}(K; H)$, given by

$$|h\rangle \otimes \langle k| = |h\rangle \langle k|\,,$$

and extended by linearity.

Tensor products are used in quantum mechanics to represent *composite systems* made by "joining" two quantum systems H, K. States on the composite system $H \otimes K$ are represented by density operators $\rho \in \mathcal{S}(H \otimes K)$, while observables are self-adjoint operators $A \in \mathcal{O}(H \otimes K)$. Let us recall therefore some basic facts on general operators $M \in \mathcal{L}(H \otimes K)$, and in particular the partial trace operation.

The tensor product construction naturally extends to operators as follows: given $A \in \mathcal{L}(H; \tilde{H})$, $B \in \mathcal{L}(K; \tilde{K})$, one defines the operator $A \otimes B \in \mathcal{L}(H \otimes H; \tilde{H} \otimes \tilde{K})$ acting on elementary tensors as

$$(A \otimes B)\,|\varphi\rangle \otimes |\psi\rangle = |A\varphi\rangle \otimes |B\psi\rangle$$

and then extends the operator by linearity. It is simple to check that $(A \otimes B)^* = A^* \otimes B^*$, hence it $A \otimes B \in \mathcal{O}(H \otimes K)$ is self-adjoint if both $A \in \mathcal{O}(H)$ and $B \in \mathcal{O}(K)$ are both self-adjoint (but the converse may not hold). The spectrum $\sigma(A \otimes B)$ is given by the pairwise products of the elements in the spectra of $\sigma(A)$ and $\sigma(B)$. In particular $A \otimes B \in \mathcal{O}_{\geq}(H \otimes K)$ is positive if both A and B are positive operators.

By choosing suitable bases, one can represent any operator $M \in \mathcal{L}(H \otimes K; \tilde{H} \otimes \tilde{K})$ as a Kronecker product of matrices (or more conveniently, as a block matrix whose entries are operators in $\mathcal{L}(H; \tilde{H})$). Assume for simplicity that $H = \tilde{H}$, $K = \tilde{K}$. Then, choosing orthonormal bases $(|i\rangle)_{i\in I} \subseteq H$, $(|j\rangle)_{j\in J} \subseteq K$ yields the representation

$$M = \sum_{i,j,k,\ell} M_{ij,k\ell}\,|i, j\rangle \langle k, \ell|\,, \tag{3.1}$$

with $M_{ij,k\ell} = \langle i \otimes j | M (k \otimes \ell)\rangle$. For fixed j and $\ell \in J$, define the operator acting on H given by

$$M_{j,\ell} = \sum_{i,k} M_{ij,k\ell}\,|i\rangle \langle k| \in \mathcal{L}(H).$$

Then, M is identified with the block matrix

$$M = (M_{j,\ell})_{j,\ell\in J}\,.$$

When $K = \mathbb{C}^d$, such block matrix representation is always understood with respect to the computational basis. For example, $M \in \mathcal{L}(H \otimes \mathbb{C}^2)$ can be represented as

$$M = \begin{pmatrix} M_{00} & M_{01} \\ M_{10} & M_{11} \end{pmatrix},$$

where each $M_{ij} \in \mathcal{L}(H)$. Notice that $M \in \mathcal{O}(H \otimes \mathbb{C}^2)$ if and only if M_{00}, $M_{11} \in \mathcal{O}(H)$ and $M_{01}^* = M_{10}$. In Section 4, we provide a criterion for positivity (Lemma 4.1).

Over a tensor product space $H \otimes K$, one defines the *partial trace* operation over H as the linear operator

$$\mathrm{tr}_H : \mathcal{L}(H \otimes K) \to \mathcal{L}(K), \quad M \mapsto \mathrm{tr}_H[M]$$

such that, for every $A \in \mathcal{L}(K)$, one has

$$\mathrm{tr}[A^* \mathrm{tr}_H[M]] = \mathrm{tr}[(\mathbb{1}_H \otimes A^*)M].$$

In other words, tr_H is the adjoint of the partial tensor product operation $A \mapsto \mathbb{1}_H \otimes A$ with respect to the Hilbert-Schmidt scalar product. Analogously, one defines a partial trace over K. By representing M as in (3.1), one has the formulas

$$\mathrm{tr}_H[M] = \sum_{j,\ell} |j\rangle \langle \ell| \, \mathrm{tr}_H[M]_{j,\ell}, \quad \mathrm{tr}_H[M]_{j,\ell} = \sum_i M_{ij,i\ell},$$

and

$$\mathrm{tr}_K[M] = \sum_{i,k} |i\rangle \langle k| \, \mathrm{tr}_K[M]_{i,k}, \quad \mathrm{tr}_K[M]_{i,k} = \sum_j M_{ij,kj}.$$

From these expressions is straightforward to check that, if $M \in \mathcal{O}(H \otimes K)$ is self-adjoint, so are $\mathrm{tr}_H[M]$ and $\mathrm{tr}_K[M]$, although we are going to give a more abstract argument below. Similarly, if $M = \rho \in \mathcal{S}(H \otimes K)$, both $\mathrm{tr}_H[\rho]$, $\mathrm{tr}_K[\rho]$ are density operators (respectively on K and H). These are called *reduced* density operators, and correspond to the elementary notion of marginal densities in probability.

It is worth mentioning here the following definitions concerning states on a composite system $H \otimes K$. We say that $\rho \in \mathcal{S}(H \otimes K)$ is *separable* if it can be represented as a convex combination

$$\rho = \sum_{x \in \mathcal{X}} p_x \rho_x \otimes \sigma_x,$$

with $\rho_x \in \mathcal{S}(H)$, $\sigma_x \in \mathcal{S}(K)$ and $(p_x)_{x \in \mathcal{X}}$ a classical probability distribution over a finite set $\mathcal{X}$. States $\rho \in \mathcal{S}(H \otimes K)$ that are *not separable* are called *entangled*. Entangled states enjoy some interesting properties that have no classical analogue and play an important role in possible advantages of quantum computation and information theory with respect to their classical counterparts.

3.2. **Markov kernels.** In elementary probability theory, given finite sets Ω and $\mathcal{X}$, a Markov kernel (from Ω to $\mathcal{X}$) is a collection $N = (N(\omega, \cdot))_{\omega \in \Omega}$ parametrized by $\omega \in \Omega$, consisting of probability distributions on $\mathcal{X}$, so that for every $\omega \in \Omega$,

$$N(\omega, x) \in [0, 1] \quad \text{for all } x \in \mathcal{X}, \text{ and } \quad \sum_{x \in \mathcal{X}} N(\omega, x) = 1. \tag{3.2}$$

This is often conveniently represented via a *stochastic matrix* $(N(\omega, x))_{\omega \in \Omega, x \in \mathcal{X}}$. There are two natural operations associated to a kernel N, corresponding to matrix-vector multiplication by the associated matrix:

(1) given a function $f : \mathcal{X} \to \mathbb{C}$, one defines the function $Nf : \Omega \to \mathbb{C}$ as

$$N(f)(\omega) = \sum_{x \in \mathcal{X}} f(x) N(\omega, x),$$

(2) given a function $p : \Omega \to \mathbb{C}$, one defines $N^\dagger p : \mathcal{X} \to \mathbb{C}$ as

$$N^\dagger(p)(x) = \sum_{\omega} p(\omega) N(\omega, x).$$

Clearly, both operations are linear and the notation $N^\dagger$ is motivated by the duality

$$\sum_{x \in \mathcal{X}} f(x) N^\dagger(p)(x) = \sum_{\omega \in \Omega} N(f)(\omega) p(\omega).$$

We use the $\dagger$ symbol instead of $*$, although one could also use $*$ in a proper sense.

Since N is a Markov kernel, it is easy to prove that both operations N and $N^\dagger$ are *positive*, i.e. they map non-negative (real-valued) functions to non-negative functions. Moreover, one has

$$N(1_\Omega) = 1_{\mathcal{X}}$$

and, by duality, for every $p : \Omega \to \mathbb{C}$,

$$\sum_{x\in\mathcal{X}} N^\dagger(p)(x) = \sum_{\omega\in\Omega} p(\omega).$$

In particular, $N^\dagger$ maps probability distributions on Ω to probability distributions on $\mathcal{X}$. In the particular case $\Omega = \mathcal{X}$, this allows to interpret N as the *transition probability* of a Markov chain, hence defining by iterated applications a discrete-time evolution on Ω.

Our aim is to investigate an analogue of the above construction for elementary quantum systems H (and later briefly present the general C^*-algebra setting). Let us take a bottom-up approach and discuss some special cases first.

3.3. **Non-sharp measurements.** A first strategy that we describe is to relax the notion of measurements, and it is interesting on its own in quantum information theory. The idea is to replace the functions $\omega \mapsto N(x,\omega)$ with suitable observables N_x.

Given an elementary quantum system H (which plays the role of the sample space Ω) and a finite set $\mathcal{X}$ (which we can still think of the possible outcomes of a measurement), we define for every $x \in \mathcal{X}$, an observable $N_x \in \mathcal{O}(H)$ such that $0 \le N_x \le \mathbb{1}_H$ (so that $\sigma(N_x) \subseteq [0,1]$) and

$$\sum_{x\in\mathcal{X}} N_x = \mathbb{1}_H.$$

Such a family is called in general a *positive operator valued measure* (POVM). In simpler terms, we are relaxing the *sharp* indicator observables (i.e., the projection operators) $\mathbb{1}_{V_x}$ associated to a measurement with the operators N_x (a similar construction could be done also in the classical case and leads to a notion of randomized, *non-sharp*, random variable). Using this intepretation, we may still define the probability of observing x, after the measurement of $(N_x)_{x\in\mathcal{X}}$ is performed, as the quantity $(N_x)_\rho = \mathrm{tr}\rho N_x$. How should we accordingly transform a density operator ρ? The answer is given by extending (2.2):

$$\rho \mapsto \sum_{x\in\mathcal{X}} \sqrt{N_x}\rho\sqrt{N_x}, \tag{3.3}$$

where $\sqrt{N_x}$ is defined via functional calculus (spectral theorem). Notice that in the (sharp) measurement case, $N_x = \mathbb{1}_{V_x}$, the square root disappears since $\mathbb{1}_{V_x}^2 = \mathbb{1}_{V_x}$.

3.4. **Unitary evolutions.** The above construction however does not exhaust all the possible transformations, in particular we want to define the analogues of a Markov kernel between two finite dimensional quantum system H and K.

Another special, but relevant, case is the transformation induced by an isometry $U : H \to K$. Indeed, we can think of using U to "embed" each state vector $|\psi\rangle$ on H into the state vector $U|\psi\rangle$ on K. The induced transformation at the level of density operators reads

$$\rho \in \mathcal{S}(H) \mapsto U\rho U^* \in \mathcal{S}(K).$$

When $H = K$, one actually postulates that a unitary evolution of the above kind is the one naturally occurring for a *closed* quantum system, i.e., if H is isolated from the rest of the universe. By contrast, all the other transformations we are going to describe are often interpreted as evolution of *open* quantum systems (i.e., when H interacts with a larger system).

We may define more general transformations of states, by taking convex combinations over a family of isometries $(U_x)_{x\in\mathcal{X}}$ with respect to a probability distribution p over the set $\mathcal{X}$, and define

$$\rho \mapsto \sum_{x\in\mathcal{X}} p_x U_x \rho U_x^* \tag{3.4}$$

We need that p is a probability distribution to ensure that the outcome is a densisty operator. When $H = K$, the usual interpretation of such a transformation is that the

system acts by first randomly sampling $x \in \mathcal{X}$ according to the distribution p, and then evolves following the unitary U_x.

3.5. **Kraus representation.** By comparing (3.4) with (3.3), we may see a formal correspondence between $\sqrt{N_x}$ between and $\sqrt{p_x}U_x$. We are thus lead to consider more general general transformations

$$\Phi : \mathcal{L}(H) \mapsto \mathcal{L}(K),$$

that admit a representation of the type

$$\Phi(A) = \sum_{x\in\mathcal{X}} B_x^* A B_x, \tag{3.5}$$

where $(B_x)_{x\in\mathcal{X}} \subseteq \mathcal{L}(K;H)$ is a family of operators, called *Kraus* (or noise) operators. Using cyclicity of the trace, it is not difficult to check that the dual operator, $\Phi^\dagger : \mathcal{L}(K) \to \mathcal{L}(H)$, abstractly defined as

$$\mathrm{tr}[A^*\Phi^\dagger(A')] = \mathrm{tr}[(\Phi(A))^* A'] \quad \text{for every } A \in \mathcal{L}(H),\ A' \in \mathcal{L}(K),$$

also enjoys a similar representation, but in terms of the dual family $(B_x^*)_{x\in\mathcal{X}} \in \mathcal{L}(H;K)$:

$$\Phi^\dagger(A') = \sum_{x\in\mathcal{X}} B_x^* A' B_x. \tag{3.6}$$

It is also not difficult to show that any $\Phi : \mathcal{L}(H) \to \mathcal{L}(K)$ represented via Kraus operators as in (3.5) is *positive*, i.e.,

$$A \in \mathcal{O}_{\geq}(H) \quad \Rightarrow \Phi(A) \in \mathcal{O}_{\geq}(K),$$

Moreover, Φ is *unital*, i.e.,

$$\Phi(\mathbb{1}_H) = \mathbb{1}_K,$$

if and only if

$$\sum_{x\in\mathcal{X}} B_x^* B_x = \mathbb{1}_H. \tag{3.7}$$

By duality, Φ is unital if and only if $\Phi^\dagger$ is *trace-preserving* (often abbreviated as TP), i.e., $\mathrm{tr}[\Phi^\dagger(A)] = \mathrm{tr}[A]$.

The following examples may look trivial, but are still worth noticing. Consider the map $\Phi : \mathcal{L}(\mathbb{C})(= \mathbb{C}) \to \mathcal{L}(H)$ given by

$$\Phi(\lambda) = \lambda \mathbb{1}_H.$$

Clearly, it is positive and unital. Moreover, writing

$$\mathbb{1}_H = \sum_{x\in\mathcal{X}} |x\rangle \langle x|,$$

for a given orthonormal basis $(|x\rangle)_{x\in\mathcal{X}}$, we see that Φ admits a representation as in (3.5) through the kraus operators $(\langle x|)_{x\in\mathcal{X}} \subseteq \mathcal{L}(H;\mathbb{C})$. The dual map $\Phi^\dagger : \mathcal{L}(H) \to \mathcal{L}(\mathbb{C})$ is then represented as (3.6), which reads

$$\Phi^\dagger(A) = \sum_{x\in\mathcal{X}} \langle x|Ax\rangle = \mathrm{tr}[A],$$

i.e., it is the trace map, which is obviously positive and trace-preserving.

It is also immediate to check that positive linear combinations of maps represented by Kraus operators admit a representation in terms of suitable Kraus operators. The same holds for compositions: if $\Phi : \mathcal{L}(H_1) \to \mathcal{L}(H_2)$ is represented by $(B_x)_{x\in\mathcal{X}} \subseteq \mathcal{L}(H_2;H_1)$, and $\Psi : \mathcal{L}(H_2) \to \mathcal{L}(H_3)$ is represented by $(C_y)_{y\in\mathcal{Y}} \subseteq \mathcal{L}(H_3;H_2)$, the composition $\Psi \circ \Phi : \mathcal{L}(H_1) \to \mathcal{L}(H_3)$ is represented via the family $(B_xC_y)_{x\in\mathcal{X},y\in\mathcal{Y}} \subseteq \mathcal{L}(H_3;H_1)$.

We end this section with a structure result showing that a $*$-homomorphism enjoys a representation in terms of Kraus operators.

Lemma 3.1. *Let H, K be finite dimensional Hilbert spaces and let*

$$\Phi : \mathcal{L}(K) \to \mathcal{L}(H)$$

be a $*$-homomorphism, i.e., Φ is linear and

$$\Phi(\mathbb{1}_K) = \mathbb{1}_H, \quad \Phi(AB) = \Phi(A)\Phi(B), \quad \Phi(A^*) = \Phi(A)^*.$$

Then, there exist suitable Kraus operators $(B_x)_{x\in\mathcal{X}} \subseteq \mathcal{L}(H;K)$ such that (3.6) hold. One has in particular $|\mathcal{X}| = \dim(H)/\dim(K)$.

The strategy of the proof is to prove that H is isomorphic to a tensor product $K \otimes \mathbb{C}^{\mathcal{X}}$, and that (up to such an isometry) $\Phi(A) = A \otimes \mathbb{1}_{\mathbb{C}^{\mathcal{X}}}$.

Proof. Fix an orthonormal basis $(|i\rangle)_{i=0}^{\dim(K)-1} \subseteq K$ and define, for every i, j, the operators $E_{ij} = \Phi(|i\rangle\langle j|)$. Using the properties of Φ, we have $E_{ik} = E_{ij}E_{jk}$, $E_{ik}^* = E_{ki}$, for every i, j, k. In particular E_{ii} is an orthogonal projection operator on a subspace, that we denote V_i. For every i, choose an isometry $U_i : K \to K$ such that $U_i|0\rangle = |i\rangle$, so that we have the identity

$$E_{ij} = \Phi(U_i|0\rangle\langle 0|U_j^*) = \Phi(U_i)E_{00}\Phi(U_j)^*,$$

showing in particular that $E_{ii} = \Phi(U_i)E_{00}\Phi(U_i)^*$. Moreover,

$$\sum_i E_{ii} = \sum_i \Phi(|i\rangle\langle i|) = \Phi(\sum_i |i\rangle\langle i|) = \Phi(\mathbb{1}_K) = \mathbb{1}_H$$

Therefore, H is decomposed as the orthogonal sum of the subspaces V_i (in particular $\dim(V_i) = \dim(H)/\dim(K)$). Fix an orthonormal basis $(|x\rangle)_{x\in\mathcal{X}} \subseteq V_0$, so that $|\mathcal{X}| = \dim(H)/\dim(K)$ and

$$E_{00} = \sum_{x\in\mathcal{X}} |x\rangle\langle x|,$$

and define, for every $x \in \mathcal{X}$, $i = 0, \dots, \dim(K) - 1$, the Kraus operator (it is simpler to think about the adjoint)

$$B_x^* = \sum_{i=0}^{\dim(K)-1} \Phi(U_i)|x\rangle\langle i| \in \mathcal{L}(K;H).$$

Writing for $A \in \mathcal{L}(K)$, $A = \sum_{i,j=0}^{d-1} A_{ij}|i\rangle\langle j|$, it is sufficient to check (3.5) with A of the form $|\ell\rangle\langle m|$ for some ℓ, m. We have then

$$\begin{aligned}
\sum_{x\in\mathcal{X}} B_x^*|\ell\rangle\langle m|B_x &= \sum_{x\in\mathcal{X}} \sum_{i,j=0}^{\dim(K)-1} \Phi(U_i)|x\rangle\langle i|\ell\rangle\langle m|j\rangle\langle x|\Phi(U_j)^* \\
&= \sum_{x\in\mathcal{X}} \Phi(U_\ell)|x\rangle\langle x|\Phi(U_m)^* \\
&= \Phi(U_\ell)\sum_{x\in\mathcal{X}} |x\rangle\langle x|\Phi(U_m)^* \\
&= \Phi(U_\ell)E_{00}\Phi(U_m)^* = E_{\ell m} = \Phi(|\ell\rangle\langle m|).
\end{aligned}$$

□

3.6. **Complete positivity.** We are thus led to the following question: is any linear positive map $\Phi : \mathcal{L}(H) \to \mathcal{L}(K)$ represented by a suitable family of Kraus operators? It turns out that this is not in general the case (see Exercise 3.3) but, more importantly, one can single out a simple additional condition which ensures the existence of a representation.

A linear map $\Phi : \mathcal{L}(H) \to \mathcal{L}(K)$ is *completely positive* (CP) if, for every $d \in \mathbb{N}$, the map

$$\Phi \otimes \mathbb{1}_{\mathcal{L}(\mathbb{C}^d)} : \mathcal{L}(H \otimes \mathbb{C}^d) \to \mathcal{L}(K \otimes \mathbb{C}^d) \tag{3.8}$$

is positive[8]. For every $M \in \mathcal{L}(H \otimes \mathbb{C}^d)$ represented in block operator as

$$M = (M_{ij})_{i,j=1}^d \subseteq \mathcal{L}(H),$$

we have that

$$\Phi \otimes \mathbb{1}_{\mathbb{C}^d}(M) = (\Phi(M_{ij}))_{i,j=1}^d \subseteq \mathcal{L}(K).$$

Complete positivity means that, if $M = (M_{ij}) \geq 0$, then $(\Phi(M_{ij}))_{i,j=1}^d$. For example, letting $d = 2$, this means that

$$\begin{pmatrix} M_{00} & M_{01} \\ M_{10} & M_{11} \end{pmatrix} \geq 0 \quad \Rightarrow \quad \begin{pmatrix} \Phi(M_{00}) & \Phi(M_{01}) \\ \Phi(M_{10}) & \Phi(M_{11}) \end{pmatrix} \geq 0.$$

In the next section we are going to provide a useful criterion (Lemma 4.1) for positivity of 2×2 block operators.

Back to the general case, since any $M \in \mathcal{O}_{\geq}(H \otimes \mathbb{C}^d)$ can be represented as $M = A^*A$ (e.g. letting $A = \sqrt{M}$ via spectral calculus), positivity of (3.8) amounts to the fact that

$$\Phi \otimes \mathbb{1}_{\mathcal{L}(\mathbb{C}^d)}(A^*A) \in \mathcal{L}(K \otimes \mathbb{C}^d)$$

is non-negative. Using a block operator representation for A, we have

$$A = \sum_{i,j=1}^d A_{ij} \otimes |i\rangle \langle j|, \quad A^* = \sum_{i,j=1}^d A_{ji}^* \otimes |i\rangle \langle j|,$$

hence

$$\Phi \otimes \mathbb{1}_{\mathcal{L}(\mathbb{C}^d)}(A^*A) = \sum_{i,j=1}^d \sum_{k=0}^{d-1} \Phi(A_{ki}^* A_{kj}) |i\rangle \langle j| \geq 0.$$

We can of course specialize the condition by restricting initially it to "rank-one" block operators of the form

$$A = \sum_{j=1}^d A_j \otimes |1\rangle \langle j|,$$

so that it simplifies to

$$\sum_{i,j=1}^d \Phi(A_i^* A_j) |i\rangle \langle j| \geq 0 \tag{3.9}$$

i.e.,

$$\sum_{i,j=1}^d \langle \psi_i | \Phi(A_i^* A_j) \psi_j \rangle \geq 0 \quad \text{for every } (A_i)_{i=1}^d \subseteq \mathcal{L}(H),\ (|\psi_i\rangle)_{i=1}^d \subseteq K. \tag{3.10}$$

because the left hand side amounts to the quantity $\langle v | M v \rangle$ for $v = \sum_{i=1}^d |\psi_i\rangle \otimes |i\rangle$.

Let us collect the following elementary facts:

(1) any map of the form $\Phi_B(A) = B^*AB$ with $B \in \mathcal{L}(K; H)$ (which we already know to be positive being represented by a single Kraus operator B) is also completely positive, because
$$\Phi_B \otimes \mathbb{1}_{\mathcal{L}(\mathbb{C}^d)} = \Phi_{B \otimes \mathbb{1}_{\mathbb{C}^d}}.$$
(2) linear combinations with positive coefficients of CP maps yield CP maps
(3) the dual of a CP map is also CP, since $\Phi^\dagger \otimes \mathbb{1}_{\mathcal{L}(\mathbb{C}^d)} = \left(\Phi \otimes \mathbb{1}_{\mathcal{L}(\mathbb{C}^d)}\right)^\dagger$,
(4) composition of CP maps is also CP.

[8]More specifically, for each $d \in \mathbb{N}$, a map such that (3.8) is positive is called d-positive. Several results, in particular in Section 4, will indeed hold for general 2-positive maps

An a consequence, we have that any map Φ enjoying a representation via Kraus operators in CP. In particular, since the trace map is CP, also the partial trace map

$$\mathrm{tr}_{H_1} = \mathrm{tr} \otimes \mathbb{1}_{\mathcal{L}(H_2)} : \mathcal{L}(H_1 \otimes H_2) \to \mathcal{L}(H_2).$$

(where we identify H_2 with some $\mathbb{C}^d$ by choosing coordinates). Also its is dual map is CP, which is the "partial tensoring" map

$$A \in \mathcal{L}(H_1) \mapsto A \otimes \mathbb{1}_{H_2} \in \mathcal{L}(H_1 \otimes H_2).$$

A completely positive, trace preserving (CPTP) map $\Phi : \mathcal{L}(H) \to \mathcal{L}(K)$ is also called a quantum operation or *quantum channel* from the system H to K. The following theorem shows that any quantum channel must have a representation in terms of suitable Kraus operators.

Theorem 3.2 (Kraus representation of quantum channels)**.** *Let H, K be finite dimensional Hilbert spaces. Any quantum channel $\Phi^\dagger$ from H to K can be represented via a finite family of Kraus operators $(B_x)_{x\in\mathcal{X}} \subseteq \mathcal{L}(H;K)$ such that $\sum_{x\in\mathcal{X}} B_x^* B_x = \mathbb{1}_H$:*

$$\Phi^\dagger(A) = \sum_{x\in\mathcal{X}} B_x A B_x^* \quad \textit{for every } A \in \mathcal{L}(H).$$

The number of elements of $\mathcal{X}$ is at most $\dim(H)\dim(K)$.

We actually focus on the dual operator $\Phi : \mathcal{L}(K) \to \mathcal{L}(H)$, which is CP and unital. The crucial point is to build an auxiliary system $\tilde{H}$ so that one can represent $\Phi(A) = U^*\pi(A)U$, i.e., a composition of a $*$-homomorphism π and a transformation associated to an isometry $U : K \to \tilde{H}$. This is a special case of the *Stinespring* dilation theorem. Applying Lemma 3.1 to π the required representation.

Proof. We define, on the set $(\mathcal{L}(K) \times H)^2$, the complex valued function

$$\beta : ((A_0, \psi_0), (A_1, \psi_1)) \mapsto \langle \psi_0 | \Phi(A_0^* A_1)\psi_1 \rangle ,$$

where $\langle \cdot | \cdot \rangle$ denotes the scalar product on H. Such β is clearly anti-bilinear in the variables (ψ_0, A_0), and bilinear in (ψ_1, A_1), hence it can be extended to a bilinear form, still denoted with β, on $(\mathcal{L}(K) \otimes H)^2$. We claim that the CP assumption yields that β is non-negative. Indeed, writing any element of $\mathcal{L}(K) \otimes H$ as a (finite) linear combination

$$v = \sum_{i=1}^{d} A_i \otimes |i\rangle ,$$

with $(|i\rangle)_{i=1}^d \subseteq H$, $(A_i)_{i=1}^d \subseteq \mathcal{L}(K)$, using bi-linearity of β, this amounts to prove that

$$\beta(v, v) = \sum_{i,j=1}^{d} \langle i | \Phi(A_i^* A_j) j \rangle \geq 0,$$

which is exactly (3.10). It is not ensured however that β is non-degenerate, i.e. $\beta(v, v) = 0$ implies $v = 0$. We thus introduce the equivalence relation $v \sim v'$ if and only if $\beta(v - v', v - v') = 0$, so that, on the quotient

$$\tilde{H} = K \otimes \mathcal{L}(H)/\sim,$$

which is still a complex vector space, we have that $\beta([v], [w]) = \beta(v, w)$ is a well-defined scalar-product (we denote here and below with $[v]$ the equivalence class of v). Next, we define

$$U : H \to \tilde{H}, \qquad U|\psi\rangle = [\mathbb{1}_K \otimes |\psi\rangle],$$

which is easily seen to be an isometry, using that $\Phi(\mathbb{1}_K) = \mathbb{1}_H$. Moreover, we have

$$U^*[A \otimes |\varphi\rangle] = \Phi(A)\,|\varphi\rangle\,, \tag{3.11}$$

because of the identity

$$\begin{aligned}\beta(U|\psi\rangle, [A\otimes|\varphi\rangle]) &= \beta([\mathbb{1}_K\otimes|\psi\rangle], [A\otimes|\varphi\rangle])\\ &= \langle\psi|\Phi(A)\varphi\rangle\,.\end{aligned}$$

Define then

$$\pi : \mathcal{L}(K)\to\mathcal{L}(\tilde{H}), \qquad \pi(A)[B\otimes|\psi\rangle] = [(AB)\otimes|\psi\rangle],$$

and extended it by linearity. In order to check that this is a good definition, it is sufficient to argue that, for any family $(B_i)_{i=1}^d \subseteq \mathcal{L}(K)$ and $(|i\rangle)_{i=1}^d \subseteq H$, the following implication holds:

$$\sum_{i,j=1}^d B_i\otimes|i\rangle \sim 0 \quad\Rightarrow\quad \sum_{i,j}(AB_i)\otimes|i\rangle\sim 0.$$

More explicitly, it reads

$$\sum_{i,j=1}^d \langle j|\Phi(B_j^*B_i)i\rangle = 0 \quad\Rightarrow\quad \sum_{i,j=1}^d \langle j|\Phi(B_j^*A^*AB_i)i\rangle = 0.$$

To prove it, let $\lambda = \max\{\sigma(A^*A)\}\geq 0$, so that $\lambda\mathbb{1}_K - A^*A\geq 0$, hence we may represent $A^*A = \lambda\mathbb{1}_K - C^*C$, hence

$$\sum_{i,j=1}^d \langle j|\Phi(B_j^*A^*AB_i)i\rangle = \lambda\sum_{i,j=1}^d \langle j|\Phi(B_j^*B_i)i\rangle - \sum_{i,j=1}^d \langle j|\Phi(B_j^*C^*CB_i)i\rangle\,.$$

By complete monotonicity, we have

$$\sum_{i,j=1}^d \langle j|\Phi(B_j^*C^*CB_i)i\rangle \geq 0,$$

and by assumption

$$\sum_{i,j=1}^d \langle j|\Phi(B_j^*B_i)i\rangle = 0,$$

hence

$$\sum_{i,j=1}^d \langle j|\Phi(B_j^*A^*AB_i)i\rangle \leq 0,$$

but it also must be non-negative, by complete monotonicity, hence it equals 0.

It is then immediate to check that, $\pi(A_0A_1) = \pi(A_0)\pi(A_1)$ and $\pi(A^*) = \pi(A)^*$, since

$$\begin{aligned}\beta(\pi(A^*)[B_0\otimes|\psi_0\rangle], [B_1\otimes|\psi_1\rangle]) &= \beta([A^*B_0\otimes|\psi_0\rangle], [B_1\otimes|\psi_1\rangle])\\ &= \langle\psi_0|\Phi((A^*B_0)^*B_1)\psi_1\rangle\\ &= \langle\psi_0|\Phi(B_0^*(AB_1))\psi_1\rangle = \beta([B_0\otimes|\psi_0\rangle], [AB_1\otimes|\psi_1\rangle]).\end{aligned}$$

Finally, one checks the identity

$$\Phi(A) = U^*\pi(A)U. \tag{3.12}$$

Indeed, the right hand side applied to $|\psi\rangle\in H$ gives

$$U^*\pi(A)U|\psi\rangle = U^*\pi(A)[\mathbb{1}_K\otimes|\psi\rangle] = U^*[A\otimes|\psi\rangle] = \Phi(A)|\psi\rangle\,,$$

where the last identity follows from (3.11).

To obtain the Kraus representation for Φ, notice that it is sufficient to write the Kraus representation for $\pi : \mathcal{L}(K)\mapsto\mathcal{L}(\tilde{H})$ using Lemma 3.1 and then compose with U. Since $\dim(\tilde{H})\leq\dim(K)^2\dim(H)$, we also obtain that $|\mathcal{X}|\leq\dim(H)\dim(K)$. □

In fact, the proof of Lemma 3.1 yields that, up to isomorphisms, one can choose $\tilde{H} = K \otimes \mathbb{C}^{\mathcal{X}}$, and $\pi(A) = A \otimes \mathbb{1}_{\mathbb{C}^{\mathcal{X}}}$. Moreover, again up to an isometry of $\tilde{H}$, one can let $U|\psi\rangle = |\psi\rangle \otimes |0\rangle$, so that by duality one obtains the following formula, also called sometimes Stinespring representation of the quantum channel:

$$\Phi^\dagger(\rho) = \mathrm{tr}_{\mathbb{C}^{\mathcal{X}}}[V(\rho \otimes |0\rangle\langle 0|)V^*],$$

where $V : H \otimes \mathbb{C}^{\mathcal{X}} \to K \otimes \mathbb{C}^{\mathcal{X}}$ is unitary. Assuming $H = K$, this has the nice physical interpretation that $\Phi^\dagger$ is a composition of a global unitary transformation on the composite system $H \otimes \mathbb{C}^{\mathcal{X}}$, with a partial trace which "forgets" the auxiliary system and yields the reduced density operator on H.

3.7. **CP maps on C^*-algebras and Stinespring dilation theorem.** It is quite natural to extend the above notions of positivity and complete positivity to operators between C^*-algebras, $\mathcal{A}$, $\mathcal{B}$. Given a linear $\Phi : \mathcal{A} \to \mathcal{B}$, we say that Φ is positive if $\Phi(a)$ is positive whenever a is positive. For complete positivity, the simplest way is to ask that the analogue of (3.9) holds: we require that, for every $d \geq 1$, $(a_i)_{i=1}^d \subseteq \mathcal{A}$, $(b_i)_{i=1}^d \subseteq \mathcal{B}$, the sum

$$\sum_{i,j=1}^d b_i^* \Phi(a_i^* a_j) b_j \tag{3.13}$$

defines a positive element in $\mathcal{B}$.

Remark 3.3. If $\mathcal{A} = \mathcal{L}(H)$, $\mathcal{B} = \mathcal{L}(K)$, the above notion coincides with CP maps defined in the previous section. Indeed, given any $(|\psi_i\rangle)_{i=1}^d \subseteq K$, one simply chooses $b_i \in \mathcal{L}(K)$ such that $|\psi_i\rangle = b_i|\psi_0\rangle$ for a fixed $|\psi_0\rangle \in K$, hence obtaining (3.10) when testing positivity (3.13) with $|\psi_0\rangle$, i.e.,

$$\left\langle \psi_0 \middle| \sum_{i,j=1}^d b_i^* \Phi(a_i^* a_j) b_j \psi_0 \right\rangle \geq 0,$$

and viceversa, given $(b_i)_{i=1}^d$ and $|\psi_0\rangle$, one simply defines $|\psi_i\rangle = b_i|\psi_0\rangle$.

We also say that Φ is unital if $\Phi(\mathbb{1}_{\mathcal{A}}) = \mathbb{1}_{\mathcal{B}}$. A simple yet important example of completely positive maps is provided by states η on $\mathcal{A}$, i.e. continuous linear positive functionals on $\mathcal{A}$, by letting $\mathcal{B} = \mathbb{C}$. Complete positivity is straightforward: given $(a_i)_{i=1}^d \subseteq \mathcal{A}$ and complex numbers $(b_i)_{i=1}^d \subseteq \mathbb{C}$,

$$\sum_{i,j} b_i^* \eta(a_i^* a_j) b_j = \eta\left(\sum_{i,j} b_i^* a_i^* a_j b_j\right) = \eta(s^* s) \geq 0,$$

where we let $s = \sum_i b_i a_i$.

One may ask then if CP unital map admit a Kraus-like representation. By repeating the proof of (3.2) (with some caveats because of possible infinite dimensional spaces!) it is not difficult to prove the following result.

Theorem 3.4 (Stinespring dilation)**.** *Let $\mathcal{A}$ be a C^*-algebra, let H be a Hilbert space and $\mathcal{B}(H)$ denote the C^*-algebra of bounded linear operators on H. Given any CP unital map $\Phi : \mathcal{A} \to \mathcal{B}(H)$, there exist*

i) a Hilbert space $\tilde{H}$,
ii) an isometry $U : H \to \tilde{H}$,
iii) a $*$-homomorphism $\pi : \mathcal{A} \to \mathcal{B}(\tilde{H})$,

such that, for every $a \in \mathcal{A}$,

$$\Phi(a) = U^* \pi(a) U,$$

and

$$\{\pi(a)U\psi \,:\, a \in \mathcal{A}, \psi \in H\} \subseteq \tilde{H} \quad \text{is dense.}$$

Such a triple $(\tilde{H}, U, \pi)$ is unique up to isomorphisms.

We omit the details of the proof, but the main technical caveat is that $\tilde{H}$ is defined as the abstract completion of analogous object in the finite-dimensional case. One also has to be careful with checking that π is well-defined – one uses the fact that $\|a\|^2 \mathbb{1}_{\mathcal{A}} - a^*a$ has non-negative spectrum, hence it is positive, i.e., there exists $c \in \mathcal{A}$ such that $c^*c = \|a\|^2 \mathbb{1}_{\mathcal{A}} - a^*a$.

A relevant consequence of Stinespring's dilation theorem is the Gelfand-Naimark-Segal (GNS) construction, which connects the Hilbert space approach to quantum mechanics with the C^*-algebra one, showing indeed that one can represent any C^*-algebra $\mathcal{A}$ together with a chosen state η in terms of operators on a Hilbert space H and a state vector $|\psi\rangle$.

Theorem 3.5 (GNS)**.** *Let $\mathcal{A}$ be a C^*-algebra and let $\eta : \mathcal{A} \to \mathbb{C}$ be a state. Then, there exists*

i) a Hilbert space H,
ii) a unit norm vector $|\psi\rangle \in H$,
iii) and a $*$-homomorphism $\pi : \mathcal{A} \to \mathcal{B}(H)$

such that, for every $a \in \mathcal{A}$,

$$\eta(a) = \langle \psi | \pi(a) \psi \rangle \,,$$

and $\{\pi(a)\,|\psi\rangle\}_{a\in\mathcal{A}} \subseteq H$ is dense. Such a triple is unique up to isomorphisms.

To prove it, apply Theorem 3.4 to $\Phi(a) = \eta(a)$ and set $|\psi\rangle = U(1)$ (also relabel $\tilde{H}$ to H).

3.8. **Quantum Markov semigroups.** Given an (elementary) quantum system H and a quantum channel Φ from H into itself, one can consider the quantum analogue of a Markov chain evolution by taking a state $\rho = \rho_0$ and letting $\rho_{n+1} = \Phi(\rho_n)$, for $n \in \mathbb{N}$. Notice that, compared with the classical case, we are only describing the evolution of the marginals densities (in fact there is no satisfactory notion of a joint density in the quantum case). When $\Phi(\rho) = U\rho U^*$ is induced by a unitary U, it can be thought as the analogue of a classical dynamical system.

Almost always in physics, but also quite often in probability, one describes dynamics in continuous time, in terms of semigroups. The quantum analogue, called quantum Markov semigroup, is defined as a family $(\Phi^t)_{t\geq 0}$ of operators such that

(1) for every $t \geq 0$, Φ^t is a quantum channel from H into itself,
(2) (semigroup law) for every $s, t \geq 0$, $\Phi^t\Phi^s = \Phi^{s+t}$,
(3) (strong continuity) for every $A \in \mathcal{L}(H)$, $t \mapsto \Phi^t(A)$ is continuous.

One defines the *generator* as $L(A) = \lim_{t\to 0^+}(\Phi^t(A) - A)/t$. If H is finite-dimensional, then L is a well-defined bounded operator and admits an explicit structure that plays the role of the Kraus representation (the so-called Lindblad form). However, it is still an open problem to completely describe unbounded generators of quantum Markov semigroups (in infinite dimensional systems). A special but fundamental case is given by a theorem by Stone, which applies to unitary semigroups Φ^t, i.e., represented $(U_t)_{t\geq 0}$ such that $\Phi^t(\rho) = U_t\rho U_t^*$. One can prove (also in infinite dimensions) that $L(A) = -i[H, A]$ for a suitable self-adjoint densely defined operator.

3.9. **Exercises.**

Exercise 3.1 (Bell states)**.** The simplest example of *entangled* states is provided by so-called Bell states in a two-qubit composite system $H = \mathbb{C}^2 \otimes \mathbb{C}^2$, defined as follows:

$$\left|\Phi^+\right\rangle = \left(|0,0\rangle + |1,1\rangle\right)/\sqrt{2}, \quad \left|\Phi^-\right\rangle = \left(|0,0\rangle + |1,1\rangle\right)/\sqrt{2},$$
$$\left|\Psi^+\right\rangle = \left(|0,1\rangle + |1,0\rangle\right)/\sqrt{2}, \quad \left|\Psi^-\right\rangle = \left(|0,1\rangle - |1,0\rangle\right)/\sqrt{2}.$$

(1) Show that the four state vectors provide an orthonormal basis for the system.
(2) Show that each of the four pure states corresponding to the Bell vectors is not separable, hence entangled.

Exercise 3.2. Consider the Pauli operators σ_x, σ_y on a single-qubit system $\mathbb{C}^2$.

(1) Find the matrix representation (with respect to the computational basis in $\mathbb{C}^4 = \mathbb{C}^2 \otimes \mathbb{C}^2$) of the operators
$$A = \sigma_x \otimes \sigma_y, \quad \text{and} \quad B = \sigma_y \otimes \sigma_x.$$
(2) Prove that A, B are self-adjoint operators and compute their spectra.
(3) Compute $[A, B]$.
(4) Assume that the system is prepared in the Bell state $|\Phi^+\rangle$. What is the probability of observing 1 if we measure A?

Exercise 3.3 (Partial transpose)**.** Given finite-dimensional quantum systems H, K and an operator $A \in \mathcal{L}(H; K)$ define its *transpose* operator as $A^\tau : \mathcal{L}(K^*) \to \mathcal{L}(H^*)$ as
$$\tau(A) : \langle\varphi| \mapsto \tau(A)(\langle\varphi|) := \langle\varphi| A,$$
i.e., $\tau(A)(\langle\varphi|) = \langle\varphi| A$ is the linear functional on H given by
$$\langle\varphi| A : |\psi\rangle \mapsto \langle\varphi|A\psi\rangle .$$

(1) Fix orthonormal bases $(|i\rangle)_{i\in I} \subseteq K$ and $(|j\rangle)_{j\in J} \subseteq H$. Write the associated matrix representation
$$A = (A_{ij})_{i\in I, j\in J} = (\langle i|Aj\rangle)_{i\in I, j\in J}$$
and compare it with the matrix representation of A^τ with respect to the bases $(\langle i|)_{i\in I} \subseteq K^*$, $(\langle j|)_{j\in J} \subseteq H$.
(2) Prove that $A \mapsto \tau(A)$ is linear, and if $A \in \mathcal{O}(H)$ is an observable, then $A^\tau \in \mathcal{O}(H^*)$, and moreover if $A \geq 0$ then $\tau(A) \geq 0$ (i.e., the map τ is positive).
(3) Show however that already if $H = K = \mathbb{C}^2$, then τ is not completely positive (in particular the *partial transpose* $\tau \otimes \mathbb{1}_{\mathcal{L}(\mathbb{C}^2)}$ is not a positive map).

Exercise 3.4 (PPT criterion)**.** Let H, K be finite dimensional quantum systems. Denoting by $\tau : \mathcal{L}(H) \to \mathcal{L}(H^*)$ the transpose map (defined in the previous exercise), prove that if $\rho \in \mathcal{S}(H \otimes K)$ is separable, then its partial transpose $\tau \otimes \mathbb{1}_{\mathcal{L}(K)}$ is a density operator (in particular, it is positive). This motivates the so-called *positive partial trace* (PPT) sufficient criterion for entanglement: a state $\rho \in \mathcal{S}(H \otimes K)$ is entangled if its partial transpose $\tau \otimes \mathbb{1}_{\mathcal{L}(K)}(\rho)$ is not positive.

Do Bell states satisfy the PPT criterion?

4. Inequalities

In this section we collect some inequalities that play a relevant role in quantum information theory. We begin with the so-called uncertainty inequalities, that are perhaps one of the most popularized aspects of quantum mechanics. The central part of this section deals will be a family of operator monotonicity inequalities, which could be thought as quantum analogues of Jensen (or Hölder) inequality. We end the section with the so-called Lieb's concavity theorem (although we prove directly its monotonicity formulation), which will be in particular applied in the study of the properties of quantum entropy.

4.1. Uncertainty inequalities.

Recall from Section 2.2 that two *compatible* measurements can be always performed in any order and yield pairs of observable outcomes with well-defined joint probabilities. Uncertainty inequalities aim to quantify how this cannot be done when the measurements are incompatible. Usually, they are stated for pairs of observables as they amount to a lower bound for the product of the standard deviations when the quantum system is on a given state.

Given an elementary quantum system H and operators $X, Y \in \mathcal{L}(H)$, we define their commutator and anti-commutator as follows
$$[X, Y] = XY - YX \qquad \{X, Y\} = XY + YX,$$

so that the following identity holds:

$$XY = \frac{1}{2}\{X,Y\} + \frac{1}{2}[X,Y]. \tag{4.1}$$

Both the commutator and the anti-commutator are bilinear expressions with respect to X and Y, and satisfy

$$[X,Y]^* = [Y^*,X^*] = -[X^*,Y^*], \quad \{X,Y\}^* = \{X^*,Y^*\}.$$

Hence, if $X, Y \in \mathcal{O}(H)$ are observables, then $\{X,Y\}, i[X,Y] \in \mathcal{O}(H)$.

Given a density operator $\rho \in \mathcal{S}(H)$, recalling the notation $(X)_\rho = \mathrm{tr}[X\rho]$, write $\tilde{X} = X - (X)_\rho \mathbb{1}_H$, $\tilde{Y} = Y - (Y)_\rho \mathbb{1}_H$ for the "centred" observables and define the covariance between X and Y as follows:

$$\mathrm{Cov}_\rho(X,Y) = \frac{1}{2}(\left\{\tilde{X},\tilde{Y}\right\})_\rho.$$

As with classical random variables, $(X,Y) \mapsto \mathrm{Cov}_\rho(X,Y)$ is bilinear, symmetric and $\mathrm{Cov}_\rho(X,X) = \sigma^2_\rho(X) = (\tilde{X}^2)_\rho \geq 0$. If we define instead the commutation

$$\mathrm{Com}_\rho(X,Y) = \frac{1}{2}(i\left[\tilde{X},\tilde{Y}\right])_\rho = \frac{1}{2}(i\,[X,Y])_\rho,$$

bi-linearity still holds, but it is anti-symmetric:

$$\mathrm{Com}_\rho(X,Y) = -\,\mathrm{Com}_\rho(Y,X).$$

Moreover, using (4.1), we have the identity

$$\mathrm{Cov}_\rho(X,Y) - i\,\mathrm{Com}_\rho(X,Y) = \mathrm{tr}[\tilde{X}\tilde{Y}\rho]. \tag{4.2}$$

Given a finite family of observables $(X_i)_{i=1,\dots n} \subseteq \mathcal{O}(H)$, we introduce the real valued quantities

$$\mathrm{Cov}_{\rho,ij} = \mathrm{Cov}_\rho(X_i,X_j), \quad \mathrm{Com}_{\rho,ij} = \mathrm{Com}_\rho(X_i,X_j),$$

which we collect into two matrices $\mathrm{Cov}_\rho, \mathrm{Com}_\rho \in \mathbb{R}^{n\times n}$, called respectively the covariance matrix (which is symmetric) and the commutation matrix (which is anti-symmetric). We claim that the following inequalities hold:

$$\mathrm{Cov}_\rho \geq \pm i\,\mathrm{Com}_\rho, \tag{4.3}$$

(these are actually two inequalities, one for the right hand side with + sign, another with − sign). To prove it, notice that by (4.2),

$$\mathrm{Cov}_\rho \; -i\,\mathrm{Com}_\rho = (\mathrm{tr}[\tilde{X}_i\tilde{X}_j\rho])_{i,j=1,\dots,n},$$

which is positive: for every $(b_i)_{i=1}^d \in \mathbb{R}^d$,

$$\sum_{i,j=1}^n b_i b_j \mathrm{tr}[\tilde{X}_i\tilde{X}_j\rho] = \mathrm{tr}\left[\left(\sum_{i=1}^d b_i X_i\right)^2 \rho\right] \geq 0.$$

Similarly,

$$\mathrm{Cov}_\rho \; +i\,\mathrm{Com}_\rho = (\mathrm{tr}[\tilde{X}_j\tilde{X}_i\rho])_{i,j=1,\dots,n}.$$

is the transposed matrix (hence also positive).

To see how (4.3) is a form of uncertainty inequality, i.e., it provides a lower bound for the product of standard deviations, we specialize it to the case $n = 2$, i.e.,

$$\begin{pmatrix} \mathrm{Cov}_\rho(X,X) & \mathrm{Cov}_\rho(X,Y) \\ \mathrm{Cov}_\rho(X,Y) & \mathrm{Cov}_\rho(Y,Y) \end{pmatrix} \geq i \begin{pmatrix} 0 & \mathrm{Com}_\rho(X,Y) \\ \mathrm{Com}_\rho(X,Y) & 0 \end{pmatrix}.$$

Since $\mathrm{Cov}_\rho(X,X) = \sigma^2_\rho(X)$, $\mathrm{Cov}_\rho(Y,Y) = \sigma^2_\rho(Y)$, this can be equivalently stated as

$$\begin{pmatrix} \sigma^2_\rho(X) & b \\ \bar{b} & \sigma^2_\rho(Y) \end{pmatrix} \geq 0,$$

where $b = \mathrm{Cov}_\rho(X,Y) - i\,\mathrm{Com}_\rho(X,Y)$. By considering the determinant, this yields the so-called *Schrödinger-Robertson uncertainty relation*

$$\sigma_\rho^2(X)\sigma_\rho^2(Y) \geq |b|^2 = |\mathrm{Cov}_\rho(X,Y)|^2 + |\mathrm{Com}_\rho(X,Y)|^2 .$$

Dropping the covariance term in the right hand side and taking the square root yields a version of Heisenberg uncertainty inequality

$$\sigma_\rho(X)\sigma_\rho(Y) \geq |\mathrm{Com}_\rho(X,Y)| .$$

Dropping instead the commutant term yields a version of the classical bound of the covariance in terms of the product of standard deviations

$$|\mathrm{Cov}_\rho(X,Y)| \leq \sigma_\rho(X)\sigma_\rho(Y).$$

4.2. **Monotonicity inequalities.** Recalling that quantum channels are the counterparts of classical Markov kernels $N = (N(\omega,x)_{\omega\in\Omega,x\in\mathcal{X}}$, it is natural (and useful in applications) to investigate whether the analogues of common functional inequalities hold true.

For example, since for every $\omega \in \Omega$, $N(\omega,\cdot)$ is a probability distribution, then for every function $f : \mathcal{X} \to \mathbb{C}$, Cauchy-Schwarz inequality easily yields the following inequality between functions on Ω:

$$|Nf|^2 \leq N(|f|^2),$$

(recall that we define $(Ng)(\omega) = \sum_{x\in\mathcal{X}} g(x)N(\omega,x)$). Therefore, it is natural to ask whether the following analogue holds, for a CP unital map $\Phi : \mathcal{L}(H) \to \mathcal{L}(\tilde{H})$ (dual to a quantum channel from a quantum system $\tilde{H}$ to H):

$$\Phi(A)^*\Phi(A) \leq \Phi(A^*A) \tag{4.4}$$

where we naturally interpret $A^*A = |A|^2$ for any operator $A \in \mathcal{L}(K)$. This is indeed the case, and the inequality is also called Kadison-Schwarz inequality. To prove it, we use complete positivity and apply $\Phi \otimes \mathbb{1}_{\mathcal{L}(\mathbb{C}^2)}$ to the positive operator acting on $H \otimes \mathbb{C}^2$, represented by the block operator as

$$M = \begin{pmatrix} A^*A & A^* \\ A & \mathbb{1}_H \end{pmatrix}. \tag{4.5}$$

The fact that M is positive should be easily seen from the block representation

$$M = \begin{pmatrix} A^* \\ \mathbb{1}_H^* \end{pmatrix} \begin{pmatrix} A & \mathbb{1}_H \end{pmatrix}.$$

However, since we are going to use more general 2×2 block operators in this section, we directly give a general criterion for positivity, based on the so-called Schur complement.

Lemma 4.1. *Let H be a finite dimensional Hilbert space, $X, Y \in \mathcal{O}(H)$, $K \in \mathcal{L}(H)$ and $Y \geq 0$ and invertible. Then, the operator $M \in \mathcal{O}(H\otimes\mathbb{C}^2)$ represented by the block matrix*

$$M = \begin{pmatrix} X & K \\ K^* & Y \end{pmatrix}$$

is positive if and only if its Schur complement

$$X - KY^{-1}K^* \in \mathcal{O}(H)$$

is positive.

Proof. M is positive if and only if, for every pair $|\psi_0\rangle$, $|\psi_1\rangle \in H$, letting $|v\rangle = |\psi_0, 0\rangle + |\psi_1, 1\rangle \in H\otimes\mathbb{C}^2$, one has $\langle v|Mv\rangle \geq 0$. Explicitly, this amounts to the inequality

$$\langle\psi_0|X\psi_0\rangle + \langle\psi_1|K^*\psi_0\rangle + \langle\psi_0|K\psi_1\rangle + \langle\psi_1|Y\psi_1\rangle \geq 0. \tag{4.6}$$

Assume that M is positive. To argue that the Schur complement is positive, given any $|\psi\rangle \in H$, choose $|\psi_0\rangle = |\psi\rangle$ and $|\psi_1\rangle = -Y^{-1}K^*|\psi_0\rangle$ in (4.6), so that it becomes

$$\langle\psi|X\psi\rangle - \langle\psi|KY^{-1}K^*\psi\rangle \geq 0,$$

which is the thesis.

Viceversa, assume that the Schur complement is positive. Then,

$$\langle \psi_0 | X \psi_0 \rangle \geq \left\langle K\psi_0 | Y^{-1} K^* \psi_0 \right\rangle ,$$

which used in (4.6) yields that it is sufficient to argue that

$$M' = \begin{pmatrix} KY^{-1}K^* & K \\ K^* & Y \end{pmatrix} \geq 0.$$

Positivity of M' then follows from the block factorization

$$M = \begin{pmatrix} KY^{-1/2} \\ Y^{1/2} \end{pmatrix} \begin{pmatrix} Y^{-1/2}K^* & Y^{1/2} \end{pmatrix} .$$

Indeed, working in (4.6) minding the above factorization, we obtain that

$$\langle v | M' v \rangle = \left\| Y^{-1/2} K^* \psi_0 + Y^{1/2} \psi_1 \right\|^2 \geq 0. \qquad \square$$

Back to the proof of (4.4), by complete positivity, the operator $\Phi \otimes \mathbb{1}_{\mathcal{L}(\mathbb{C}^2)}$ applied to the positive M defined in (4.5) yields a positive operator, which can be represented by the block matrix

$$\Phi \otimes \mathbb{1}_{\mathcal{L}(\mathbb{C}^2)}(M) = \begin{pmatrix} \Phi(A^*A) & \Phi(A^*) \\ \Phi(A) & \mathbb{1}_{\tilde{H}} \end{pmatrix},$$

where we used the fact that $\Phi(\mathbb{1}_H) = \mathbb{1}_{\tilde{H}}$. Using Lemma 4.1 again, we conclude that inequality (4.4) holds.

Before we proceed with further inequalities, one should be aware that many "natural" ones, i.e., valid for the case of functions, fail when extended naively to operators. As a striking example, it is not true in general that $A \leq B$ implies $A^2 \leq B^2$. However, there are several others that are *operator monotone*, when acting on positive operators. For example, if A, $B \in \mathcal{O}_>(H)$ are self-adjoint positive and invertible operators on H, then

$$A \leq B \quad \Rightarrow \quad A^{-1} \geq B^{-1}. \tag{4.7}$$

The proof is a simple application of Lemma 4.1. Consider the operator on $H \otimes \mathbb{C}^2$ represented as the block operator

$$M = \begin{pmatrix} B & \mathbb{1}_H \\ \mathbb{1}_H & A^{-1} \end{pmatrix}. \tag{4.8}$$

Since $B \geq A = \mathbb{1}_H (A^{-1})^{-1} \mathbb{1}_H$, the Schur complement is positive hence $M \geq 0$ is positive. But of course we can also apply the criterion reversing the roles of A and B (this would correspond to invert the order of the vectors in the standard basis of $\mathbb{C}^2$), or equivalently

$$M' = \begin{pmatrix} A^{-1} & \mathbb{1}_H \\ \mathbb{1}_H & B \end{pmatrix} \geq 0.$$

Again by (4.1), this time we obtain the inequality

$$A^{-1} \geq \mathbb{1}_H B^{-1} \mathbb{1}_H = B^{-1},$$

which is (4.7). Let us further notice that, having represented the condition $A \geq B$ in terms of positivity of M in (4.8), yields also a monotonicity inequality with respect to composition with CP unital maps $\Phi : \mathcal{L}(H) \to \mathcal{L}(\tilde{H})$. Indeed, if M in (4.7) is positive, then composing with $\Phi \otimes \mathbb{1}_{\mathcal{L}(\mathbb{C}^2)}$ yields

$$\begin{pmatrix} \Phi(B) & \mathbb{1}_{\tilde{H}} \\ \mathbb{1}_{\tilde{H}} & \Phi(A^{-1}) \end{pmatrix} \geq 0,$$

hence, again by Lemma 4.1, we proved that

$$A \leq B \quad \Rightarrow \quad \Phi(B) \geq \Phi(A^{-1})^{-1}.$$

(assuming that $\Phi(A^{-1})$ is invertible). In particular, when $B = A$ and using the already established monotonicity of the inverse, this gives

$$\Phi(A)^{-1} \leq \Phi(A^{-1}), \tag{4.9}$$

which classically would be an application of Jensen inequality with the convex function $z \mapsto z^{-1}$ for $z > 0$.

Another elementary example of operator monotone function is the square root: given positive operators $A \in \mathcal{O}_{\geq}(H)$, it holds

$$A \leq B \Rightarrow \sqrt{A} \leq \sqrt{B}. \tag{4.10}$$

To prove it, assume for simplicity that A is invertible, and write $A = X^2$, $B = Y^2$ for positive X, Y, so that the statement becomes

$$X^2 \leq Y^2 \quad \Rightarrow \quad X \leq Y.$$

Multiplying both sides of the assumption by X^{-1} (which preserves positivity), we have the inequality

$$\mathbb{1}_H \leq X^{-1}Y^2X^{-1} = K^*K, \tag{4.11}$$

where we define $K = YX^{-1}$. Since $X^{-1/2}KX^{1/2} = X^{-1/2}YX^{-1/2}$, and conjugation preserves the spectrum, it follows that

$$\sigma(K) = \sigma(X^{-1/2}YX^{-1/2}).$$

If λ belongs to the set above, it is an eigenvalue for K, for some eigenvector $|\psi\rangle \in H$, i.e., $K|\psi\rangle = \lambda|\psi\rangle$. Then, by (4.11) we deduce that

$$|\lambda|^2 \langle\psi|\psi\rangle = \langle K\psi|K\psi\rangle = \langle\psi|K^*K\psi\rangle \geq \langle\psi|\psi\rangle,$$

hence $|\lambda|^2 \geq 1$. On the other side, $X^{-1/2}YX^{-1/2}$ is a positive operator, hence $\lambda \geq 0$, and we conclude that $\lambda \geq 1$. This inequality holds for any eigenvalue $\lambda \in \sigma(X^{-1/2}YX^{-1/2})$, thus

$$X^{-1/2}YX^{-1/2} \geq \mathbb{1}_H.$$

Multiplying both sides by $X^{1/2}$, we obtain the thesis[9].

In view of (4.5) and (4.9), it seems natural to conjecture that, for any CP unital map $\Phi : \mathcal{L}(H) \to \mathcal{L}(\tilde{H})$, and $A \in \mathcal{O}_{\geq}(H)$,

$$\sqrt{\Phi(A)} \geq \Phi(\sqrt{A}). \tag{4.12}$$

This is indeed the case, as it follows by writing first (4.5) with $\sqrt{A}$ instead of A, then taking the square root both sides.

The square root example[10] turns out to be special cases of a generalized family of *t-weighted geometric means* for operators. For $t \in [0,1]$, and positive operators A, $B \in \mathcal{O}_{\geq}(H)$, we define

$$A\sharp_t B = A^{1/2}(A^{-1/2}BA^{-1/2})^t A^{1/2},$$

where $(A^{-1/2}BA^{-1/2})^t$ is defined via functional calculus. Notice that the above definition is actually valid only if A is invertible, but we are going to prove below in (4.15) that we can exchange the roles of A and B (up to replacing t with $1-t$). In fact, for our purposes, we may always assume that both A and B are invertible.

The operator $A\sharp_t B$ generalizes the geometric mean: to see it, consider the case $H = \mathbb{C}$, so that A, B are positive real numbers. Then,

$$A\sharp_t B = A^{1-t}B^t. \tag{4.13}$$

Given $U \in \mathcal{U}(H)$ unitary, it is also straightforward to check that

$$(U^*AU)\sharp_t(U^*BU) = U^*(A\sharp_t B)U.$$

[9] Is there a more streamlined proof relying e.g. upon Lemma 4.1?

[10] but actually also the inverse, see Exercise 4.6

It follows that, if A and B commute, since they can be simultaneously diagonalized through the same unitary U, we have that (4.13) also holds.

Notice that

$$\mathbb{1}_H \sharp_t A = A^t.$$

We collect two useful algebraic identities concerning operator means: for $s, t \in [0,1]$,

$$A\sharp_s(A\sharp_t B) = A\sharp_{st} B, \tag{4.14}$$

which can be proved by straightforward substitution with the definitions, and

$$A\sharp_t B = B\sharp_{1-t} A. \tag{4.15}$$

To prove it, assume for simplicity that both A and B are invertible, and let

$$X = A^{-1/2} B A^{-1/2}, \quad Y = B^{-1/2} A B^{-1/2}, \quad U = A^{1/2} B^{-1/2} Y^{-1/2},$$

so that $X, Y \in \mathcal{O}(H)$ and $U \in \mathcal{U}(H)$, since

$$UU^* = A^{1/2} B^{-1/2} Y^{-1} B^{-1/2} A^{1/2} = A^{1/2} B^{-1/2} B^{1/2} A^{-1} B^{1/2} B^{-1/2} A^{1/2} = \mathbb{1}_H.$$

Moreover,

$$\begin{aligned} UY^{-1}U^* &= A^{1/2} B^{-1/2} Y^{-2} B^{-1/2} A^{1/2} = A^{1/2} B^{-1/2} (B^{1/2} A^{-1} B^{1/2})^2 B^{-1/2} A^{1/2} \\ &= A^{-1/2} B A^{-1/2} = X, \end{aligned}$$

so that, by spectral calculus,

$$\begin{aligned} X^t &= (UY^{-1}U^*)^t = UY^{-t}U^* \\ &= A^{1/2} B^{-1/2} Y^{-t-1} B^{-1/2} A^{1/2}. \end{aligned}$$

Multiplying both sides by $A^{1/2}$ yields (4.15), since

$$\begin{aligned} A\sharp_t B &= A^{1/2} X^t A^{1/2} \\ &= A B^{-1/2} Y^{-t-1} B^{-1/2} A = B^{1/2} Y Y^{-t-1} Y B^{1/2} \\ &= B^{1/2} Y^{1-t} B^{1/2} = B\sharp_{1-t} A. \end{aligned}$$

We now establish monotonicity inequalities for these operators means (we refer e.g. to [Car10] for further results on operator inequalities).

Proposition 4.2. *Let H, K be finite dimensional Hilbert spaces, let $A, A', B, B' \in \mathcal{O}_{\geq}(H)$, $\Phi : \mathcal{L}(H) \to \mathcal{L}(K)$ be CP, and $t \in [0,1]$. It holds*

$$A' \geq A, B' \geq B \quad \Rightarrow \quad A'\sharp_t B' \geq A\sharp_t B, \tag{4.16}$$

and

$$\Phi(A)\sharp_t \Phi(B) \geq \Phi(A\sharp_t B). \tag{4.17}$$

Proof. Let us begin with the case $t = 1/2$. We argue that

$$A\sharp_{1/2} B \geq T \tag{4.18}$$

if and only if the following holds:

$$\text{there exists } W \in \mathcal{O}_{\geq}(H) \text{ such that } \quad M = \begin{pmatrix} A & W \\ W & B \end{pmatrix} \geq 0, \quad \text{and} \quad W \geq T. \tag{4.19}$$

In other words, $A\sharp_{1/2} B$ is the largest $W \in \mathcal{O}(H)$ that one can take while keeping M positive. Such a characterization in terms of a semidefinite linear problem turns out to be extremely useful to obtain the thesis, but also for computational aspects: similar representations in fact hold for rational $t \in [0,1]$, see e.g. [FS17].

For simplicity (but without loss of generality), assume that both A and B are invertible. If (4.18) holds, choose $W = A\sharp_t B$ and notice that $M \geq 0$ by Lemma 4.1, since

$$WB^{-1}W = A^{1/2} (A^{-1/2} B A^{-1/2})^{1/2} A^{1/2} B^{-1} A^{1/2} (A^{-1/2} B A^{-1/2})^{1/2} = A^{1/2} A^{1/2} = A.$$

To prove the converse, assume (4.19), so that $M \geq 0$ and Lemma 4.1 entails

$$B \geq WA^{-1}W \Rightarrow A^{-1/2}BA^{-1/2} \geq (A^{-1/2}WA^{-1/2})^2.$$

By (4.10), it follows that

$$\left(A^{-1/2}BA^{-1/2}\right)^{1/2} \geq A^{-1/2}WA^{-1/2} \Rightarrow A\sharp_{1/2}B = A^{1/2}\left(A^{-1/2}BA^{-1/2}\right)^{1/2}A^{1/2} \geq W,$$

hence (4.18), since $W \geq T$.

Let us now deduce the thesis for the case $t = 1/2$. From (4.19), we see immediately that (4.16) holds, since it always holds

$$M' = \begin{pmatrix} A' & W \\ W & B' \end{pmatrix} = \begin{pmatrix} A'-A & 0 \\ 0 & B'-B \end{pmatrix} + \begin{pmatrix} A & W \\ W & B \end{pmatrix} \geq \begin{pmatrix} A & W \\ W & B \end{pmatrix} = M.$$

Choosing $W = A\sharp_{1/2}B$, we have that $M \geq 0$, hence $M' \geq 0$ and by (4.19) we deduce that $A'\sharp_t B' \geq A\sharp_t B$. Similarly, let $T = A\sharp_{1/2}B$ and apply (4.19), which yields $W \geq T$ (actually $W = T$ in this case) such that $M \geq 0$. By positivity, $\Phi(W) \geq \Phi(T)$, and by complete positivity,

$$\Phi \otimes \mathbb{1}_{\mathcal{L}(\mathbb{C}^2)}(M) = \begin{pmatrix} \Phi(A) & \Phi(W) \\ \Phi(W) & \Phi(B) \end{pmatrix} \geq 0,$$

hence, again by (4.19), $\Phi(A)\sharp_{1/2}\Phi(B) \geq \Phi(A\sharp_{1/2}B)$.

To prove the general case $t \in [0,1]$, we notice first that it is sufficient to establish (4.16) and (4.17) for dyadic $t = p/2^n$, $p \in \{0,\dots,2^n\}$, for every $n \geq 1$. Indeed, both inequalities involve continuous functions of t, hence, by density, the thesis will then follow from dyadic t's to the whole interval.

Next, we argue by induction with respect to n, having already established the case $n = 1$, (essentially the case $t = 1/2$, the other cases being trivial), we now assume that the thesis, i.e., (4.16) and (4.17), holds true for every $t = p/2^{n-1}$, $p \in \{0,\dots,2^{n-1}\}$ and prove it for every $t = q/2^n$, $q \in \{0,\dots,2^n\}$. Actually it is enough to prove it for such t's in the interval $[0,1/2]$, i.e., such that $q \leq 2^{n-1}$ because then the case $t \in [1/2,1]$ follows by applying (4.15) (reversing the roles of A and B). Assuming $t \in [0,1/2]$, using (4.14), we write

$$A\sharp_t B = A\sharp_{1/2}(A\sharp_{2t}B),$$

so that $2t = q/2^{n-1}$ and we can apply the inductive assumption: given $A' \geq A$, $B' \geq B$, we have

$$A'\sharp_{2t}B' \geq A\sharp_{2t}B \quad \text{and} \quad \Phi(A)\sharp_{2t}\Phi(B) \geq \Phi(A\sharp_{2t}B).$$

Using also the thesis in the case $t = 1/2$, we deduce

$$A'\sharp_t B' = A'\sharp_{1/2}(A'\sharp_{2t}B') \geq A\sharp_{1/2}(A\sharp_{2t}B) = A\sharp_t B$$

and

$$\begin{aligned}\Phi(A)\sharp_t\Phi(B) &= \Phi(A)\sharp_{1/2}(\Phi(A)\sharp_{2t}\Phi(B)) \\ &\geq \Phi(A)\sharp_{1/2}(\Phi(A\sharp_{2t}B)) \\ &\geq \Phi(A\sharp_{1/2}A\sharp_{2t}B) = \Phi(A\sharp_t B),\end{aligned}$$

and the proof is completed. □

4.3. **Lieb's concavity theorem.** In this final section, we prove a fundamental inequality involving the trace of operators and their transformation through quantum channels (or their adjoint CP unital maps). Historically, it is referred as Lieb's concavity theorem, although we are going to directly state and prove it in a *monotonicity* formulation, inspired by the exposition in [Car22]. It is then straightforward to deduce from it the usual concavity formulation, see Exercise 4.8.

Theorem 4.3 (Lieb's concavity theorem, monotonicity version)**.** *Let H, $\tilde{H}$ be finite dimensional quantum systems and $\Phi : \mathcal{L}(\tilde{H}) \to \mathcal{L}(H))$ be CP and unital, so that $\Phi^\dagger$ is a quantum channel from H to $\tilde{H}$. Let $X, Y \in \mathcal{O}_{\geq}(H)$ be positive, $K \in \mathcal{L}(\tilde{H})$ and $t \in [0,1]$. Then,*

$$\mathrm{tr}[\Phi(K)^* X^{1-t} \Phi(K) Y^t] \leq \mathrm{tr}[K^* \Phi^\dagger(X)^{1-t} K \Phi^\dagger(Y)^t]. \tag{4.20}$$

In the case $K = \mathbb{1}_{\tilde{H}}$, we have $\Phi(\mathbb{1}_{\tilde{H}}) = \mathbb{1}_H$, hence (4.20) becomes

$$\mathrm{tr}[\Phi^\dagger(X^{1-t} Y^t)] \leq \mathrm{tr}[\Phi^\dagger(X)^{1-t} \Phi^\dagger(Y)^t], \tag{4.21}$$

where we used that

$$\mathrm{tr}[X^{1-t} Y^t] = \mathrm{tr}[\Phi(\mathbb{1}_H) X^{1-t} Y^t] = \mathrm{tr}[\Phi^\dagger(X^{1-t} Y^t)].$$

In this form, the inequality looks like a traced Hölder inequality with exponents $p = 1/(1-t)$ and dual $p' = 1/t$. Moreover, by Proposition 4.2, we already know that

$$\Phi^\dagger(X \sharp_t Y) \leq \Phi^\dagger(X) \sharp_t \Phi^\dagger(Y),$$

which looks like (4.21), without the trace and replacing $X^{1-t} Y^t$, $\Phi^\dagger(X)^{1-t} \Phi^\dagger(Y)^t$ with the corresponding operator geometric means. Thus, we are very close to (4.21) (even without the trace) but X and Y do not necessarily commute. The main idea is to move to a "higher" level (i.e., think of the spaces $\mathcal{L}(H)$ and $\mathcal{L}(\tilde{H})$ as the basic Hilbert spaces) so that some form of commutativity is restored.

Proof. Recall that we consider $\mathcal{L}(H)$ and $\mathcal{L}(\tilde{H})$ as Hilbert spaces endowed with the Hilbert-Schmidt scalar product ($\langle A|B\rangle = \mathrm{tr}[A^* B]$). Thus, write $|X\rangle, |Y\rangle \in \mathcal{L}(H)$, and $|K\rangle \in \mathcal{L}(\tilde{H})$. The operator Φ is linear from $\mathcal{L}(H)$ to $\mathcal{L}(\tilde{H})$, hence we write $\Phi \in \mathcal{L}(\mathcal{L}(H); \mathcal{L}(\tilde{H})) = \mathcal{L}(\mathcal{L}(H))$, i.e., in Dirac notation, $\Phi |K\rangle = |\Phi(K)\rangle$. We use it together with its adjoint (which is indeed $\Phi^\dagger$, but we write it as Φ^* here) to define a map

$$\tilde{\Phi} : \mathcal{L}(\mathcal{L}(H)) \to \mathcal{L}(\mathcal{L}(\tilde{H})), \quad A \mapsto \tilde{\Phi}(A) = \Phi^* A \Phi,$$

which is clearly CP, since it is already given in terms of a single Kraus operator (notice that here we do not even need that Φ itself is CP).

We introduce two further operators: the left-multiplication by $X \in \mathcal{L}(H)$

$$L_X \in \mathcal{L}(\mathcal{L}(H)), \quad L_X : \mathcal{L}(H) \to \mathcal{L}(H), \quad L_X |A\rangle \mapsto |XA\rangle,$$

and the right multiplication by $Y \in \mathcal{L}(H)$,

$$R_Y \in \mathcal{L}(\mathcal{L}(H)), \quad R_Y : \mathcal{L}(H) \to \mathcal{L}(H), \quad R_Y |A\rangle \mapsto |AY\rangle.$$

Clearly, $L_X L_{X'} = L_{XX'}$ while $R_Y R_{Y'} = R_{Y'Y}$. It holds $L_X^* = L_{X^*}$, since

$$\langle L_X A|A'\rangle = \mathrm{tr}[(XA)^* A'] = \mathrm{tr}[A^*(X^* A')] = \langle A|L_{X^*} A'\rangle,$$

and similarly $R_Y^* = R_{Y^*}$. Therefore, if $X \in \mathcal{O}_{\geq}(H)$, we have that $L_X = L_{\sqrt{X}}^* L_{\sqrt{X}} \geq 0$, and similarly $R_Y \geq 0$ if $Y \in \mathcal{O}_{\geq}(H)$. Moreover, any operator L_X commutes with any R_Y, since

$$L_X R_Y |A\rangle = L_X |AY\rangle = |XAY\rangle = R_Y |XA\rangle = R_Y L_X |A\rangle,$$

so that we can use (4.13) with $A = L_X$, $B = R_Y$, to obtain

$$(L_X) \sharp_t (R_Y) = L_X^{1-t} R_Y^t = L_{X^{1-t}} R_{Y^t},$$

where the identities $L_X^{1-t} = L_{X^{1-t}}$, $R_Y^t = R_{Y^t}$ follow from spectral calculus (one can argue that $f(L_X) = L_{f(X)}$ and similarly $f(R_Y) = R_{f(Y)}$ for a general f).

For operators $\tilde{X}, \tilde{Y} \in \mathcal{L}(\tilde{H})$, we define similarly $L_{\tilde{X}}$, $R_{\tilde{Y}}$. The thesis follows from Proposition 4.2 in this framework, with the operators

$$A = \tilde{\Phi}(L_X), A' = L_{\Phi^\dagger(X)}, B = \tilde{\Phi}(R_Y), B' \in R_{\Phi^\dagger(Y)} \in \mathcal{L}(\tilde{H}).$$

Indeed, notice first that $L_{\Phi^\dagger(X)} \geq \tilde{\Phi}(L_X)$, since

$$\begin{aligned}\left\langle K|\tilde{\Phi}(L_X)K\right\rangle &= \langle K|\Phi^* L_X \Phi(K)\rangle = \langle \Phi(K)|X\Phi(K)\rangle \\ &= \mathrm{tr}[\Phi(K)^* X \Phi(K)] = \mathrm{tr}[X^{1/2}\Phi(K)\Phi(K)^* X^{1/2}] \\ &\leq \mathrm{tr}[X\Phi(KK^*)] = \mathrm{tr}[K^*\Phi^\dagger(X)K] \\ &= \left\langle K|L_{\Phi^\dagger(X)}K\right\rangle,\end{aligned}$$

where the inequality follows from (4.4) (with $A = K^*$), multiplying both sides by $X^{1/2}$ and taking the trace. Similarly, we have $R_{\Phi^\dagger(Y)} \geq \tilde{\Phi}(R_Y)$. Therefore, combining (4.17) and (4.16), we have

$$\begin{aligned}L_{\Phi^\dagger(X)^{1-t}} R_{\Phi^\dagger(Y)^t} &= L_{\Phi^\dagger(X)} \sharp_t R_{\Phi^\dagger(Y)} \\ &\geq \tilde{\Phi}(L_X)\sharp_t \tilde{\Phi}(R_X) \geq \tilde{\Phi}(L_X \sharp_t R_Y). \\ &= \Phi^* L_{X^{1-t}} R_{Y^t} \Phi.\end{aligned}$$

To conclude, it is sufficient to take the scalar product with $K \in \mathcal{L}(\tilde{H})$. We have,

$$\left\langle K|L_{\Phi^\dagger(X)^{1-t}} R_{\Phi^\dagger(Y)^t} K\right\rangle = \mathrm{tr}[K^*\Phi^\dagger(X)^{1-t} X \Phi^\dagger(Y)^t],$$

while

$$\langle K|\Phi^* L_{X^{1-t}} R_{Y^t}\Phi K\rangle = \left\langle \Phi(K)|X^{1-t}\Phi(K)Y^t\right\rangle = \mathrm{tr}[\Phi(K)^* X^{1-t}\Phi(K)Y^t]. \qquad \square$$

4.4. **Exercises.**

Exercise 4.1 (Uncertainty inequality for Pauli operators)**.** Consider a pure density operator $\rho \in \mathcal{S}(\mathbb{C}^2)$ on a single qubit system and write explicitly (4.3) for the Pauli operators, in terms of the vector $b = b(\rho)$ of the Bloch parametrization (2.7). Investigate when equality may occur. What about equality cases in the Schrödinger-Robertson inequality for a pair of Pauli operators?

Exercise 4.2 (Lieb-Ruskai monotonicity theorem)**.** Let H, $\tilde{H}$ be finite-dimensional quantum systems, $\Phi : \mathcal{L}(H) \to \mathcal{L}(\tilde{H})$ be CP , $K \in \mathcal{L}(H)$ and $X \in \mathcal{O}_{>0}(H)$ be positive and invertible such that $\Phi(X)$ is also invertible. Then,

$$\Phi(K)^*\Phi(X)^{-1}\Phi(K) \leq \Phi(K^* X^{-1} K).$$

Exercise 4.3. Show that $f(A) = A^2$ is not operator monotone on $\mathcal{O}_{\geq}(\mathbb{C}^2)$, i.e., it does not hold in general

$$A \leq B \quad \Rightarrow \quad A^2 \leq B^2.$$

Exercise 4.4. Let H be a finite dimensional Hilbert space, $X, Y \in \mathcal{O}_{>}(H)$, $K \in \mathcal{L}(H)$. Then, the operator $M \in \mathcal{O}(H \otimes \mathbb{C}^2)$ represented by the block matrix

$$M = \begin{pmatrix} X & K \\ K^* & Y \end{pmatrix}$$

is positive if and only if there exists $Z \in \mathcal{L}(H)$ with operator norm $\|Z\| \leq 1$ (i.e. $Z^*Z \leq \mathbb{1}_H$) such that $K = \sqrt{X} Z \sqrt{Y}$.

Exercise 4.5. Given A, $B \in \mathcal{O}_{\geq}(H)$, s, $t \in [0,1]$, show that

$$(A\sharp_s B)\sharp_t B) = A\sharp_{s+t-st} B.$$

Exercise 4.6. One can extend the definition of $A\sharp_t B$ for any $t \in \mathbb{R}$, provided that A, $B \in \mathcal{O}_{>}(H)$ are positive and invertible. In turns out that monotonicity inequalities hold true also in the range $t \in [-1, 0] \cup [1, 2]$ (with a reverse inequality than the case $t \in [0, 1]$.

(1) Show that (4.15) and (4.14) hold for every $t \in \mathbb{R}$.

(2) Show that, for every $A, B, C \in \mathcal{O}_>(H)$, s, $t \in \mathbb{R}$,

$$C\sharp_s A \leq C\sharp_t B \quad \Leftrightarrow \quad C\sharp_{-s} A \geq C\sharp_{-t} A$$

(3) Show that, for $t \in [0,1]$, given A, B, $T \in \mathcal{O}_>(H)$, the inequality

$$A\sharp_{-t} B \leq T$$

is equivalent to the following condition:

$$\text{there exists } W \in \mathcal{O}(H) \text{ such that } A\sharp_t B \geq W \text{ and } M = \begin{pmatrix} T & A \\ A & W \end{pmatrix} \geq 0.$$

(Hint: write $A\sharp_t B = A\sharp_{-1}(A\sharp_t B)$ and notice that $A\sharp_t B = AB^{-1}A$.)

(4) Deduce that, for $t \in [-1,0] \cup [0,1]$ and for every CP map $\Phi : \mathcal{L}(H) \to \mathcal{L}(K)$ and A, $B \in \mathcal{O}_>(H)$, the inequality

$$\Phi(A\sharp_t B) \geq \Phi(A)\sharp_t \Phi(B)$$

holds (provided that $\Phi(A)$, $\Phi(B)$ are invertible).

Exercise 4.7 (Lieb's theorem for negative exponents)**.** Let H, $\tilde{H}$ be finite dimensional quantum systems and $\Phi : \mathcal{L}(H) \to \mathcal{L}(\tilde{H})$) be a quantum channel from H to $\tilde{H}$. Let X, $Y \in \mathcal{O}_>(H)$ be positive, $K \in \mathcal{L}(H)$ and $t \in [0,1]$. Then,

$$\mathrm{tr}[\Phi(K)^* \Phi(X)^{t-1} \Phi(K) \Phi(Y)^{-t}] \leq \mathrm{tr}[K^* X^{1-t} K Y^t], \tag{4.22}$$

provided that $\Phi(X)$, $\Phi(Y)$ are invertible.

Exercise 4.8. Deduce from Theorem 4.3 the original concavity formulation of Lieb's theorem: for $t \in [0,1]$, $K \in \mathcal{L}(H)$, the map on $\mathcal{O}_\geq(H) \times \mathcal{O}_\geq(H)$

$$(X, Y) \mapsto \mathrm{tr}[K^* X^{1-t} K Y^t]$$

is jointly concave. (*Hint: consider the partial tensor map $\Phi(A) = A \otimes \mathbb{1}_{\mathbb{C}^2}$ and deduce mid-point concavity*)

Exercise 4.9. Consider a state η on the Weyl algebra and let $f : \mathbb{R}^2 \to \mathbb{C}$, $f(r,s) = \eta(W(r,s))$ denote its characteristic function, which we assume to be twice differentiable at 0 (by 2.13 assume without loss of generality with $\nabla f(0) = 0$). Recall also the notation (2.8) for the symplectic form ω from 2.14.

(1) Prove that, for every $u, v \in \mathbb{R}^2$,

$$f(u-v) e^{\frac{i}{2}\omega(v,u)} - 2f(0) + f(v-u) e^{\frac{i}{2}\omega(u,v)}.$$

(2) Prove that $\Sigma := -\nabla^2 f(0)$ is a real symmetric matrix and that

$$\Sigma \geq i\frac{\omega}{2}. \tag{4.23}$$

(*Hint: replace $u \mapsto \varepsilon u$, $v \mapsto \varepsilon$ in the above point to deduce that a certain block matrix is positive. Then, use the equivalence between (4.18) and (4.19).*)

(3) Specialize to the case $\rho = |\psi\rangle\langle\psi|$ being the pure state associated to a smooth, compactly supported wavefunction $\psi \in L^2(\mathbb{R}, dx)$, to deduce from (4.23) the original formulation of Heisenberg's uncertainty inequality.

Exercise 4.10. Recall that a state η on the Weyl algebra is called a quantum (bosonic) *Gaussian* state if its characteristic function is a quadratic polynomial of the variables, i.e.,

$$f(r,s) = \exp\left(a + b\begin{pmatrix} r \\ s \end{pmatrix} + (r,s) C \begin{pmatrix} r \\ s \end{pmatrix}\right),$$

for some $a \in \mathbb{C}$, $b \in \mathbb{C}^2$ and $C \in \mathbb{C}^{2\times 2}$.

(1) Show that necessarily $a = 0$ and $C \in \mathbb{R}^{2\times 2}$ is real and symmetric.

(2) Using the previous exercise, deduce that

$$C \leq -i\omega,$$

denoting with C the real quadratic form on $\mathbb{R}^2$ associated to C.

5. Distances

In this section, we address the following natural problem: how to compare two states of a quantum system? The answer of course depends on the application one has in mind, e.g. on the restrictions upon the measurements one can perform, and justifies a large variety of "distances" (which actually may fail to satisfy all the axioms of proper distances). This variety should not be a surprise, since a similar picture emerges when comparing classical probability distributions. In fact, the quantities we are going to introduce are the quantum analogues of three widely used distances in probability and statistics:

(1) the total variation distance, which will lead to the *trace distance* in the quantum case,
(2) the Hellinger distance, or Bhattacharyya coefficient, yielding the quantum *fidelity* between two states,
(3) the Kantorovich-Wasserstein (or earth mover's) distance, which actually has more than one quantum analogue.

We leave aside one of the most relevant notion of information-theoretic "distance", the Kullback-Leibler divergence, or relative entropy, to be addressed in Section 6 together with related entropic quantities.

Besides their basic properties (e.g., are they actual distances? are there inequalities between them?) a natural question that we address is their behaviour with respect to composition with quantum channels between systems, by showing that they contract (or expand in a controlled way) with respect to such action. This will also require the application of the inequalities from the previous section.

5.1. **Trace distance.** Given two probability distributions p, q over a (finite) set Ω, their *total variation* distance is simply given by half of the ℓ^1-norm of their difference:

$$\|p - q\|_{TV} = \frac{1}{2} \sum_{\omega \in \Omega} |p(\omega) - q(\omega)| \in [0, 1],$$

Clearly, it defines a proper distance on the set of probability distributions. It is a simple exercise to check that, given a Markov kernel $N = (N(\omega, x))_{\omega \in \Omega, x \in \mathcal{X}}$ from Ω to a finite set $\mathcal{X}$, its action on probability distributions provides a non-expanding map, with respect to the total variation:

$$\|(N^\dagger p) - (N^\dagger q)\|_{TV} \leq \|p - q\|_{TV}. \tag{5.1}$$

In the particular case of $N(\omega, x) = \delta_{X(\omega)}(x)$ where X is a random variable with values in X, since $(N^\dagger)p(x) = \mathbb{P}_p(X = x)$ is the distribution of X, one obtains that

$$\sum_{x \in \mathcal{X}} |\mathbb{P}_p(X = x) - \mathbb{P}_q(X = x)| \leq \|p - q\|_{TV}. \tag{5.2}$$

This can be used e.g. to argue via fixed point arguments that any Markov chain (i.e., when $\Omega = \mathcal{X}$) admits at least one invariant distribution and convergence to equilibrium, under suitable assumptions (i.e., when $N^\dagger$ is a proper contraction).

Another useful point of view of the total variation distance is provided by its dual representation:

$$\|p - q\|_{TV} = \sup_{V \subseteq \Omega} \sum_{\omega \in V} p(\omega) - q(\omega). \tag{5.3}$$

The inequality $\geq$ is obvious, to see the converse use $V = \{\omega \in \Omega : p(\omega) > q(\omega)\}$ and notice that

$$\|p - q\|_{TV} = \sum_{\omega \in V} p(\omega) - q(\omega).$$

We can also relax the right hand side of (5.3), obtaining a further equivalent representation:

$$\|p - q\|_{TV} = \sup_{a:\Omega \to [0,1]} \sum_{\omega \in V} a(\omega) \left(p(\omega) - q(\omega)\right). \tag{5.4}$$

The usefulness and simplicity of the total variation distance motivates the introduction its quantum analogue, called *trace distance*.

On a finite-dimensional quantum system H, given states ρ, $\sigma \in \mathcal{S}(H)$, one defines

$$D_{\mathrm{tr}}(\rho, \sigma) = \frac{1}{2}\mathrm{tr}|\rho - \sigma|$$

where $|\rho - \sigma|$ is defined via spectral calculus on $\rho - \sigma$, which is self-adjoint:

$$|\rho - \sigma| = \sum_{\lambda \in \sigma(\rho - \sigma)} |\lambda| \mathbb{1}_{(\rho-\sigma)=\lambda}.$$

The same expression also yields $D_{\mathrm{tr}}(\rho, \sigma) \in [0, 1]$.

Remark 5.1. Up to the factor $1/2$, the trace distance can be seen as the $p = 1$ case of the general p-Schatten norm defined (for $p \geq 1$) on operators $A \in \mathcal{L}(H)$ as

$$\|A\|_p := \mathrm{tr}[|A|^p]^{1/p} = \mathrm{tr}[(A^*A)^{p/2}]^{1/p}.$$

One writes $\mathcal{L}^p(H)$ for the space $\mathcal{L}(H)$ (we work in finite dimensions) endowed with the Schatten p-norm, which define analogues of Lebesgue spaces. The case $p = \infty$ corresponds to the operator norm

$$\|A\|_\infty = \sup_{|\psi\rangle \in H \setminus \{0\}} \frac{\langle A\psi | A\psi \rangle}{\langle \psi | \psi \rangle}.$$

One can argue that the following duality holds, for every $A \in \mathcal{L}(H)$ and conjugate p, q, i.e. $1/p + 1/q = 1$:

$$\|A\|_p = \sup_{\|B\|_q \leq 1} \mathrm{tr}[B^*A],$$

hence in particular (see Exercise 5.1) that

$$\|A\|_1 = \sup_{\|B\|_\infty \leq 1} \mathrm{tr}[BA]. \tag{5.5}$$

Clearly, the trace distance $D_{\mathrm{tr}}(\rho, \sigma)$ is symmetric $D_{\mathrm{tr}}(\rho, \sigma) = D_{\mathrm{tr}}(\sigma, \rho)$ and $D_{\mathrm{tr}}(\rho, \sigma) = 0$ if and only if $\rho = \sigma$. Moreover, if both ρ and σ are diagonal states with respect to the same orthonormal basis $(|i\rangle)_{i \in I}$, i.e.,

$$\rho = \sum_{i \in I} p_i \left|i\right\rangle \left\langle i\right|, \quad \sigma = \sum_{i \in I} q_i \left|i\right\rangle \left\langle i\right|,$$

for classical probability densities $(p_i)_{i \in I}$, $(q_i)_{i \in I}$, then

$$D_{\mathrm{tr}}(\rho, \sigma) = \|p - q\|_{TV}.$$

Moreover if ρ, $\sigma\mathcal{S}(H)$ are pure states associated respectively to state vectors $|\psi\rangle$, $|\varphi\rangle$, one can prove (see 5.3) that

$$D_{\mathrm{tr}}(\rho, \sigma) = \sqrt{1 - |\langle \psi | \varphi \rangle|^2}. \tag{5.6}$$

In order to prove that the trace distance enjoys the triangle inequality, we establish first the following dual representation, completely analogue to (5.3):

$$D_{\mathrm{tr}}(\rho, \sigma) = \sup_{V < H} \mathrm{tr}[\mathbb{1}_V(\rho - \sigma)] \tag{5.7}$$

The right hand side can be equivalently written using a probabilistic notation:

$$\sup_{V<H} \mathrm{tr}[\mathbb{1}_V(\rho-\sigma)] = \sup_{V<H} (\mathbb{1}_V)_\rho - (\mathbb{1}_V)_\sigma = \sup_{V<H} \mathbb{P}_\rho(V) - \mathbb{P}_\sigma(V).$$

In physical terms, the right hand side can be interpreted as the largest "discrepancy" between the probabilities of observing that V holds, when V can be any subspace of H (which recall that we interpret as a logical proposition about the system H).

To prove (5.7), notice first that

$$0 \leq (\rho-\sigma)^+ = |\rho-\sigma| - (\rho-\sigma)$$

(by spectral calculus applied to $\rho-\sigma$ and $f(x) = x^+ = |x| - x$), so that

$$(\mathbb{1}_V)_\rho - (\mathbb{1}_V)_\sigma = \mathrm{tr}[\mathbb{1}_V(\rho-\sigma)] \leq \mathrm{tr}[\mathbb{1}_V|\rho-\sigma|] \leq \mathrm{tr}[|\rho-\sigma|] = D_{\mathrm{tr}}(\rho,\sigma), \tag{5.8}$$

where in the last inequality we used that $\mathbb{1}_V \leq \mathbb{1}_H$. This yields inequality $\geq$ in (5.7). To prove equality, we choose $V = \{\rho-\sigma > 0\}$ i.e., the subspace spanned by all the eigenspaces corresponding to positive eigenvalues of $\rho-\sigma$, we obtain the thesis. Indeed, again by spectral calculus,

$$\mathbb{1}_V = \sum_{\lambda\in\sigma(\rho-\sigma),\lambda>0} \mathbb{1}_{\{\rho-\sigma=\lambda\}},$$

hence

$$(\rho-\sigma)\mathbb{1}_V = \sum_{\lambda\in\sigma(\rho-\sigma),\lambda>0} \lambda\mathbb{1}_{\{\rho-\sigma=\lambda\}} = (\rho-\sigma)^+.$$

On the other hand, since $\mathrm{tr}[\rho-\sigma] = 0$, we have that

$$\mathrm{tr}[(\rho-\sigma)^+] = \mathrm{tr}[(\rho-\sigma)^-],$$

hence

$$\frac{1}{2}\mathrm{tr}[|\rho-\sigma|] = \frac{1}{2}\left(\mathrm{tr}[(\rho-\sigma)^+] + \mathrm{tr}[(\rho-\sigma)^-]\right) = \mathrm{tr}[(\rho-\sigma)^+] = \mathrm{tr}[\mathbb{1}_V(\rho-\sigma)].$$

Using (5.7), it is straightforward to check the triangle inequality. Indeed, for ρ, σ, $\tau \in \mathcal{S}(H)$, and $V < H$,

$$\mathbb{P}_\rho(V) - \mathbb{P}_\tau(V) = (\mathbb{P}_\rho(V) - \mathbb{P}_\sigma(V)) + (\mathbb{P}_\sigma(V) - \mathbb{P}_\tau(V)) \leq D_{\mathrm{tr}}(\rho,\sigma) + D_{\mathrm{tr}}(\sigma,\tau),$$

hence

$$D_{\mathrm{tr}}(\rho,\tau) \leq D_{\mathrm{tr}}(\rho,\sigma) + D_{\mathrm{tr}}(\sigma,\tau).$$

The argument leading to inequality $\geq$ in (5.7) holds if we replace $\mathbb{1}_V$ with any observable $A \in \mathcal{O}(H)$ so that $0 \leq A \leq \mathbb{1}_H$ (i.e., such that $\sigma(A) \subseteq [0,1]$), which includes of course the case $A = \mathbb{1}_V$. Therefore, we obtain the following analogue of (5.4):

$$D_{\mathrm{tr}}(\rho,\sigma) = \sup_{A\in\mathcal{O}(H),\sigma(A)\subseteq[0,1]} (A)_\rho - (A)_\sigma. \tag{5.9}$$

Using such characterization, we deduce the quantum analogue of (5.1). Precisely, we prove that given any quantum channel $\Phi^\dagger : \mathcal{L}(H) \to \mathcal{L}(K)$ from a quantum system H to a quantum system K, the trace distance decreases:

$$D_{\mathrm{tr}}(\Phi^\dagger(\rho),\Phi^\dagger(\sigma)) \leq D_{\mathrm{tr}}(\rho,\sigma) \quad \text{for every } \rho,\, \sigma \in \mathcal{S}(H).\ . \tag{5.10}$$

To see this, let $A \in \mathcal{O}(K)$ be such that $\sigma(A) \subseteq [0,1]$, i.e., $0 \leq A \leq \mathbb{1}_K$. Then, the dual map $\Phi : \mathcal{L}(K) \to \mathcal{L}(H)$, applied to A yields $0 \leq \Phi(A) \leq \Phi(\mathbb{1}_K) = \mathbb{1}_H$, being positive (it is actually CP) and unital. By definition of dual map,

$$(\Phi(A))_\rho = \mathrm{tr}[\Phi(A)\rho] = \mathrm{tr}[A\Phi^\dagger(\rho)] = (A)_{\Phi^\dagger(\rho)}.$$

Thus,

$$(A)_{\Phi^\dagger(\rho)} - (A)_{\Phi^\dagger(\sigma)} = (\Phi(A))_\rho - (\Phi(A))_\sigma \leq D_{\mathrm{tr}}(\rho,\sigma).$$

Being $A \in \mathcal{O}(K)$ with $\sigma(A) \subseteq [0,1]$ arbitrary, we deduce (5.10).

Example 5.2 (The partial trace decreases the trace distance)**.** Let ρ, $\sigma \in \mathcal{S}(H \otimes K)$ and let $\rho^H = \mathrm{tr}_K[\rho]$, $\sigma^H = \mathrm{tr}_K[\sigma] \in \mathcal{S}(H)$. Then

$$D_{\mathrm{tr}}(\rho^H, \sigma^H) \leq D_{\mathrm{tr}}(\rho, \sigma).$$

To obtain the analogue of (5.2) in the case of a quantum measurement $X = (\mathbb{1}_{X=x})_{x\in\mathcal{X}}$, i.e.,

$$\|\mathbb{P}_\rho(X = \cdot) - \mathbb{P}_\sigma(X = \cdot)\|_{TV} \leq D_{\mathrm{tr}}(\rho, \sigma), \tag{5.11}$$

we argue as follows. Consider the quantum channel

$$\rho \mapsto \Phi^\dagger(\rho) = \sum_{x\in\mathcal{X}} \mathbb{1}_{\{X=x\}} \rho \mathbb{1}_{\{X=x\}},$$

and given any $a : \mathcal{X} \to [0, 1]$, consider the observable

$$A = \sum_{x\in\mathcal{X}} a(x)\mathbb{1}_{\{X=x\}},$$

with $\sigma(A) \in [0, 1]$, so that

$$(A)_{\Phi^\dagger(\rho)} = \sum_{x\in\mathcal{X}} a(x)\mathbb{P}_\rho(X = x),$$

hence

$$\sum_{x\in\mathcal{X}} a(x)\left(\mathbb{P}_\rho(X = x) - \mathbb{P}_\sigma(X = x)\right) = (A)_{\Phi^\dagger(\rho)} - (A)_{\Phi^\dagger(\sigma)} = (\Phi(A))_\rho - (\Phi(A))_\sigma$$
$$\leq D_{\mathrm{tr}}(\rho, \sigma).$$

Since $a : \mathcal{X} \to [0, 1]$ is arbitrary, by (5.4) we conclude that (5.11) holds.

5.2. **Fidelity.** The total variation distance has the slight drawback[11] that it is modelled upon the ℓ^1 norm which is not strictly convex, nor smooth. One would prefer a smoother notion of distance, e.g. modelled after an ℓ^2 norm. A possible partial solution is to introduce the (squared) Hellinger distance between two probability distributions $(p(\omega))_{\omega\in\Omega}$, $(q(\omega))_{\omega\in\Omega}$, defined as

$$H^2(p, q) = \frac{1}{2}\sum_{\omega\in\Omega} |\sqrt{p(\omega)} - \sqrt{q(\omega)}|^2. \tag{5.12}$$

Up to a factor $1/2$, this amounts to the squared ℓ^2-norm between the square roots of the two distributions (hence it is easily seen to be a distance). By developing the square, we see that

$$H^2(p, q) = \frac{1}{2}\sum_{\omega\in\Omega} |\sqrt{p(\omega)} - \sqrt{q(\omega)}|^2 = 1 - \sum_{\omega\in\Omega} \sqrt{p(\omega)q(\omega)} \leq 1.$$

This identity often motivates the introduction of the *Bhattacharyya coefficient* (BC)

$$BC(p, q) = \sum_{\omega\in\Omega} \sqrt{p(\omega)q(\omega)} \in [0, 1],$$

so that

$$H(p, q) = 1 - BC(p, q),$$

hence the larger $BC(p, q)$ is, the closer are the two distributions. We can easily compare the Hellinger distance with the total variation distance:

$$H^2(p, q) \leq \|p - q\|_{TV} \leq \sqrt{2H(p, q)}. \tag{5.13}$$

It turns out that H satisfies a property similar to (5.1), i.e.,

$$H(N^\dagger p, N^\dagger q) \leq H(p, q).$$

[11] to be fair, sometimes this may actually be an advantage, e.g. in LASSO methods in statistics

for any Markov kernel $(N(\omega,x)_{\omega\in\Omega,x\in\mathcal{X}}$ from Ω to $\mathcal{X}$. Again, this entails that for every random variable X with values in $\mathcal{X}$,

$$H\left(\mathbb{P}_p(X=\cdot),\mathbb{P}_q(X=\cdot)\right)\leq H\left(p,q\right).$$

These can be rephrased in terms of the BC quantity, simply by reversing the inequalities:

$$BC(N^\dagger p,N^\dagger q)\geq BC(p,q),\quad BC\left(\mathbb{P}_p(X=\cdot),\mathbb{P}_q(X=\cdot)\right)\geq BC\left(p,q\right). \tag{5.14}$$

It turns out that in the quantum setting the analogues of Hellinger distance and Bhattacharyya coefficient have a natural interpretation, in particular for pure states $\rho=|\psi\rangle\langle\psi|$, because the square root can be somehow replaced with the state vector. It is infact more common to deal with the analogue of the (squared) Bhattacharyya coefficient, which is called the *fidelity* between quantum states, ρ, $\sigma\in\mathcal{S}(H)$, defined as

$$F(\rho,\sigma)=\left(\mathrm{tr}[\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right)^2.$$

If ρ and σ commute, then clearly

$$\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}=\sqrt{\rho}\sqrt{\sigma},$$

thus, when ρ, σ are diagonal operators with respect to a common basis $(|i\rangle)_{i\in I}$, i.e.,

$$\rho=\sum_{i\in I}p_i\,|i\rangle\langle i|\,,\quad \sigma=\sum_{i\in I}q_i\,|i\rangle\langle i|\,,$$

for classical probability densities $(p_i)_{i\in I}$, $(q_i)_{i\in I}$, then

$$F(\rho,\sigma)=\sum_{i\in I}\sqrt{p_i}\sqrt{q_i}=BC^2(p,q).$$

The analogue of the Hellinger distance is commonly defined as

$$D_B(\rho,\sigma)^2=2\left(1-\sqrt{F(\rho,\sigma)}\right),$$

and called in this context the *Bures* metric (it is in fact a proper distance, see [Hol19, section 10.2.3]).

When both states $\rho=|\psi\rangle\langle\psi|$, $\sigma=|\varphi\rangle\langle\varphi|$ one has (since $\sqrt{\rho}=\rho$)

$$\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}=\sqrt{|\psi\rangle\,|\langle\psi|\varphi\rangle|^2\,|\psi\rangle}=\rho\,|\langle\psi|\varphi\rangle|^2\,,$$

hence

$$F(\rho,\sigma)=|\langle\psi|\varphi\rangle|^2$$

coincides with a squared amplitude, that we interpreted in section 2 as the probability that one observes that the system is in state σ, given that it is infact in the state ρ (or viceversa).

There are several equivalent ways to rewrite the fidelity:

$$F(\rho,\sigma)=\mathrm{tr}[\left|\sqrt{\rho}\sqrt{\sigma}\right|]^2=\frac{1}{2}\left(\mathrm{tr}[(\rho\sharp_{1/2}\sigma^{-1})\sigma]+\mathrm{tr}[(\sigma\sharp_{1/2}\rho^{-1})\rho]\right),$$

which highlight that it is in general symmetric, $F(\rho,\sigma)=F(\sigma,\rho)$. Clearly, $F(\rho,\rho)=1$, and one can prove that $F(\rho,\sigma)=1$ if and only if $\rho=\sigma$.

Like the trace distance, also the fidelity is well-behaved with respect to the action of quantum channels. One can prove indeed the analogues of (5.14), which read in this case as

$$F(\Phi^\dagger(\rho),\Phi^\dagger(\sigma))\geq F(\rho,\sigma), \tag{5.15}$$

for any channel $\Phi^\dagger$ from between two quantum systems, and

$$BC^2(\mathbb{P}_\rho(X=\cdot),\mathbb{P}_\sigma(X=\cdot))\geq F(\rho,\sigma),$$

for every measurement $X = (V_x)_{x\in\mathcal{X}}$ taking values in $\mathcal{X}$. In fact, one can argue that

$$F(\rho,\sigma) = \inf\left\{BC^2(\mathbb{P}_\rho(X=\cdot),\mathbb{P}_\sigma(X=\cdot)) \,:\, X=(V_i)_{i=1,\dots,d} \text{ measurement}\right\}, \tag{5.16}$$

where $d = \dim(H)$.

The key argument relies on the following variational representation of the fidelity:

$$F(\rho,\sigma) = \sup\left\{|\mathrm{tr}[X]| \,:\, X\in\mathcal{L}(H), \text{ such that } M = \begin{pmatrix} \rho & X \\ X^* & \sigma \end{pmatrix} \geq 0\right\}. \tag{5.17}$$

To see this identity, the simple consequence of Lemma 4.1 noticed in Exercise 4.4 and the duality 5.5: the condition $M \geq 0$ coincides with the representation $X = \sqrt{\rho}Z\sqrt{\sigma}$ for some Z with operator norm $\|Z\| \leq 1$, so that

$$\sup_X |\mathrm{tr}[X]| = \sup_Z \left|\mathrm{tr}[\sqrt{\rho}Z\sqrt{\sigma}]\right| = \sup_Z \left|\mathrm{tr}[\sqrt{\sigma}\sqrt{\rho}Z]\right| = \mathrm{tr}[\left|\sqrt{\sigma}\sqrt{\rho}\right|] = F(\rho,\sigma).$$

Given a channel $\Phi^\dagger$, and $X\in\mathcal{L}(H)$ such that

$$M = \begin{pmatrix} \rho & X \\ X^* & \sigma \end{pmatrix} \geq 0,$$

by complete positivity, we have

$$\begin{pmatrix} \Phi^\dagger(\rho) & \Phi^\dagger(X) \\ \Phi^\dagger(X^*) & \Phi^\dagger(\sigma) \end{pmatrix} \geq 0,$$

and, since $\Phi^\dagger$ is trace preserving,

$$\mathrm{tr}[\Phi^\dagger(X)] = \mathrm{tr}[X].$$

Thus,

$$F(\Phi^\dagger(\rho),\Phi^\dagger(\sigma)) \geq |\mathrm{tr}[\Phi^\dagger(X)]| = |\mathrm{tr}[X]|.$$

Maximizing upon X yields (5.15).

Finally, fidelity (or the Bures metric) and related to the trace distance via the following analogues of (5.13), called *Fuchs–van de Graaf inequalities*

$$1 - \sqrt{F} \leq D_{\mathrm{tr}} \leq \sqrt{1-F}.$$

5.3. **Quantum optimal transport.** The classical optimal transport problem, originated by the works by G. Monge and L. Kantorovich (see [Vil09]) searches for the most efficient way to move two probability distributions $(p(x))_{x\in\mathcal{X}}$, $(q(x))_{x\in\mathcal{X}}$, with respect to an average displacement cost $(c(x,y))_{x,y\in\mathcal{X}}$ (often, given in terms of function of a distance d defined $\mathcal{X}$). The usefulness of such a problem is that it measures the discrepancy between the two distributions taking into account the "geometry" induced by c, yielding more flexibility.

The precise definition is given by the following variational problem:

$$W^c(p,q) := \inf_{\pi\in\mathcal{C}(p,q)} \sum_{x,y\in\mathcal{X}} c(x,y)\pi(x,y) \tag{5.18}$$

where $\mathcal{C}(p,q)$ denotes the set of so-called *couplings* between the two probabilities p, q, given by joint probability distributions $\pi = (\pi(x,y))_{x,y\in\mathcal{X}}$ such that their marginals are p, q, i.e.,

$$\forall x,y\in\mathcal{X}, \quad 0\leq\pi(x,y)\leq 1, \sum_{x\in\mathcal{X}}\pi(x,y) = q(y), \sum_{y\in\mathcal{X}}\pi(x,y) = p(x). \tag{5.19}$$

We can also give a different, more dynamical, description of a coupling, in terms given by the conditional probabilities

$$N(x,y) = \pi(y|x) = \frac{\pi(x,y)}{p(x)}, \tag{5.20}$$

(of course provided that $p(x) > 0$ for every $x\in\mathcal{X}$). Then, N defines a Markov kernel from $\mathcal{X}$ into itself such that, with the usual notation $N^\dagger p = q$. We can thus interpret N

(or better $N^\dagger$) as a generalized function, called a *transport plan*, pushing the probability distribution p into q.

A further point of view, is provided by the *dual formulation* of the problem (5.18). Indeed, both the cost function and the constraints are linear with respect to π, hence one can invoke duality from linear programming theory and write

$$W^c(p,q) = \sup\left\{ \sum_{x\in\mathcal{X}} f(x)p(x) + \sum_{y\in\mathcal{X}} g(y)q(y) \;:\; f(x)+g(y) \le c(x,y)\ \forall x,y\in\mathcal{X} \right\}, \tag{5.21}$$

Any pair of functions (f,g) that attains the supremum is called a pair of *Kantorovich potentials* for the problem. Notice that inequality $\geq$ follows straightforwardly. In the special case of $c(x,y) = d(x,y)$ being a distance function, it is simple to prove that one can actually restrict the dual problem to potentials that are 1-Lipschitz functions, i.e., of the form $(f,-f)$ for some $f:\mathcal{X}\to\mathbb{R}$ such that

$$|f(x)-f(y)| \le d(x,y) \quad \text{for every } x,y\in\mathcal{X}.$$

The problem (5.21) reduces to the so-called Kantorovich problem

$$W^d(p,q) = \sup_{f \text{ is 1-Lip}} \left\{ \sum_{x\in\mathcal{X}} f(x)\,(p(x)-q(x)) \right\}, \tag{5.22}$$

and defines the so-called *Wasserstein distance of order* 1 associated to the distance d.

Besides providing a fundamental problem in operation research and combinatorial optimization (it can be seen indeed a linear programming relaxation of the so-called optimal matching problem on weighted graphs, [PL86]), in the last three decades optimal transport theory evolved providing a variety of novel mathematical tools with applications in many fields, from the analysis of PDE's and Riemannian geometry [Vil09], to statistics and machine learning, see [PC+19].

In quantum (or more general non-commutative) settings, the first proposals for a optimal transport date back to the 1990's [CL92; KW98], but in recent years new formulations have been investigated [Agr13; CM14; GMP16; Gos21; De +21], motivated by the fact that classical optimal transport admits several equivalent formulations (and we already described some in the elementary setting above). We briefly describe three such proposals in the setting of elementary quantum systems.

5.3.1. *Optimal Transport via quantum couplings.* Given a quantum system H, it is quite natural, in order define an analogue of (5.18), where p and q are replaced by density operators $\rho,\sigma\in\mathcal{S}(H)$, to introduce a cost operator $C\in\mathcal{O}(H_1\otimes H_2)$, where $H_1 = H_2 = H$ (we label them just to differentiate) and a notion of quantum couplings $\mathcal{C}(\rho,\sigma)$ given by density operators $\Pi\in\mathcal{S}(H_1\otimes H_2)$ such that the corresponding reduced density operators are

$$\rho = \mathrm{tr}_{H_2}[\Pi], \quad \sigma = \mathrm{tr}_{H_1}[\Pi].$$

One defines the optimal transport cost as

$$W_C(\rho,\sigma) = \inf_{\Pi\in\mathcal{C}(p,q)} \mathrm{tr}[C\Pi].$$

(where C conveniently stands both for the operator C but also for "coupling")

As a special case of cost operator one can introduce a "squared-distance" like observable, given by

$$C = \sum_{i\in I} (A_i\otimes \mathbb{1}_{H_2} - \mathbb{1}_{H_1}\otimes A_i)^2, \tag{5.23}$$

for a family of $(A_i)_{i\in I}\subseteq\mathcal{O}(H)$. With such a choice, clearly $W^C(\rho,\sigma)\geq 0$ and $W_C(\rho,\sigma) = W_C(\sigma,\rho)$, however further properties depend upon the choice of the family $(A_i)_{i\in I}$. An one might expect, if they are all compatible, then the situation becomes closer to the

classical optimal transport problem, but incompatible observables may lead to interesting phenomena, some of them are widely open questions: for example, under which conditions $\sqrt{W_C(\rho,\sigma)}$ is an actual distance? can one characterize, or at least find simple conditions to estimate how much a quantum channel $\Phi^\dagger$ expands the cost, i.e., provide a constant $\|\Phi^\dagger\|_{W_C}$ such that

$$W_C(\Phi^\dagger(\rho),\Phi^\dagger(\sigma)) \leq \|\Phi^\dagger\|_{W_C} W_C(\rho,\sigma), \quad \text{for every } \rho,\sigma \in \mathcal{S}(H)?$$

This notion of quantum optimal transport is fruitfully used in [GMP16] and subsequent works in the infinite-dimensional (CCR) setting, letting A_i being position or momentum operators, to investigate quantitatively semiclassical limits, i.e., when the Planck constant $\hbar \to 0$. Other cost operators, have also been proposed and investigated, e.g. projection operators [BEŻ22].

5.3.2. *Optimal Transport via quantum channels.* A different take on the problem is provided by the identification (5.20) of a coupling $\pi \in \mathcal{C}(p,q)$ with a plan N such that $N^\dagger p = q$. In [DT21], it is proposed to replace $N^\dagger$ with a quantum channel $\Phi^\dagger$ from H into itself such that $\Phi^\dagger(\rho) = \sigma$. Notice that the family of such *quantum plans* defines a nice closed and convex set in $\mathcal{L}(\mathcal{L}(H))$. However, it is less obvious now how to define the cost functional.

For simplicity, let us restrict to the quadratic case, i.e. the cost is a "sum of squared differences" for a family $(A_i)_{i\in I} \subseteq \mathcal{O}(H)$ (but one could also consider general costs as in the previous section). Back to the classical case where A_i corresponds to a function f_i, this amounts to

$$c(x,y) = \sum_{i\in I} (f_i(x) - f_i(y))^2 = \sum_{i\in I} f_i^2(x) - f_j^2(y) - 2f_i(x)f_j(y),$$

so that, using (5.19),

$$\begin{aligned}\sum_{x,y\in\mathcal{X}} c(x,y)\pi(x,y) &= \sum_{i\in I}\sum_{x,y\in\mathcal{X}} (f_i(x) - f_i(y))^2\pi(x,y)\\ &= \sum_{i\in I}\sum_{x\in\mathcal{X}} f_i^2(x)p(x) - \sum_{y\in\mathcal{X}} f_i^2(y) - \sum_{x,y\in\mathcal{X}} 2f_i(x)f_i(y)\pi(x,y).\end{aligned}$$

The first two terms are simply mean values of f_i^2 with respect to p and q, hence their quantum analogues are easily written. The problem comes with the third term, which we can rewrite however using only p and the kernel N:

$$\sum_{x,y} f_i(x)f_i(y)\pi(x,y) = \sum_{x\in\mathcal{X}} f_i(x)p(x)(Nf_i)(x).$$

In the quantum case, we choose to rewrite the quantity as follows:

$$\mathrm{tr}[A_i\sqrt{\rho}\Phi(A_i)\sqrt{\rho}] = \mathrm{tr}[(\sqrt{\rho}A_i)^*\Phi(A_i)\sqrt{\rho}] = \langle\sqrt{\rho}A_i|\Phi(A_i)\sqrt{\rho}\rangle\,, \tag{5.24}$$

where we used the Hilbert-Schmidt scalar product in $\mathcal{L}(H)$. We are thus lead to the following expression for the cost in terms of Φ, ρ and σ only:

$$\mathrm{Cost}(\Phi,\rho,\sigma) = \sum_{i\in I} \mathrm{tr}[A_i^2\rho] + \mathrm{tr}[A_i^2\sigma] - 2\mathrm{tr}[A_i\sqrt{\rho}\Phi(A_i)\sqrt{\rho}].$$

It turns out that minimization of the above quantity with respect to quantum plans $\Phi^\dagger$ from ρ to σ, yields an optimal transport cost

$$W_P(\rho,\sigma) = \inf_{\Phi^\dagger(\rho)=\sigma} \mathrm{Cost}(\Phi,\rho,\sigma),$$

which in general is different from the one defined in the previous section (we also use P to remember it is defined via "plans"). Although it may be not evident from the definition, it holds

$$W_P(\rho,\sigma) = W_P(\sigma,\rho).$$

This quantity shares many similarities with W_C in terms of general properties, with two notable differences: one always give a lower bound

$$W_P(\rho,\sigma) \geq \frac{1}{2}\left(W_P(\rho,\rho) + W_P(\sigma,\sigma)\right), \quad \text{for every } \rho,\, \sigma \in \mathcal{S}(H), \tag{5.25}$$

and show a *modified* triangle inequality:

$$\sqrt{W_P(\rho,\tau)} \leq \sqrt{W_P(\rho,\sigma)} + \sqrt{W_P(\sigma,\sigma)} + \sqrt{W_P(\sigma,\tau)}, \quad \text{for every } \rho,\sigma,\tau \in \mathcal{S}(H). \tag{5.26}$$

Both proofs are not difficult relying upon Lieb's theorem Theorem 4.3, see 5.7. Motivated by (5.25), one can then introduce the quantity

$$(\rho,\sigma) \mapsto \sqrt{W_P(\rho,\sigma) - \left(\frac{1}{2}\left(W_P(\rho,\rho) + W_P(\sigma,\sigma)\right)\right)}.$$

It is an open problem to determine whether this defines an actual distance (possibly allowing it to be degenerate, in the sense that it may be null even if $\rho \neq \sigma$).

5.3.3. *Optimal transport via Lipschitz operators.* This third way instead begins from the observation that one can use the dual problem (5.22) to *define* $W^d(p,q)$, provided that one specifies a suitable notion of 1-Lipschitz function. Such observation in the non-commutative case actually dates back at least to [CL92]. It was recently noticed [De +21] that it leads to a particularly simple yet useful theory in the setting of product systems $H = \bigotimes_{i\in I} H_i$, providing a quantum analogue of the optimal transport problem with respect to the Hamming distance on product sets $\Pi_{i\in I}\mathcal{X}_i$, defined as

$$d_{\mathrm{Ham}}((x_i)_{i\in I},(y_i)_{i\in I}) = \sum_{i\in I} 1_{\{x_i\neq y_i\}}.$$

In simple terms, the Hamming distance between two sequences counts the number of positions in which they differ. Since our aim is to argue by duality, we notice first that a function $f : \Pi_{i\in I}\mathcal{X}_i \to \mathbb{R}$ is 1-Lipschitz with respect to the Hamming distance if and only if, for every $i \in I$, one has

$$|f(x) - f(y)| \leq 1$$

whenever the sequences x, y differ only at the position i (let us write $x \sim_i y$) . For every $i \in I$, we thus introduce the oscillation at position $i \in I$,

$$\partial_i f = \sup_{x\sim_i y} |f(x) - f(y)| \tag{5.27}$$

Hence, the Lipschitz constant of f is

$$\|f\|_{\mathrm{Lip}} = \max_{i\in I} \partial_i f,$$

and with this notion we can define the optimal transport problem with cost given by d_{Ham} by (5.22).

If we forget about the product structure $\mathcal{X} = \Pi_{i\in I}\mathcal{X}_i$ and think of the whole $\mathcal{X}$ as a single factor, the Hamming distance reduces to the trivial distance

$$d_{\mathrm{Tri}}(x,y) = 1_{\{x\neq y\}}.$$

A 1-Lipschitz function f with respect to the trivial distance is, up to an additive constant, any function $a : \mathcal{X} \to [0,1]$. We find therefore that (5.22) reduces to (5.4), hence the total variation distance coincides with the optimal transport cost with respect to the trivial distance. Moreover, the following inequalities hold:

$$d_{\mathrm{Tri}}(x,y) \leq d_{\mathrm{Ham}}(x,y) \leq |I| d_{\mathrm{Tri}}(x,y),$$

which easily imply

$$\|p-q\|_{TV} \leq W^{d_{\mathrm{Ham}}}(p,q) \leq |I|\|p-q\|_{TV}. \tag{5.28}$$

The analogue of the quantum setting of the above construction on a composite system $H = \bigotimes_{i\in I} H_i$, where f are replaced with observables $A \in \mathcal{O}(H)$, stems from the observation that, in the classical case, one has the equivalent expression

$$\partial_i f = 2 \inf_{g_i} \sup_x |f(x) - g_i(x)|, \tag{5.29}$$

where the infimum runs among all the functions $g_i : x \mapsto g_i(x)$ which do not depend upon the coordinate x_i. This allows to formulate a quantum analogue of the oscillation (which we call *dependence*) at position i, :

$$\partial_i A = \inf \left\{ 2 \|A - G_i \otimes \mathbb{1}_{H_i}\|_\infty \,:\, G_i \in \mathcal{O}(\bigotimes_{j\neq i} H_j) \right\},$$

where the structure $G_i \otimes I_{H_i}$ encodes the fact that the observable "is not a function" of position i. The *quantum Lipschitz constant* of $A \in \mathcal{O}(H)$ is defined as

$$\|A\|_L := \max_{i\in I} \partial_i A,$$

and the quantity

$$\begin{aligned} \|\rho-\sigma\|_{W_1} &= \sup\{\mathrm{tr}[A(\rho-\sigma)] \,:\, \|A\|_L \leq 1\} \\ &= \sup\{(A)_\rho - (A)_\sigma \,:\, \|A\|_L \leq 1\} \end{aligned}$$

is called *quantum Wasserstein distance of order* 1 between the states $\rho, \sigma \in \mathcal{S}(H)$.

If one forgets about the product structure of $H = \bigotimes_{i\in I} H_i$ and considers it as a single system, the notion of Lipschitz observable trivializes, since one can only add subtract a multiple of the identity operator $\mathbb{1}_H$ on H. Thus, any $A \in \mathcal{O}(H)$ with Lipschitz constant ≤ 1 is, up to adding a multiple of $\mathbb{1}_H$, an observable such that $\sigma(A) \subseteq [0,1]$. It follows from (5.9) that in such a case the *quantum Wasserstein distance of order* 1 reduces to the trace distance. More generally, one can prove the following analogue of (5.28):

$$D_{\mathrm{tr}}(\rho,\sigma) \leq \|\rho-\sigma\|_{W_1} \leq |I| D_{\mathrm{tr}}(\rho,\sigma).$$

If the states are product states $\rho = \otimes_{i\in I}\rho_i$, $\sigma = \otimes_{i\in I}\sigma_i$ (with respect to the decomposition $H = \bigotimes_{i\in I} H_i$), then

$$\|\rho-\sigma\|_{W_1} = \sum_{i\in I} D_{\mathrm{tr}}(\rho_i,\sigma_i). \tag{5.30}$$

As a quantum analogue of the Hamming distance between two state vectors $|\psi\rangle$, $|\varphi\rangle \in H$, although one can use the as a replacement *quantum Wasserstein distance of order* 1 between the associated pure states:

$$d_{\mathrm{Ham}}(|\psi\rangle, |\varphi\rangle) = \| |\psi\rangle\langle\psi| - |\varphi\rangle\langle\varphi| \|_{W_1}.$$

In particular, if $|\psi\rangle = \otimes_{i\in I} |\psi_i\rangle$, $|\varphi\rangle = \otimes_{i\in I} |\varphi_i\rangle$, since by (5.6)

$$D_{\mathrm{tr}}(|\psi_i\rangle\langle\psi_i|, |\varphi_i\rangle\langle\varphi_i|) = \sqrt{1 - |\langle\psi_i|\varphi_i\rangle|^2}$$

we have

$$d_{\mathrm{Ham}}(\otimes_{i\in I}|\psi_i\rangle, \otimes_{i\in I}|\varphi_i\rangle) = \sum_{i\in I} \sqrt{1 - |\langle\psi_i|\varphi_i\rangle|^2}.$$

5.4. **Exercises.**

Exercise 5.1. Show that (5.5) holds. *(Hint: write the trace in terms of a basis of eigenvectors of A, and use Cauchy-Schwarz inequality.)*

Exercise 5.2. Show that (5.16) holds.

Exercise 5.3. On a single qubit system $H = \mathbb{C}^2$, recall the parametrization of states (2.7) in terms of vectors $b = (b_x, b_y, b_z) \in \mathbb{R}^3$ with $b_x^2 + b_y^2 + b_z^2 \leq 1$.

(1) Find an explicit expression of the trace distance $D_{\mathrm{tr}}(\rho, \sigma)$ in terms of the associated vectors b_ρ, b_σ.
(2) Deduce from the previous point the validity of (5.6) *(Hint: restrict to the system spanned by $|\psi\rangle, |\varphi\rangle \in H$.)*
(3) Find an explicit expression of the for the fidelity $F(\rho, \sigma)$ in terms of the associated vectors b_ρ, b_σ.

Exercise 5.4. Show through a explicit examples that it is false in general that *any* channel Φ from a system H to a system K does not expand W_C, i.e.,

$$W_C(\Phi^\dagger(\rho), \Phi^\dagger(\sigma)) \leq W_C(\rho, \sigma),$$

and similarly for the distance W_P.

Exercise 5.5. Prove the validity of (5.25) using this argument:

(1) By writing explicitly the costs, reduce (5.25) to the validity of the inequality

$$\mathrm{tr}[A\sqrt{\rho}\Phi(A)\sqrt{\rho}] \leq \frac{1}{2}\mathrm{tr}[A\sqrt{\rho}\Phi(A)\sqrt{\rho}] + \frac{1}{2}\mathrm{tr}[A\sqrt{\sigma}A\sqrt{\sigma}],$$

for any quantum plan $\Phi^\dagger(\rho) = \sigma$.
(2) To obtain it, apply first Cauchy-Schwarz inequality and then Theorem 4.3.
(3) Conclude that the identity channel $\Phi^\dagger(\tau) = \tau$, $\forall \tau \in \mathcal{S}(H)$ is always optimal when $\sigma = \rho$.

Exercise 5.6. Prove the validity of (5.26).

Exercise 5.7. Consider two composite quantum system $H = \bigotimes_{i\in I} H_i$, $K = \bigotimes_{i\in I} K_i$ and a family $(\Phi_i^\dagger)_{i\in I}$ of quantum channels such that $\Phi_i : \mathcal{L}(H_i) \to \mathcal{L}(K_i)$ and set $\Phi^\dagger := \bigotimes_{i\in I} \Phi_i^\dagger$. Prove that, for any pair of states $\rho, \sigma \in \mathcal{S}(H)$, it holds

$$\|\Phi^\dagger(\rho) - \Phi^\dagger(\sigma)\|_{W_1} \leq \|\rho - \sigma\|_{W_1}.$$

Exercise 5.8. Compute the Wasserstein distance of order 1 between any two Bell states on the composite system $H = \mathbb{C}^2 \otimes \mathbb{C}^2$, e.g. $\|\rho - \sigma\|_{W_1}$ where

$$\rho = \frac{1}{2}\left(|00\rangle + |11\rangle\right)\left(\langle 00| + \langle 11|\right),$$

$$\sigma = \frac{1}{2}\left(|01\rangle + |10\rangle\right)\left(\langle 01| + \langle 10|\right).$$

6. Entropy

In this section we introduce entropy, a central quantity in information theory, since the seminal work by Shannon [Sha48]. Interestingly, its quantum counterpart was introduced by Von Neumann some years before Shannon, but this should not come as a surprise, since entropy in physics had been previously considered, first in thermodynamics (Clausius) and then in statistical mechanics (by Boltzmann and Planck). The main difference between the "information-theoretic" entropy and the "thermodynamical" one is that the former is interpreted as a way to quantify the information contained in a random variable (considered as a signal), while the latter is a measure of disorder. This is not a substantial difference, as in both cases entropy is essentially useful way of counting the possible configurations of a system (by weighting them with the information available to an observer).

6.1. **Classical entropy.** Given a probability distribution p over a (finite) set Ω, we define its *Shannon entropy* as

$$S(p) = -\sum_{\omega\in\Omega} p(\omega)\log p(\omega),$$

where we set by continuity $0\log 0 = 0$. The logarithm is traditionally with respect to basis 2, so that entropy is measured in *bits*. Let us also notice that $S(p) \geq 0$, and that $p \mapsto S(p)$ is concave, since $[0,1] \ni z \mapsto -z\log z \geq 0$ is concave.

Example 6.1. The entropy of a uniform distribution over n values is

$$S((1/n)_{i=1}^n) = -n\cdot\frac{1}{n}\log(1/n) = \log n,$$

so that for the uniform distribution over n-bits sequences $\Omega = \{0,1\}^n$ is n. The larger n, the larger is the entropy.

Example 6.2. The entropy of a probability distribution over two values (i.e., a Bernoulli law) is

$$S((\alpha, 1-\alpha)) = -\alpha\log\alpha - (1-\alpha)\log(1-\alpha) = h_2(\alpha).$$

The function h_2 is also called *binary entropy function*. Notice that $\alpha \mapsto h_2(\alpha)$ is concave, $h_2(\alpha) = h_2(1-\alpha)$, hence it attains its maximum for $\alpha = 1/2$.

Although Shannon's entropy is indeed a function on probability distributions, it is common to refer to the entropy of a random variable $X : \Omega \mapsto \mathcal{X}$ as the entropy of its law:

$$S(X) = S((\mathbb{P}(X=x))_{x\in\mathcal{X}}) = -\sum_{x\in\mathcal{X}} \mathbb{P}(X=x)\log(\mathbb{P}(X=x)).$$

This is slightly ambiguous, since one should also specify the probability distribution p on Ω which is then used to define the law of X. If it is not clear from the context, one should use the more precise notation $S(X)_p$ instead.

In which sense $S(X)$ measures the information content of a random variable X? The example of a uniform distribution is particularly clear. Imagine that X codifies a signal to be transmitted from a source (Alice) to a receiver (Bob), who preliminarily agreed only that *any* binary string of length k will be equally possible. Then, before receiving the message, Bob assumes that X is uniformly distributed over 2^k values, hence its entropy $S(X) = k$, which is indeed the amount of information transmitted by Alice and gained by Bob, after receiving the message (assuming that no noise has distorted the message during the communication). Alternatively, $S(X)$ is a measure of the current state of *ignorance* of the subject (Bob) about X (i.e., before observing it): if $S(X)$ is large, Bob is very unsure about X, hence it will gain a lot of information after observing X. With this interpretation in mind, many properties of entropy become reasonable (but of course they still require a mathematical proof). Indeed, it holds[12] for any random variable X with values in a finite set $\mathcal{X}$,

$$0 \leq S(X) \leq \log|\mathcal{X}| \tag{6.1}$$

with $S(X) = 0$ if and only if X is constant (minimal ignorance) $S(X) = \log|\mathcal{X}|$ if and only if the law of X is uniform on the whole set (maximal ignorance).

Back to the example, if Bob happens to receive some information about X through observation of another random variable Y (possibly correlated with X), how should he update the entropy of X, i.e., its state of ignorance about X? Clearly, after Bob observes $Y = y$, then the answer is simply given by updating the law of X to the conditional distribution

$$x \mapsto \mathbb{P}(X=x|Y=y) = \frac{\mathbb{P}(X=x, Y=y)}{\mathbb{P}(Y=y)},$$

[12] most of the properties in this subsection are stated without proof, since we are going to prove them in the quantum case, which imply as well those for the classical case. The only exception is Lemma 6.3.

which would lead to the quantity

$$S(X)_{\mathbb{P}|Y=y} = -\sum_{x\in\mathcal{X}} \mathbb{P}(X=x|Y=y)\log \mathbb{P}(X=x|Y=y).$$

In information theory, however, the point of view is that of an engineer (as Shannon was after all) who is examining the entire process of sending/receiving information between Alice and Bob, hence it is not interested in the actual value $Y=y$ that Bob may have observed in a specific instance, but rather on the *average* with respect to the possible values he may observe. This motivated Shannon to define the *conditional entropy* of X given Y as follows:

$$S(X|Y) = \sum_{y\in\mathcal{Y}} S(X)_{\mathbb{P}|Y=y}\mathbb{P}(Y=y). \tag{6.2}$$

Notice that $S(X|Y)$, being the average of positive quantities, is itself positive, a fact that in the quantum case will no longer hold. From the very definitions, it is simple to check that

$$S(X|Y) = S((X,Y)) - S(Y), \tag{6.3}$$

which can be equivalently stated as the following *chain rule* for the entropy:

$$S(X,Y) = S(Y) + S(X|Y),$$

which is reminiscent of the product rule of probabilities

$$\mathbb{P}(X=x,Y=y) = \mathbb{P}(Y=y)\mathbb{P}(X=x|Y=y),$$

where products are replaced with sums because of the logarithm.

Let us collect for later use the following result.

Lemma 6.3 (Fano's inequality)**.** *Let X, X' be random variables taking values in the same set $\mathcal{X}$, and let $p=\mathbb{P}(X\neq X')$. Then,*

$$S(X|X') \leq h_2(p) + p\log(|\mathcal{X}|-1).$$

Proof. We have

$$S(X|X') = \sum_{x\in\mathcal{X}} S_{\mathbb{P}|X'=x}(X)\mathbb{P}(X'=x).$$

For every $x\in\mathcal{X}$, we introduce the random variable $I_{X=x}$ and use the chain rule to decompose

$$\begin{aligned} S(X)_{\mathbb{P}|X'=x} &= S((X,I_{X=x}))_{\mathbb{P}|X'=x} \\ &= S(I_{X=x})_{\mathbb{P}|X'=x} + S(X|I_{X=x})_{\mathbb{P}|X'=x}. \end{aligned}$$

For the first term, we simply have

$$S(I_{X=x})_{\mathbb{P}|X'=x} = h_2(\mathbb{P}(X=x|X'=x)).$$

Summation upon $x\in\mathcal{X}$ and concavity of h_2 yields

$$\begin{aligned} \sum_{x\in\mathcal{X}} h_2(\mathbb{P}(X=x|X'=x))\mathbb{P}(X'=x) &\leq h_2\left(\sum_{x\in\mathcal{X}} \mathbb{P}(X=x|X'=x)\mathbb{P}(X'=x)\right) \\ &= h_2(\mathbb{P}(X=X')) \\ &= h_2(p). \end{aligned}$$

For the second term, by definition of conditional entropy,

$$\begin{aligned} &S(X|I_{X=x})_{\mathbb{P}|X'=x} \\ &\qquad = S(X)_{\mathbb{P}|X=x,X'=x}\mathbb{P}(X=x|X'=x) + S(X)_{\mathbb{P}|X\neq x,X'=x}\mathbb{P}(X\neq x|X'=x) \\ &\qquad = S(X)_{\mathbb{P}|X\neq x,X'=x}\mathbb{P}(X\neq x|X'=x) \\ &\qquad \leq \log(|\mathcal{X}|-1)\mathbb{P}(X\neq x|X'=x), \end{aligned}$$

having used (6.1), since $X \neq x$ entails that X takes valued in the set $\mathcal{X} \setminus \{x\}$. Summation upon $x \in \mathcal{X}$ yields

$$\sum_{x\in\mathcal{X}} S(X|I_{X=x})_{\mathbb{P}|X'=x}\mathbb{P}(X'=x) \leq \log(|\mathcal{X}|-1)\mathbb{P}(X\neq x|X'=x)\mathbb{P}(X'=x)$$
$$= \mathbb{P}(X\neq X')\log(|\mathcal{X}|-1),$$

hence the thesis. □

Assume now that Y is a noisy version of the initial message X sent by Alice, which was distorted by the communication channel between the two. Then, a natural question from the point of view of the engineer is how much is the change in Bob's ignorance about X after he receives Y. This led Shannon to the definition of the *mutual information*, as

$$I(X;Y) = S(X) - S(X|Y). \tag{6.4}$$

We intuitively expect that $I(X;Y)$ should be positive since observing Y, Bob's ignorance about X should become smaller, and this is indeed the case. Notice also that one can rewrite

$$I(X;Y) = S(X) - (S(X,Y) - S(Y)) = S(X) + S(Y) - S(X,Y),$$

which shows that the mutual information is symmetric with respect to X, Y. The expression

$$I(X;Y) = S(Y) - S(Y|X)$$

can be interpreted as the information content of the signal received by Bob, minus the part $S(Y|X)$ due to noise in the communication channel. An explicit expression of $I(X;Y)$ in terms of the joint and marginal densities can be also easily obtained:

$$I(X;Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \mathbb{P}(X=x,Y=y)\log\left(\frac{\mathbb{P}(X=x,Y=y)}{\mathbb{P}(X=x)\mathbb{P}(Y=y)}\right). \tag{6.5}$$

The denominator $\mathbb{P}(X=x)\mathbb{P}(Y=y)$ would be the joint density in the case that X, Y were independent variables, and in that case $I(X;Y)=0$, which agrees with our intuition (if Bob receives pure noise, its ignorance about X will not change at all).

The expression (6.5) suggests replace the denominator with a general probability density. This of course can be done also in the case of a single random variable (or a single probability distribution). Indeed, one defines the *relative entropy* (or Kullback-Leibler divergence) of p with respect to q (both defined on a set $\mathcal{X}$) as the quantity

$$\begin{aligned} D_{KL}(p||q) &= \sum_{x\in\mathcal{X}} p(x)\log(p(x)/q(x)) \\ &= \sum_{x\in\mathcal{X}} p(x)\left(\log p(x) - \log q(x)\right) \\ &= -S(p) + \sum_{x\in\mathcal{X}} p(x)\log q(x), \end{aligned}$$

provided that $p(x)=0$ whenever $q(x)=0$ (i.e., p is absolutely continuous with respect to q), otherwise we let $D_{KL}(p||q)=\infty$. The relative entropy can be conveniently thought as a "distance" (or a distinguishability measure) between p, however it is not symmetric,

$$D_{KL}(p||q) \neq D_{KL}(q||p) \quad \text{(in general)}.$$

It is a very relevant quantity in information theory, being related to the opposite of Shannon's entropy of p, but enjoying several monotonicity and convexity properties. Indeed, given any Markov kernel $N(x,y)_{x\in\mathcal{X},y\in\mathcal{Y}}$, from $\mathcal{X}$ to $\mathcal{Y}$, the relative entropy *decreases*:

$$D_{KL}(N^\dagger p||N^\dagger q) \leq D_{KL}(p||q), \tag{6.6}$$

i.e., the two transformed probabilities $N^\dagger p$, $N^\dagger q$ become more difficult to distinguish. Notice that, by taking any kernel such that $N^\dagger p = N^\dagger q$, we obtain that

$$0 = D_{KL}(N^\dagger p || N^\dagger q) \leq D_{KL}(p||q). \tag{6.7}$$

The above property implies many other entropic inequalities. It yields easily that

$$(p, q) \mapsto D_{KL}(p||q) \quad \text{is jointly convex,}$$

which can be very useful in minimization problems.

Example 6.4. Let $E : \mathcal{X} \to \mathbb{R}$ and for $m \in \mathbb{R}$ consider the following problem: what is the probability distribution p on $\mathcal{X}$ which *maximizes* Shannon's entropy $S(p)$ with the constraint that the mean

$$\sum_{x \in \mathcal{X}} E(x)p(x) = m \tag{6.8}$$

is given ? Of course, one needs $\min E \leq m \leq \max E$. In such a case, the answer is provided by a Gibbs distribution $p_\beta(x) = e^{-\beta E}/z$, where $\beta \in [-\infty, +\infty]$, z are parameters: β (if positive) may be physically interpeted as an *inverse temperature*, while $z = z(\beta) = \sum_{x \in \mathcal{X}} e^{-\beta E(x)}$ is simply a normalization constant, also called partition function in the statistical physics literature). To prove it, let β be such that p_β has mean m: such a β exists if $\min H < m < \max H$, since the mean of H with respect to p_β is continuous as a function of β and as $\beta \to \infty$ it converges to $\min H$, while for $\beta \to -\infty$ it converges to $\max H$. Then, since $\log p_\beta = -\beta E - \log z$, for every distribution p such that (6.8) holds, we have

$$D_{KL}(p||q_\beta) = -S(p) + \beta m + \log z(\beta).$$

Thus, maximizing $p \mapsto S(p)$ is equivalent to minimize the relative entropy as a function of p, which is convex, non-negative and has a minimum at $p = q_\beta$, hence the claim is proved. One could also argue that, by strict convexity, q_β is the only maximum entropy distribution (the only degeneracies can be at $m \in \{\min E, \max E\}$). In particular, taking $E(x) = 0$, we obtain that the uniform distribution maximizes the entropy, hence the second inequality in (6.1).

Let us collect further consequences of (6.6). Recall that the mutual information between variables X, Y is equivalent to the relative entropy between the (joint) law of (X, Y) and the product of the marginal laws of X and Y,

$$I(X; Y) = D_{KL}(\mathbb{P}_{XY} || \mathbb{P}_X \otimes \mathbb{P}_Y). \tag{6.9}$$

Therefore, we obtain immediately that $I(X; Y) \geq 0$. A fundamental inequality is the *data processing* inequality, which states that, whenever (X, Y, Z) are three random variables that constitute a Markov chain, i.e., X and Z are conditionally independent given Y, it holds

$$I(X; Z) \leq I(X; Y), \tag{6.10}$$

and by symmetry $I(X; Z) \leq I(Y; Z)$ as well. Recalling that $I(X; Y)$ denotes the change in the ignorance about X after Bob receives Y, this inequality shows that any further transformation (Z) of the signal received Y (without accessing X) can only decrease such a change (of course, the larger I, the better).

To see this, consider the joint probability distribution $\mathbb{P}_{XYZ}$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. The Markov chain assumption yield that there is a Markov kernel $(N(y, z))_{y \in \mathcal{Y}, z \in \mathcal{Z}}$ from $\mathcal{Y}$ to set $\mathcal{Z}$ such that $\mathbb{P}_Z = N^\dagger \mathbb{P}_Z$, and moreover

$$\mathbb{P}_{XYZ}(x, y, z) = \mathbb{P}_{XY}(x, y)N(y, z). \tag{6.11}$$

We extend N to a kernel from $\mathcal{X} \times \mathcal{Y}$ to $\mathcal{X} \times \mathcal{Z}$ by acting trivially on $\mathcal{X}$,

$$\tilde{N}((x, y), (x', z)) = \delta_x(x')N(y, z),$$

in the quantum setting, such a kernel corresponds to the channel $\mathbb{1} \otimes N^\dagger$. Then, (6.11) yields that $\tilde{N}^\dagger \mathbb{P}_{XY} = \mathbb{P}_{XZ}$, while

$$\tilde{N}^\dagger(\mathbb{P}_X \otimes \mathbb{P}_Y) = \mathbb{P}_X \otimes N^\dagger \mathbb{P}_Y = \mathbb{P}_X \otimes \mathbb{P}_Z.$$

Plugging these identities in (6.6) yields (6.10). As a simple example of (6.10), consider the case that $Z = f(Y)$ is a function of Y. Then,

$$I(X; f(Y)) \leq I(X; Y).$$

In particular, if one replaces Y with a joint variable (Y, Z) is a joint variable and $f(y, z) = y$, we obtain that

$$I(X; Y) \leq I(X; (Y, Z)).$$

Since $I(X; Y) = S(X) - S(X|Y)$, $I(X; (Y, Z)) = S(X) - S(X|(Y, Z))$, this inequality is equivalent to

$$S(X|(Y, Z)) \leq S(X|Y)$$

which be also equivalently restated as the *strong subadditivity* property of the Shannon entropy

$$S(X, Y, Z) \leq S(X, Y) + S(Y, Z) - S(Y),$$

which improves upon the ordinary subadditivity

$$S(X, Y) \leq S(X) + S(Y),$$

which is simply equivalent to $I(X; Y) \geq 0$.

Using these inequalities we can slightly extend Lemma 6.3: given X and Y, assume that $X' = f(Y)$ and set $p = \mathbb{P}(X \neq X')$. Then,

$$S(X|Y) \leq h_2(p) + p \log(|\mathcal{X}| - 1). \tag{6.12}$$

Indeed, since $(Y, f(Y)) = (Y, X')$ is a function of Y and viceversa, then $S(X|Y) = S(X|(Y, X')$ hence

$$S(X|Y) = S(X|(Y, X')) \leq S(X|X') \leq h_2(p) + p \log(|\mathcal{X}| - 1),$$

having applied Lemma 6.3 in the last inequality.

6.2. **Quantum entropy.** Let us consider a finite-dimensional quantum system H. Given a state $\rho \in \mathcal{S}(H)$, its *von Neumann entropy* is defined as

$$S(\rho) = -\mathrm{tr}[\rho \log \rho],$$

where $\rho \log \rho$ is defined via functional calculus. Since the spectrum (with multiplicities) of ρ is contained in $[0, 1]$, it follows immediately that $S(\rho) \geq 0$, with $S(\rho) = 0$ if and only if ρ is pure (since $-z \log z$ equals zero if and only if $z \in \{0, 1\}$. To mimic the classical notation with random variables, one often writes $S(H)_\rho$ or simply $S(H)$ if the state ρ is understood.

Similarly as in the classical case, to establish properties of the von Neumann entropy, it is more convenient to consider the *quantum relative entropy* of ρ with respect to another state $\sigma \in \mathcal{S}(H)$, which is defined as

$$S(\rho||\sigma) = \mathrm{tr}[\rho(\log \rho - \log \sigma)],$$

where the operators $\rho \log \rho$ and $\log \sigma$ are defined via functional calculus, with the only caveat that we require that the kernel of σ to be contained in the kernel of ρ and we interpret $\rho(\log \rho - \log \sigma) = 0$ on the kernel of ρ. This is the equivalent of requiring absolute continuity of ρ with respect to σ. If this does not happen, we set $S(\rho||\sigma) = \infty$. If ρ and σ commute, i.e., they can be simultaneously diagonalized, then

$$S(\rho||\sigma) = D_{KL}(p||q),$$

where p is the classical probability distribution associated to the spectrum of ρ, and q to σ. Thus, the properties we are going to prove below include those of the Kullback-Leibler divergence as special cases.

Our first result is the quantum analogue of (6.6).

Theorem 6.5 (data processing inequality, DPI)**.** *Let H, $\tilde{H}$ be quantum systems and let $\Phi^\dagger$ be a quantum channel from H to $\tilde{H}$. Then, for any ρ, $\sigma \in \mathcal{S}(H)$, it holds*

$$S(\Phi^\dagger(\rho)||\Phi^\dagger(\sigma)) \leq S(\rho||\sigma). \tag{6.13}$$

The proof relies upon a differentiation argument which is often used with entropic inequalities, hence it may be worth explaining in general abstract terms. Assume that one has two functions $f, g : [a, b] \to \mathbb{R}$ such that, for $t \in [a, b]$

$$f(t) \leq g(t) \quad \text{and} \quad f(a) = g(a).$$

If both f and g are (right-)differentiable at $t = a$, then it follows that

$$f'(a) \leq g'(a).$$

Indeed, write

$$f(t) = f(a) + f'(a)(t-a) + o(t-a) \leq g(a) + g'(a)(t-a) + o(t-a) = g(t),$$

hence

$$f'(a)(t-a) \leq g'(a)(t-a) + o(t-a),$$

so that dividing by $t - a > 0$ and letting $t \to a$ yields the required inequality.

Proof. Let us recall the special case (4.21) of Lieb's concavity theorem, for $K = \mathbb{1}_{\tilde{H}}$, and $X = \rho$, $Y = \sigma$, $t \in [0, 1]$,

$$\mathrm{tr}[\rho^{1-t}\sigma^t] \leq \mathrm{tr}[\Phi^\dagger(\rho)^{1-t}\Phi^\dagger(\sigma)^t].$$

For $t = 0$, the above inequality becomes in fact an identity (Φ is trace preserving). Assuming for simplicity that ρ, σ, $\Phi^\dagger(\rho)$, $\Phi^\dagger(\sigma)$ are all invertible, then both sides in the inequality are smooth functions of t, and the thesis follows by differentiation at $t = 0$. Indeed, we have

$$\left.\frac{d}{dt}\right|_{t=0^+} \mathrm{tr}[\rho^{1-t}\sigma^t] \leq \left.\frac{d}{dt}\right|_{t=0^+} \mathrm{tr}[\Phi^\dagger(\rho)^{1-t}\Phi^\dagger(\sigma)^t].$$

We compute

$$\left.\frac{d}{dt}\right|_{t=0^+} \mathrm{tr}[\rho^{1-t}\sigma^t] = \mathrm{tr}[-\rho\log\rho + \rho\log\sigma] = -S(\rho||\sigma),$$

and similarly for the right hand side. To remove the invertibility assumptions, one should perturb ρ with a small convex combination with an invertible state, e.g. $\mathbb{1}_H/\dim(H)$, and similarly σ, but also perturb the channel $\Phi^\dagger$ by taking a small convex combination with the trivial channel $\Phi_0^\dagger(\rho) = \mathbb{1}_{\tilde{H}}/\dim(\tilde{H})$. We leave as an exercise to check that the thesis in the general case follows by a limiting argument. □

From the data processing inequality, we obtain several other properties of the quantum relative entropy and von Neumann entropy.

(1) By considering any trivial channel that maps any state into the same state, e.g. $\Phi^\dagger(\rho) = \mathbb{1}_H/\dim(H)$, it follows that $S(\rho||\sigma) \geq 0$ for every ρ, $\sigma \in \mathcal{S}(H)$.

(2) The quantum relative entropy is jointly convex, i.e.,

$$(\rho, \sigma) \mapsto S(\rho||\sigma)$$

is convex. This can be seen by proving mid-point convexity only (since it is continuous), i.e.,

$$S((\rho_0 + \rho_1)/2||(\sigma_0 + \sigma_1)/2) \leq (S(\rho_0||\sigma_0) + S(\rho_1||\sigma_1))/2.$$

In turn, this follows from an application of (6.13) to the partial trace channel $\Phi^\dagger(M) = \mathrm{tr}_2[M]$ to the states on the system $H \otimes \mathbb{C}^2$,

$$\rho = \begin{pmatrix} \rho_0 & 0 \\ 0 & \rho_1 \end{pmatrix}, \quad \sigma = \begin{pmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{pmatrix}.$$

(3) Given an observable $E \in \mathcal{O}(H)$, by translating the argument from Example 6.4 in the quantum setting, it follows that any Gibbs state $\rho_\beta = e^{-\beta E}/z$ for $\beta \in \mathbb{R}$, $z = \mathrm{tr}[e^{-\beta E}] > 0$ is a maximizer of von Neumann entropy among the states ρ such that $(E)_\rho = \mathrm{tr}[E\rho_\beta]$ is fixed. In particular, von Neumann entropy always satisfies the inequalities

$$0 \leq S(H)_\rho \leq \log(\dim(H)),$$

akin to (6.1) (of course this can be seen using much simpler arguments).

We next consider the quantum analogue of the conditional entropy. This is perhaps the most delicate quantity, since one cannot use (6.2) as no quantum analogue of the conditional density is available. However, one can directly take (6.3) as a definition: given a composite system $H \otimes K$ and a density operator $\rho \in \mathcal{S}(H \otimes K)$ with reduced density operator $\rho_H = \mathrm{tr}_K[\rho] \in \mathcal{S}(H)$, one defines the *quantum conditional entropy* as

$$S(K|H)_\rho = S(\rho) - S(\rho_H) = S(H \otimes K)_\rho - S(H)_{\rho_H}.$$

(with some abuse of notation, one may also write $S(H)_\rho$ instead of $S(H)_{\rho_H}$ and $S(HK)_\rho$ instead of $S(H \otimes K)_\rho$). Clearly, this definition ensures the validity of the chain rule

$$S(H \otimes K)_\rho = S(H)_{\rho_H} + S(K|H)_\rho.$$

However, it may *fail* to be positive (hence it cannot be in general represented as in (6.2)). This is a consequence of the existence of entangled states, and indeed negativity of this entropy can be used to spot entangled states. Without entering too much into the subject, we provide examples through the following general result, called *purification of a state*.

Proposition 6.6 (purification of a state)**.** *Given any state $\rho \in \mathcal{S}(H)$ on a quantum system, there exists an auxiliary quantum system K and a pure state $|\Psi\rangle\langle\Psi| \in \mathcal{S}(H \otimes K)$ such that*

$$\mathrm{tr}_K[|\Psi\rangle\langle\Psi|] = \rho.$$

Proof. Let $K = H^*$ be the dual of H, and consider the canonical isomorphism

$$|\psi\rangle \otimes \langle\varphi| \mapsto |\psi\rangle\langle\varphi|$$

between $H \otimes H^*$ and $\mathcal{L}(H)$. We can thus find $|\Psi\rangle \in H \otimes K$ corresponding to $\sqrt{\rho} \in \mathcal{L}(H)$. We claim that such $|\Psi\rangle$ satisfies the thesis. Indeed, by choosing an orthonormal basis $(|i\rangle)_{i\in I}$ of eigenvectors of ρ (with eigenvalues $(p_i)_{i\in I}$, we have

$$\sqrt{\rho} = \sum_{i\in I} \sqrt{p_i}\,|i\rangle\langle i|\,,$$

hence

$$|\Psi\rangle = \sum_{i\in I} \sqrt{p_i}\,|i\rangle \otimes \langle i|\,.$$

Thus,

$$|\Psi\rangle\langle\Psi| = \sum_{i,j\in I} \sqrt{p_i p_j}(|i\rangle \otimes \langle i|)(\langle j| \otimes |j\rangle),$$

and taking the partial trace yields

$$\mathrm{tr}_K[|\Psi\rangle\langle\Psi|] = \sum_{i\in I} p_i\,|i\rangle\langle i| = \rho. \qquad \square$$

By considering a purification $|\Psi\rangle\langle\Psi| \in \mathcal{S}(H \otimes K)$ for a given state $\rho \in \mathcal{S}(H)$, we see that the relative entropy must be negative, since the chain rule gives

$$0 = S(H \otimes K)_{|\Psi\rangle\langle\Psi|} = S(H)_\rho + S(K|H)_{|\Psi\rangle\langle\Psi|}.$$

Finally, we discuss the quantum analogue of the mutual information. To define it, one may use either (6.4) or (6.9), which lead to the same quantity, called the *quantum*

mutual information: given a composite quantum system $H \otimes K$ and a density operator $\rho \in \mathcal{S}(H \otimes K)$ with reduced density operators $\rho_H \in \mathcal{S}(H)$, $\rho_K \in \mathcal{S}(K)$, one defines

$$\begin{aligned} I(H;K)_\rho &= S(\rho||\rho_H \otimes \rho_K) \\ &= S(H)_{\rho_H} - S(H|K)_\rho \\ &= S(H)_{\rho_H} + S(K)_{\rho_K} - S(H \otimes K)_\rho. \end{aligned}$$

We leave as an exercise to check indeed that these are all equivalent expressions. The data processing inequality entails the following result: given a state $\rho \in \mathcal{S}(H \otimes K)$ and a quantum channel $\Phi^\dagger$ from K to $\tilde{K}$, then

$$I(H;\tilde{K})_{\mathbb{1}_{\mathcal{L}(H)} \otimes \Phi^\dagger(\rho)} \leq I(H;K)_\rho. \tag{6.14}$$

Indeed, it is sufficient to apply (6.13) to the channel $\mathbb{1}_{\mathcal{L}(H)} \otimes \Phi^\dagger$ from $H \otimes K$ to $H \otimes \tilde{K}$, and notice that

$$\mathrm{tr}_{\tilde{K}}[\mathbb{1}_{\mathcal{L}(H)} \otimes \Phi^\dagger(\rho)] = \mathrm{tr}_K[\rho], \quad \mathrm{tr}_H[\mathbb{1}_{\mathcal{L}(H)} \otimes \Phi^\dagger(\rho)] = \Phi^\dagger(\rho),$$

(one can check these identities with $M \otimes N$ in place of ρ, and then argue by linearity).

Replacing K with a composite system $K \otimes L$ and letting $\Phi^\dagger = \mathrm{tr}_L$ be the partial trace channel, we obtain that, for every $\rho \in \mathcal{S}(H \otimes K \otimes L)$,

$$I(H;K)_{\rho_{HK}} \leq I(H;K \otimes L)_\rho,$$

which can be equivalently rewritten and the *strong subadditivity* of von Neumann entropy

$$S(H \otimes K \otimes L) \leq S(H \otimes K) + S(K \otimes L) - S(K).$$

6.3. **Spin systems and specific quantities.** Consider an infinite (countable) set I and a collection of (elementary) quantum systems $(H_i)_{i \in I}$. In this section we show how to rigorously define a notion of infinite tensor product

$$\bigotimes_{i \in I} H_i, \tag{6.15}$$

relying upon the C^*-algebra framework. The physical interpretation of such a system is that I corresponds to a set of sites (think e.g. of a lattice-like structure, $I = \mathbb{Z}^d$), each occupied by a particle, and we are considering the joint system of the whole structure. The system H_i may e.g. describe the spin of the particle at the site i, hence the name *spin systems*Another interpretation is that $I = \mathbb{Z}$ or $I = \mathbb{N}$ is a set of "times" and we are describing the "path space" of a dynamical system. Considering infinite sets I is particularly relevant in order to describe phenomena, such as phase transitions, which would otherwise difficult to investigate in finite systems. Notice that the probabilistic counterpart of such a construction is that of product set and of measures on the product space (in particular, infinite product measures) and similar considerations apply as well (think e.g. of a collection of i.i.d. random variables).

Back to the construction, for any finite set $\Lambda \subseteq I$, define the algebra

$$\mathfrak{U}_\Lambda = \mathcal{L}\left(\bigotimes_{i \in \Lambda} H_i\right).$$

Given $\Lambda' \subseteq \Lambda \subseteq I$ (both finite), one has a natural injective $*$-homomorphism

$$\pi_{\Lambda',\Lambda} : \mathfrak{U}_{\Lambda'} \to \mathfrak{U}_\Lambda, \quad A \mapsto A \otimes \mathbb{1}_{\Lambda \setminus \Lambda'},$$

which allows essentially to identify $\mathfrak{U}_{\Lambda'}$ as a sub-algebra of $\mathfrak{U}_\Lambda$. The union

$$\mathfrak{U}_I^{loc} = \bigcup_{\Lambda \subseteq I} \mathfrak{U}_\Lambda$$

defines the *strictly local algebra* associated to the spin system. The operator norm is well defined on $\mathfrak{U}_I^{loc}$, but it does not define a Banach space structure on it. However,

it is sufficient to consider its (abstract) completion to have a well-defined structure of C^*-algebra[13] $\mathcal{U}_I$, associated to the spin system $(H_i)_{i\in I}$.

For every finite $\Lambda \subseteq I$, the inclusion maps extend to $\mathcal{U}_I$ and define $*$-homomorphisms

$$\pi_\Lambda : \mathcal{U}_\Lambda \to \mathcal{U}_I.$$

By duality, it induces a "partial trace" map $\mathrm{tr}_{I\setminus\Lambda}$. In particular, given $\rho : \mathcal{U}_I \to \mathbb{C}$ one has a well-defined *reduced* density operator $\rho_\Lambda \in \mathcal{S}(\bigotimes_{i\in\Lambda} H_i)$. Consistency between the inclusion maps yield by duality the identity between the partial traces

$$\mathrm{tr}_{\Lambda\setminus\Lambda'} \circ \mathrm{tr}_{I\setminus\Lambda} = \mathrm{tr}_{I\setminus\Lambda'},$$

hence

$$\rho_{\Lambda'} = \mathrm{tr}_{\Lambda\setminus\Lambda'}[\rho_\Lambda] \quad \text{for every } \Lambda' \subseteq \Lambda. \tag{6.16}$$

Viceversa, given any collection of density operators $(\rho_\Lambda)_{\Lambda\subseteq I}$ with $\rho_\Lambda \subseteq \mathcal{S}(\bigotimes_{i\in\Lambda} H_i)$ over the finite subsets $\Lambda \subseteq I$ such that (6.16) holds, they define uniquely a functional ρ over $\mathfrak{U}_I^{loc}$, which extends by continuity to a state on $\mathfrak{U}_I$ (since each ρ_Λ is a state, one has by (5.5) that $|\mathrm{tr}[A\rho_\lambda]| \leq \|A\|_\infty$). For example, product states $\rho = \bigotimes_{i\in I} \rho_i$ are thus well defined for every family $(\rho_i)_{i\in I}$ of density operators with $\rho_i \in \mathcal{S}(H_i)$, and provide the analogue for the infinite product construction in probability.

It is common to restrict the study to the homogeneous settings, where each H_i's are equal (e.g. to a copy of $\mathbb{C}^d$). In such a case, any bijection $\tau : I \to I$ induces a $*$-isomorphism between $\mathcal{U}_\Lambda$ and $\mathcal{U}_{\tau(\Lambda)}$, defined on tensor product operators as

$$\Phi_\tau(\bigotimes_{i\in\Lambda} A_i) = \bigotimes_{i\in\tau(\Lambda)} A_{\tau^{-1}(i)}.$$

E.g. if $\tau(1) = 2$, $\tau(2) = 1$, this amounts to swapping the roles of the operators $\tau(A_1 \otimes A_2) = A_2 \otimes A_1$. Clearly, such a family of operators acts consistently on the families $\mathfrak{U}_\Lambda$, hence can be used to define a well-defined $*$-automorphism Φ_τ of $\mathfrak{U}_I$ (i.e., an $*$-isomorphism of $\mathcal{U}_I$ into itself).

By duality, a state ρ on $\mathcal{U}_I$ is τ-invariant if $\rho(\Phi_\tau(A)) = \rho(A)$ for every $A \in \mathcal{U}_I$ (or equivalently, for every $A \in \mathcal{U}_I^{loc}$). If I has additional structure, e.g. it is a group or a homogeneous space, one can isolate the sets of *invariant* states with respect to the $*$-automorphism induced by the group action. For example, states that are invariant with respect to any permutation $\tau : I \to I$ (i.e., that is non-trivial only on a finite subset) provide the analogues of exchangeable probabilities. If $I = \mathbb{Z}$, invariant states with respect to translations (i.e., with respect to $\tau(n) = n + 1$) are the analogues of (the laws) of stationary processes. These states are particularly relevant for modelling reasons.

It turns out that defining the entropy as $\lim_{\Lambda\uparrow I} S(\rho_\Lambda)$ is not very useful, on invariant states ρ, since already on simple examples, e.g. a product state $\bigotimes_{i\in I} \rho_i$ with $\rho_i = \rho_0$ for every i, one obtains $S(\rho_\Lambda) = |\lambda| S(\rho_0)$. This motivates the introduction of an entropy density with respect to the number of sites, called *specific entropy*. When $I = \mathbb{Z}$, it is defined on any invariant state ρ as

$$s(\rho) = \lim_{n\to\infty} \frac{S(\rho_{[1,n]})}{n}.$$

Existence of the limit follows from a sub-additivity argument.

For other "extensive" quantities, one can define a similarly a notion of density with respect to the number of sites. For example, in [De +21], it is introduced a notion of specific quantum Wasserstein distance of order 1 between two invariant states, $\rho, \sigma \in \mathfrak{U}_\mathbb{Z}$ as

$$w_1(\rho, \sigma) = \lim_{n\to\infty} \frac{\|\rho_{[1,n]} - \sigma_{[1,n]}\|_{W_1}}{n}.$$

[13] such direct limit construction is also called *uniformly hyperfinite* (UHF) algebra

6.4. **Exercises.**

Exercise 6.1. For $\alpha, \beta \in [0,1]$, show directly by an application of Jensen inequality to the (strictly) convex function $z \mapsto z \log z$ that

$$D_{KL}((\alpha, 1-\alpha)||(\beta, 1-\beta)) \geq 0,$$

with equality if and only if $\alpha = \beta$. Deduce that, for some constant $c > 0$ (independent of α and β), it holds

$$D_{KL}((\alpha, 1-\alpha)||(\beta, 1-\beta)) \geq c|\alpha - \beta|^2. \tag{6.17}$$

Exercise 6.2. Complete the proof of Theorem 6.5 for general ρ, σ and channel $\Phi^\dagger$.

Exercise 6.3 (Quantum Pinsker inequality)**.** Using (6.17) and the dual characterization of the trace distance (5.7), show that, there exists a constant $c > 0$ such that, for any finite-dimensional quantum system H and density operators $\rho, \sigma \in \mathcal{S}(H)$, it holds

$$S(\rho||\sigma) \geq c D_{\mathrm{tr}}(\rho, \sigma)^2.$$

Exercise 6.4. Compute the von Neumann entropy of a single qubit state ρ in terms of its Bloch ball representation (2.7).

7. A quantum coding theorem

In this final section, we address a quantum version of the fundamental Shannon's limit, which quantifies the maximum rate of data that can theoretically be transferred through a noisy communication channel. This result, first presented [Sha48] is considered to be a foundational one in information theory. The question whether one may go beyond Shannon's limit, by exploiting quantum mechanical features is both natural and of great interest in applications.

7.1. **The classical case.** Recall the example from the last section, where Alice and Bob agree upon sending a message codified as a random variable X taking values in set $\mathcal{X}$, through a noisy communication channel, so that Bob receives a possibly distorted version Y (maybe even taking values in a different alphabet $\mathcal{Y}$). The channel itself is modelled as a Markov kernel N from $\mathcal{X}$ to $\mathcal{Y}$, so that if $p = p_X$ is the law of the variable X, then $N^\dagger(p)$ is the law of Y. More generally, one can define the joint law of (X, Y) as the distribution on $\mathcal{X} \times \mathcal{Y}$ as

$$\mathbb{P}(X = x, Y = y) = p(x)N(x, y). \tag{7.1}$$

Alice and Bob preliminarily agree to use iterated applications of the channel and transmit the message via a suitable *coding* procedure. Let us give some precise definition: given $n \geq 1$, one defines the composite memoryless channel $N^{\otimes n}$ as the kernel from $\mathcal{X}^n$ to $\mathcal{Y}^n$,

$$N^{\otimes n}((x_i)_{i=1}^n, (y_i)_{i=1}^n) = \prod_{i=1}^n N(x_i, y_i).$$

The idea is encode a message through a *code* (W, V), which consists of

i) a *codebook* $W : \{1, \dots, m\} \to \mathcal{X}^n$ consisting of m codewords $(W(i))_{i=1}^m$, each of a fixed length n, to be transmitted by Alice via the composite channel $N^{\otimes n}$,
ii) a *decision rule*, $V : \mathcal{Y}^n \to \{0, 1, \dots, m\}$, which assigns, to each observed word $y \in \mathcal{Y}^n$, Bob's *estimate* about the word transmitted by Alice: if $V(y) = i$, $i \neq 0$, then Bob decodes y as $W(i)$; if $V(y) = 0$, then Bob makes no decision.

Given a code (W, V) with m codewords (the "size" of the code), each of length n, Alice needs n applications of the channel (up to a constant, n bits of information) to send m codewords, i.e., $\log m$ bits of information. The *transmission rate*, i.e. the number of bits of information per application of the channel is therefore $\log m/n$.

Of course, what matters in the picture is that the message sent by Alice is correctly decoded by Bob. For each $i \in \{1, \dots, m\}$, the probability that Bob decodes correctly the word, given that Alice sent word i, is

$$\mathbb{P}(V(y) = i|W(i)) = \sum_{y \in \{V=i\}} N^{\otimes n}(W(i), y) = \sum_{y \in \{V=i\}} \prod_{j=1}^{n} N(y_j|W(i)_j).$$

We introduce the following indicators:

(1) the *maximal error probability*

$$p_e(W, V) = \max_{i=1,\dots,m} (1 - \mathbb{P}(V(y) = i|W(i)),$$

(2) the *mean error probability*

$$\bar{p}_e(W, V) = \frac{1}{m} \sum_{i=1}^{m} (1 - \mathbb{P}(V(y) = i|W(i)).$$

Clearly,

$$\bar{p}_e(W, V) \leq p_e(W, V).$$

A simple application of Markov inequality yields the following fact: given any code (W, V) with $2m$ codewords, there exists a sub-code $(\tilde{W}, \tilde{V})$ (i.e., $\tilde{W}$ obtained by restricting W to a subset of $I \subseteq \{1, \dots, 2m\}$, and setting $\tilde{V} = V$ on $\{V = i\}$ for $i \in I$, $\tilde{V} = 0$ otherwise) of size at least m such that

$$p_e(W, V) \leq 2\bar{p}_e(\tilde{W}, \tilde{V}). \tag{7.2}$$

Given n and m, denote by

$$p_e(n, m) = \min_{(W,V)} p_e(W, V), \quad \bar{p}_e(n, m) = \min_{(W,V)} \bar{p}_e(W, V),$$

where the minimum is among all *codes* with m codewords of length n. With this notation, we say that $r > 0$ is an *achievable transmission rate* for the channel N if

$$\lim_{n \to \infty} p_e(n, 2^{nr}) = 0.$$

Roughly speaking, by using suitable codes, Alice can transmit to Bob 2^{nr} words, using n applications of the channel, with an infinitesimal error as $n \to \infty$. Notice that, by (7.2), one can equivalently use $\bar{p}_e$ instead of p_e in the definition above.

From the engineer's perspective, it is then natural to define the *(operational) channel capacity* $\mathcal{C}(N)$ as the supremum among the achievable transmission rates r, so that it holds

- (direct statement) for every $r < \mathcal{C}(N)$,

$$\lim_{n \to \infty} \bar{p}_e(n, 2^{nr}) = 0,$$

- (weak converse) for every $r > \mathcal{C}(N)$,

$$\limsup_{n \to \infty} \bar{p}_e(n, 2^{nr}) > 0,$$

  i.e., Alice cannot use codes to transmit information to Bob with a rate r.

Almost immediately from the definition, we have that the channel capacity is an *additive* quantity, i.e.,

$$\mathcal{C}(N^{\otimes k}) = k\mathcal{C}(N), \quad \text{for every } k \geq 1. \tag{7.3}$$

Indeed, given any code (W, V) of 2^{nr} codewords, each of length n, can be used to obtain a code (W^k, V^k) (essentially by concatenating k independent copies of the code), consisting

of 2^{nkr} codewords, of length nk, such that, for every $y_1, \dots, y_k \in (\mathcal{Y}^n)^k$, and $i_1, \dots, i_k \in \{0, 1, \dots, m\}^k$,

$$\begin{aligned}\mathbb{P}(V^k(y_1, y_2, \dots, y_k) = (i_1, \dots, i_k) | X^k = (W(i_1), \dots, W(i_k))) \\ = \prod_{j=1}^{k} \mathbb{P}(V(y_j) = i_j | X = W_{i_j}).\end{aligned}$$

Letting $n \to \infty$, we obtain that for every achievable rate r for N, the rate kr is achievable for $N^{\otimes k}$, hence $k\mathcal{C}(N) \le \mathcal{C}(N^{\otimes k})$. On the other side, given a code (W, V) using the channel $N^{\otimes k}$, consisting of $m = 2^{ns}$ codewords each of length n (i.e., requiring n applications of the channel $N^{\otimes k}$), this can be also interpreted as a code with respect to the original channel N with length kn, hence for every achievable rate s for $N^{\otimes k}$ we have that s/k is achievable for N, i.e. $\mathcal{C}(N^{\otimes k})/k \le \mathcal{C}(N)$.

Shannon's noisy channel coding theorem, also called *Shannon's limit* precisely characterizes $\mathcal{C}(N)$ in terms of entropic quantities, in particular related to a *single* application of the channel. Recall that we introduced the mutual information

$$I(X; Y) = S(X) - S(X|Y) = S(Y) - S(Y|X)$$

as a measure of how much the information gain of Bob, about X, after he receives Y. If we use the joint law (7.1), we have the identity

$$I(X; Y) = S\left(\sum_{x \in \mathcal{X}} p(x) N(x, \cdot)\right) - \sum_{x \in \mathcal{X}} p(x) S(N(x, \cdot)),$$

which shows that $I(X; Y)$ is (in this particular case) a concave function of the probability distribution $p = p_X$ of the message X that Alice may choose to send. Taking once again the point of view of the engineer whose aim is to optimize the communication, it is natural to maximize $I(X; Y)$ among all such distributions, thus defining the *information channel capacity* as

$$\mathcal{C}_I(N) = \max_p I(X; Y) = \max_p \left\{ S\left(\sum_{x \in \mathcal{X}} p(x) N(x, \cdot)\right) - \sum_{x \in \mathcal{X}} p(x) S(N(x, \cdot)) \right\}.$$

It turns out that this quantity coincides with the operational channel capacity $\mathcal{C}(N)$.

Theorem 7.1 (Shannon's limit)**.** *It holds*

$$\mathcal{C}(N) = \mathcal{C}_I(N). \tag{7.4}$$

Let us give only a brief sketch of the main ideas proof, as its fundamental ideas are further elaborated in the quantum case. The argument is split into two parts: the weak converse statement, i.e., inequality $\le$ in (7.4), is obtained via an application of Fano's inequality; the direct statement, i.e., inequality $\ge$, instead is obtained via a in *random coding* argument.

Weak converse. Let (W, V) be any code of size m and codewords of length n. Then, we turn W itself into a random variable $X^n = W$ taking m values among those of $\mathcal{X}^n$, by assuming uniform distribution on the m codewords. By applying the composite memoryless channel $N^{\otimes n}$ (i.e., formally applying (7.1) with $N^{\otimes n}$ instead of N and the uniform law over the m codewords), we have that V becomes a random variable with values in $\{0, 1, \dots, m\}$, which we can use to define an estimator W' of W (in case $V = 0$, estimate W by any rule, it does not matter), with error probability

$$\mathbb{P}(W' \neq W) \le \frac{1}{m} \sum_{i=1}^{m} (1 - \mathbb{P}(V = i | W(i)) = \bar{p}_e(W, V).$$

Fano's inequality (6.12) (cleverly applied only to the m values of W instead of the whole $\mathcal{X}^n$)

$$S(W|V) \leq h_2(\bar{p}_e(W,V)) + \bar{p}_e(W,V)\log(m-1) \leq 1 + \bar{p}_e(W,V)\log m.$$

Since W is uniform over m values, we have $S(W) = \log m$, hence

$$\mathcal{C}_I(N^{\otimes n}) \geq I(W;V) = S(W) - S(W|V) \geq \log m - \bar{p}_e(W,V))\log m - 1,$$

i.e.,

$$\bar{p}_e(W,V)) \geq 1 - \frac{\mathcal{C}_I(N^{\otimes n}) + 1}{\log m}.$$

To turns this into an asymptotic lower bound, write $m = 2^{nr}$ so that

$$\bar{p}_e(n, 2^{nr}) \geq 1 - \frac{\mathcal{C}_I(N^{\otimes n}) + 1}{nr}.$$

Letting $n \to \infty$, this leads to the inequality

$$\limsup_{n\to\infty} \bar{p}_e(n, 2^{nr}) \geq 1 - \frac{1}{r}\liminf_{n\to\infty} \frac{\mathcal{C}_I(N^{\otimes n})}{n},$$

which shows that any r such that

$$r > \liminf_{n\to\infty} \frac{\mathcal{C}_I(N^{\otimes n})}{n}$$

is not an achievable transmission rate, hence

$$\mathcal{C}(N) \leq \liminf_{n\to\infty} \frac{\mathcal{C}_I(N^{\otimes n})}{n}.$$

To conclude this part, we need to argue that

$$\liminf_{n\to\infty} \frac{\mathcal{C}_I(N^{\otimes n})}{n} = \mathcal{C}_I(N). \tag{7.5}$$

As we are going to argue similarly in the quantum case, let us first notice that is straightforward to prove $\mathcal{C}_I$ is super-additive, i.e.,

$$\mathcal{C}_I(N^{\otimes(k+h)}) \geq \mathcal{C}_I(N^{\otimes k}) + \mathcal{C}_I(N^{\otimes h}) \tag{7.6}$$

as a simple consequence of tensorization properties of relative entropy. It then follows from Fekete's subadditive lemma that the lim sup above is actually a limit and

$$\lim_{n\to\infty} \mathcal{C}_I(N^{\otimes n})/n = \inf_{n\geq 1} \mathcal{C}_I(N^{\otimes n})/n.$$

To prove (7.6), consider two distributions p_k, p_h that are maximizers respectively for $\mathcal{C}_I(N^{\otimes k})$ and $\mathcal{C}_I(N^{\otimes h})$. Then, considering the product distribution $p_k \otimes p_h$ to define a random variable $X^{k+h} = (X^k, X^h)$, on the set $\mathcal{X}^{k+h}$, to which we apply the kernel $N^{\otimes k+h}$ yielding variables $Y^{k+h} = (Y^k, Y^h)$, we see that the joint distribution of (X^k, Y^k), (X^h, Y^h) are independent. Therefore, by a general property of the relative entropy,

$$\begin{aligned} \mathcal{C}_I(N^{\otimes k+h}) &\geq I(X^{k+h};Y^{k+h})_{p_k\otimes p_h} \\ &= I(X^k;Y^k)_{p_k} + I(X^h;Y^h)_{p_h} \\ &= \mathcal{C}_I(N^{\otimes k}) + \mathcal{C}_I(N^{\otimes h}). \end{aligned} \tag{7.7}$$

Therefore, to obtain (7.5), we need to argue that the information channel capacity is additive, i.e., equality holds above. To see this, recall that $p \mapsto I(X;Y)_p$ is a concave function, hence to ensure that a distribution p is a maximizer it is sufficient to be a stationary point. It then follows from the identity

$$I(X^{k+h};Y^{k+h})_{p_k\otimes p_h} = I(X^k;Y^k)_{p_k} + I(X^h;Y^h)_{p_h}$$

that $p_k \otimes p_h$ is also a stationary point, hence equality holds in (7.7).

Direct statement. Here we only sketch the strategy in the proof, which consists of building a *random code*, since we are going more into details in the quantum case.

Ultimately, we need to build a connection between *information-theoretical* quantities such as entropy and conditional entropy to *operational* ones, i.e., related to an i.i.d. family of random variables $(X_i)_{i=1}^n$. This is obtained via the law of large numbers, ensuring that

$$S(p) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) = -\mathbb{E}\left[\log p(X)\right]$$

can be approximated by the empirical average

$$-\frac{1}{n}\sum_{i=1}^n \log p(X_i) = -\frac{1}{n}\log\left(\prod_{i=1}^n p(X_i)\right).$$

If we interpret $W = (X_i)_{i=1}^n$ a random *word* with letters sampled from the alphabet $\mathcal{X}$ (each letter with law p) this connection yields that a *typical word* will have probability of occurrence approximatively equal to

$$\mathbb{P}(W = w) = \prod_{i=1}^n p(X_i) \approx 2^{-nS(p)},$$

i.e., for sufficiently large n, the random variable W effectively behaves as a uniformly distributed variable over $2^{nS(p)}$ values. In fact, this observation is the starting point in the proof of another fundamental result by Shannon, the source coding theorem.

Back to the noisy channel setting, we need to build a code (W, V) to allow communication between Alice and Bob with any rate $r < \mathcal{C}_I(N)$. We begin by sampling $m = 2^{nr}$ independent words $(W(i))_{i=1}^m$, each of length n, according to a single letter distribution p (which eventually should be taken as the maximizer for $\mathcal{C}_I(N)$).

We define the decision rule V. After receiving y, Bob will argue as follows:

(1) First, he checks if y is a typical word for Y, otherwise he sets $V(y) = 0$ (effectively restricting to $\approx 2^{nS(Y)}$ words)
(2) For every $i \in \{1, \dots, m\}$, he checks (in sequence) if y is *conditionally typical* for the word $W(i)$ in the codebook, i.e.,

$$\mathbb{P}(y|W(i)) \approx 2^{-nS(Y|X)}.$$

He stops at the first affirmative case decodes $V(y) = i$. If no word is conditionally typical, he sets $V(0) = 0$.

The output of the channel effectively behaves as a uniformly distributed variable over $2^{nS(Y)}$ values, while on average, for each codeword in W, we have $2^{nS(Y|X)}$ conditionally typical outputs. By independence, the typical outputs should be well-separated. Thus, we expect to be able to build a code of size m,

$$m \approx \frac{2^{nS(Y)}}{2^{nS(Y|X)}} = 2^{n(S(Y)-S(Y|X))} = 2^{nI(X;Y)_p},$$

with asymptotically small error. Finally, choosing p in order to maximize $I(X;Y)_p$ would lead to $\mathcal{C}(N) \geq \mathcal{C}_I(N)$.

7.2. **The quantum case.** In the quantum framework, it is tempting to immediately replace the Markov kernel N with a quantum channel Φ, however, to keep the exposition simple, we focus on the following intermediate case of a *classical* to *quantum* channel, which transforms each letter in the input alphabet $\mathcal{X}$ for Alice into a quantum state in a suitable finite-dimensional quantum system (which belongs to Bob):

$$\Phi : \mathcal{X} \ni x \mapsto \Phi_x \in \mathcal{S}(H).$$

Of course, we can also extend the action of Φ to a quantum channel from the quantum system $\mathbb{C}^{\mathcal{X}}$ (denoting its canonical basis as $(|x\rangle)_{x\in\mathcal{X}}$) into H, given by the Kraus representation

$$\Phi(\rho) = \sum_{x\in\mathcal{X}} \sqrt{\Phi_x}\,\langle x|\rho x\rangle\,\sqrt{\Phi_x},$$

for every $\rho \in \mathcal{S}(\mathcal{X}^{\mathcal{X}})$. With this notation, we see that $\Phi_x = \Phi(\langle x|\,|x\rangle)$ (we can also identify $\langle x|\,|x\rangle$ as the Dirac delta distribution at $x \in \mathcal{X}$).

Of course, after receiving a quantum state, Bob will have to perform some measurement on the state in order to extract the information about Alice's message. We consider general *non-sharp* measurements $M = (M_y)_{y\in\mathcal{Y}}$ given by POVM's, i.e., $M_y \in \mathcal{O}_{\geq 0}(H)$ and such that

$$\sum_{y\in\mathcal{Y}} M_y = \mathbb{1}_H.$$

This family strictly includes that of measurements $V = (\mathbb{1}_{V_y})_{y\in\mathcal{Y}}$, but its use greatly simplifies the mathematical derivation. Although one slightly loses the analogy with the classical case, where the decision rule was a function: it is like allowing for probabilistic decision rules (i.e., given by Markov kernels) in the classical case. Recall anyway that POVM allows us to construct a quantum channel, hence have an operational meaning. Precisely, it is useful to associated to M a *quantum* to *classical* channel Φ_M, from H to $\mathbb{C}^{\mathcal{Y}}$, given by

$$\mathcal{S}(H) \ni \rho \mapsto \sum_{y\in\mathcal{Y}} \mathrm{tr}[M_y\rho]\,|y\rangle\,\langle y|\,,$$

(notice that the transformed state is always diagonal in the standard basis, hence in this sense it is classical). We leave as an exercise to check that Φ_M is a channel, i.e., CP and trace preserving, i.e., $(\mathrm{tr}[M_y\rho])_{y\in\mathcal{Y}}$ is a classical probability distribution. We interpret $\mathrm{tr}[M_y\rho] = \mathbb{P}_\rho(M = y)$ as the probability that, measuring M, we obtain y (given that the system is in the state ρ).

Following the analogy with the classical case, we assume that repeated applications of the channel are memoryless, in the sense that a word $w = (x_i)_{i=1}^n \in \mathcal{X}^n$ sent by Alice arrives to Bob as the product state

$$\Phi_w = \Phi_{x_1} \otimes \Phi_{x_2} \otimes \ldots \otimes \Phi_{x_n} \in \mathcal{S}(H^{\otimes n}).$$

In other words, n applications of Φ correspond to a single application of the channel $\Phi^{\otimes n}$.

Following the scheme of the classical case, Alice and Bob will agree to encode a message through a *code* (W, M), which consists of

i) a (classical) *codebook* $W : \{1,\ldots,m\} \to \mathcal{X}^n$ consisting of m codewords $(W(i))_{i=1}^m$, each of a fixed length n, to be transmitted by Alice,
ii) a *quantum decision rule*, consisting of a non-sharp measurement

$$M = (M_i)_{i=0,\ldots m} \subseteq \mathcal{O}_{\geq 0}(H^{\otimes n}),$$

such that

$$\sum_{i=0}^{m} M_i = \mathbb{1}_{H^{\otimes n}}$$

As in the classical case, for a code of m codewords, each of length n, we say that its *transmission rate* is $\log m/n$. For each $w \in \mathcal{X}^n$ and $j \in \{1,\ldots,m\}$, we define the probability that Bob decodes the word j, given that Alice sent the word w, as

$$\mathbb{P}(\text{“measures } M \text{ and observes } j\text{”}|\text{“Alice sent the word } w\text{”}) = \mathbb{P}_{\Phi_w}(M = j) = \mathrm{tr}[M_j\Phi_w].$$

Using this quantity, we define

(1) the *maximal error probability*

$$p_e(W, M) = \max_{j=1,\ldots,m} (1 - \mathrm{tr}[M_j\Phi_{W(j)}]),$$

(2) the *mean error probability*

$$\bar{p}_e(W,M) = \frac{1}{m}\sum_{i=1}^{m}\left(1 - \mathrm{tr}[M_j \Phi_{W(j)}]\right).$$

As in the classical case, the two errors are comparable (at least in the large m asymptotics), we set

$$p_e(n,m) = \min_{(W,M)} p_e(W,M), \quad \bar{p}_e(n,m) = \min_{(W,M)} \bar{p}_e(W,M),$$

and we say that $r > 0$ is an *achievable transmission rate* for the channel Φ if

$$\lim_{n\to\infty} p_e(n, 2^{nr}) = 0.$$

The *(operational classical) channel capacity* $\mathcal{C}(\Phi)$ is defined as the supremum among the achievable transmission rates r, so that it holds

- (direct statement) for every $r < \mathcal{C}(\Phi)$,
$$\lim_{n\to\infty} \bar{p}_e(n, 2^{nr}) = 0,$$
- (weak converse) for every $r > \mathcal{C}(\Phi)$,
$$\limsup_{n\to\infty} \bar{p}_e(n, 2^{nr}) > 0,$$
i.e., Alice cannot use codes to transmit classical information to Bob with a rate r without errors.

As in the classical case, we have that the channel capacity is an *additive* quantity, i.e.,

$$\mathcal{C}(\Phi^{\otimes k}) = k\mathcal{C}(\Phi), \quad \text{for every } k \geq 1.$$

What should be a candidate for the information channel capacity in this setting? For a classical-quantum channel $\Phi = (\Phi_x)_{x\in\mathcal{X}}$ and a probability distribution p on $\mathcal{X}$, we introduce the quantity

$$\chi(\Phi)_p = S\left(\sum_{x\in\mathcal{X}} p(x)\Phi_x\right) - \sum_{x\in\mathcal{X}} p(x) S(\Phi_x),$$

where S denotes von Neumann entropy. Clearly, this is the analogue of the mutual information of the classical noisy channel, when Alice sends X according to the distribution p. Notice that, by concavity of von Neumann entropy, which follows from the identity

$$S(\rho) = -S(\rho||\mathbb{1}_H/\dim(H)) - \log\dim(H).$$

and convexity of relative entropy, $p \mapsto \chi(\Phi)_p$ is concave. By further pursuit of the analogy, we define the χ-*capacity* of Φ as

$$\mathcal{C}_\chi(\Phi) = \max_p \chi(\Phi)_p = \max_p \left\{ S\left(\sum_{x\in\mathcal{X}} p(x)\Phi_x\right) - \sum_{x\in\mathcal{X}} p(x) S(\Phi_x) \right\}.$$

Schumacher, Westmoreland and independently Holevo proved that such quantity coincides with the channel capacity $\mathcal{C}(\Phi)$, thus providing a quantum analogue of Shannon's limit.

Theorem 7.2 (HSW)**.** *It holds*

$$\mathcal{C}(\Phi) = \mathcal{C}_\chi(\Phi).$$

Notice that, in any case,

$$\mathcal{C}_\chi(\Phi) \leq \max_p S\left(\sum_{x\in\mathcal{X}} p(x)\Phi_x\right) \leq \log\dim H,$$

this imposing a non-trivial limit on the transmission rate of information. As in the classical case, we split the argument considering first the weak converse and then the direct statement.

Weak converse. For the weak converse, our aim is to reduce to the classical bound, by considering the entire composition of *channel* and *measurement* as a single classical channel with kernel $N(x,y) = \mathrm{tr}[M_y \Phi_x]$. Indeed, notice that

$$\chi(\Phi)_p = I(\mathbb{C}^{\mathcal{X}}; H)_\rho = S(\rho || \rho_{\mathbb{C}^{\mathcal{X}}} \otimes \rho_H), \tag{7.8}$$

where we define the state $\rho \in \mathcal{S}(\mathbb{C}^{\mathcal{X}} \otimes H)$ as

$$\rho = \sum_{x \in \mathcal{X}} p(x) \left|x\right\rangle \left\langle x\right| \otimes \Phi_x. \tag{7.9}$$

Indeed, the reduced density operators are

$$\rho_{\mathbb{C}^{\mathcal{X}}} = \mathrm{tr}_H[\rho] = \sum_{x \in \mathcal{X}} p(x) \left|x\right\rangle \left\langle x\right|, \quad \rho_H = \mathrm{tr}_{\mathbb{C}^{\mathcal{X}}}[\rho] = \sum_{x \in \mathcal{X}} p(x) \Phi_x,$$

we have

$$\log(\rho_{\mathbb{C}^{\mathcal{X}}} \otimes \rho_H) = \sum_{x \in \mathcal{X}} \log p(x) \left|x\right\rangle \left\langle x\right| \otimes \mathbb{1}_H + \mathbb{1}_{\mathbb{C}^{\mathcal{X}}} \otimes \log \rho_H.$$

Moreover,

$$\log \rho = \sum_{x \in \mathcal{X}} \left|x\right\rangle \left\langle x\right| \left(\log p(x) \otimes \mathbb{1}_H + \log \Phi_x\right),$$

hence

$$\log \rho - \log(\rho_{\mathbb{C}^{\mathcal{X}}} \otimes \rho_H) = \sum_{x \in \mathcal{X}} \left|x\right\rangle \left\langle x\right| \log \Phi_x - \mathbb{1}_{\mathbb{C}^{\mathcal{X}}} \otimes \log \rho_H.$$

Multiplying by ρ and taking the trace, we obtain (7.8), since

$$\mathrm{tr}[\rho \sum_{x \in \mathcal{X}} \left|x\right\rangle \left\langle x\right| \log \Phi_x] = \sum_{x \in \mathcal{X}} p_x \mathrm{tr}[\Phi_x \log \Phi_x] = -\sum_{x \in \mathcal{X}} p_x S(\Phi_x).$$

and

$$\mathrm{tr}[\rho \mathbb{1}_{\mathbb{C}^{\mathcal{X}}} \otimes \log \rho_H] = \mathrm{tr}[\rho_H \log \rho_H] = -S(\rho_H).$$

Given any quantum to classical channel Φ_M corresponding to a non-sharp measurement $M = (M_y)_{y \in \mathcal{Y}}$, we have by the data processing inequality (6.14) that

$$\mathcal{C}_\chi(\Phi) \geq I(\mathbb{C}^{\mathcal{X}}; H)_\rho \geq I(\mathbb{C}^{\mathcal{X}}; \mathbb{C}^{\mathcal{Y}})_{\mathbb{1}_{\mathcal{L}(\mathbb{C}^{\mathcal{X}})} \otimes \Phi_M(\rho)} = I(X;Y)_p,$$

where the joint distribution of (X,Y) is given by (7.1) and $N(x,y) = \mathrm{tr}[M_y \Phi_x]$.

Of course, we can repeat the argument starting from any probability distribution p over $\mathcal{X}^n$ and using $\Phi^{\otimes n}$ instead of Φ and a measurement $M \subseteq \mathcal{O}_{\geq 0}(H^{\otimes n})$, so that

$$\mathcal{C}_\chi(\Phi^{\otimes n}) \geq \chi(\Phi^{\otimes n})_\rho \geq I(X^n; Y^n)_p.$$

Given any code (W, M), consider as in the classical case a uniform distribution over the m codewords (of length n), and let p denote its law, thus inducing random variables X^n (over the codewords) and Y^n (over $\{0, 1, \ldots, m\}$). Arguing as in the classical case, we use Fano's inequality to obtain

$$\bar{p}_e(W, M)) \geq 1 - \frac{I(X^n; Y^n)_p + 1}{\log m} \geq 1 - \frac{\mathcal{C}_\chi(\Phi^{\otimes n}) + 1}{\log m}.$$

Letting $m = 2^{nr}$, we conclude that any rate r such that

$$r > \liminf_{n \to \infty} \frac{\mathcal{C}_\chi(\Phi^{\otimes n})}{n}$$

is not admissible, hence

$$\mathcal{C}(\Phi) \leq \liminf_{n \to \infty} \frac{\mathcal{C}_\chi(\Phi^{\otimes n})}{n}.$$

However, we can prove that $\mathcal{C}_\chi(\Phi^{\otimes n}) = n\mathcal{C}_\chi(\Phi)$ is additive arguing as in the classical case (we used only concavity and tensorization properties of the entropy, which hold in the quantum case as well), hence the lower bound follows.

Remark 7.3. One has in general the inequality

$$\mathcal{C}_\chi(\Phi) \geq \sup_{M,p} \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x)\mathrm{tr}[M_y\Phi_x] \log\left(\frac{\mathrm{tr}[M_y\Phi_x]}{p(x)\sum_{x'\in\mathcal{X}} p(x')\mathrm{tr}[M_y\Phi_{x'}]}\right),$$

where maximization is all probability distributions p over the alphabet $\mathcal{X}$ and a non-sharp measurements $M = (M_y)_{y\in\mathcal{Y}}$. However, it is known that in general it cane be *strict*. Indeed, one could prove that the right hand side coincides with the operational channel capacity of Φ when Bob is restricted to only use measurements of product type

$$M_j = M_{j_1} \otimes M_{j_2} \otimes \ldots \otimes M_{j_n}.$$

This could be seen as a (dual) manifestation of entanglement, since even if Alice's messages are presented to Bob as product states, hence separable, using general non-product observables can be an advantage for Bob.

Direct statement. As in the classical case, the strategy is to find a suitable *random* codebook W and rely on the law of large numbers to suitably approximate a product state with a "uniform" state over a "typical" subspace.

Let us give the following definition (due to Schumacher and Jozsa in the quantum setting). Given a state $\rho \in \mathcal{S}(H)$, $n \geq 1$ and $\delta > 0$, its δ-typical subspace on $H^{\otimes n}$ consists of the span of the eigenvectors of $\rho^{\otimes n}$ with eigenvalues λ such that

$$2^{-nS(p)-n\delta} \leq \lambda \leq 2^{-nS(p)+n\delta}.$$

Using functional calculus, we may rewrite the orthogonal projector over the δ-typical subspace of ρ as

$$P^{\delta,n} = \mathbb{1}_{\left\{\left|\frac{1}{n}\sum_{i=1}^n \log\rho_i - S(p)\right| < \delta\right\}},$$

where $\rho^i = \mathbb{1}_{H^{\otimes(i-1)}} \otimes \rho \otimes \mathbb{1}_{H^{\otimes(n-1)}}$ are the copies (on each different factor, hence all commuting) of the state ρ.

Then, the *asymptotic equipartition property* states that, for every δ, $\varepsilon > 0$,

(1) For every n, the dimension of the δ-typical subspace of ρ is bounded from above
$$\mathrm{tr}[P^{\delta,n}] \leq 2^{n(S(p)+\delta)},$$
(2) For n sufficiently large, the contribution of vectors δ-typical can be made small, i.e.,
$$\mathrm{tr}[(1 - P^{\delta,n})\rho^{\otimes n}] \leq \varepsilon.$$
(3) For n sufficiently large, the dimension of the δ-typical subspace of ρ is bounded from below
$$\mathrm{tr}[P^{\delta,n}] \geq (1-\varepsilon)2^{n(S(p)-\delta)},$$

The proof of this properties is an application of Markov inequality.

We also need a notion of *conditionally typical subspace* associated to the classical to quantum channel $(\Phi_x)_{x\in\mathcal{X}}$. Given a probability distribution p over $\mathcal{X}$, write

$$S(H|\mathbb{C}^{\mathcal{X}})_\rho = \sum_{x\in\mathcal{X}} p(x)S(\Phi_x),$$

for the the quantum conditional entropy of the state ρ defined as in (7.9). For every $n \geq 1$, $\delta > 0$ and $w = (x_i)_{i=1}^n \in \mathcal{X}^n$, define the *conditionally typical subspace* of Φ given w (and p) as the linear span of the eigenvectors of $\Phi_w = \otimes_{i=1}^n \Phi_{x_i}$ whose eigenvalues λ satisfy

$$2^{-nS(H|\mathbb{C}^{\mathcal{X}})_\rho - n\delta} \leq \lambda \leq 2^{-nS(H|\mathbb{C}^{\mathcal{X}})_\rho + n\delta},$$

Again, via functional calculus, we may rewrite the orthogonal projector over the δ-typical subspace of Φ given w (and p) as

$$P_w^{\delta,n} = \mathbb{1}_{\left\{\left|\frac{1}{n}\sum_{i=1}^n \log \Phi_{x_i} - S(H|\mathbb{C}^{\mathcal{X}})_\rho\right| < \delta\right\}}.$$

We have the following properties:

(1) For every n, and w, we have

$$P_w^{\delta,n} \leq 2^{n(S(H|\mathbb{C}^{\mathcal{X}})_\rho + \delta)} \Phi_w,$$

(2) For n sufficiently large, and $\varepsilon > 0$,

$$\mathbb{E}\left[\mathrm{tr}[(1 - P_W^{\delta,n})\Phi_W]\right] \leq \varepsilon,$$

where $W = (X_i)_{i=1}^n$ are i.i.d. with common distribution p.

We are now in a position to describe the construction of the quantum decision rule associated to a codebook $(w_j)_{j=1}^m$ consisting of m words of length n. The construction will ultimately be performed using a probability distribution p over $\mathcal{X}$ maximizing $\chi(\Phi)_p$, but we may assume for the moment that p is fixed but general. Let us also fix δ and n in what follows, so we avoid to write them in the typical subspace projectors. For $j \in \{1, \ldots, m\}$, we would like to define

$$M_j = P_{w_j} P,$$

where P denotes the projector on the δ typical subspace associated to

$$\rho_H = \sum_{x \in \mathcal{X}} p_x \Phi_x,$$

and P_{w_j} denotes the projector on the δ-*conditionally typical subspace* of Φ given w_j (and ρ). Intuitively, this corresponds to the following decision rule by Bob: first he observes whether the state is a typical output; given an affirmative answer, he observes the conditional typical subspace to which the state belongs. However, this definition would not give rise to positive operators, hence we may correct the definition by squaring, i.e., letting

$$M_j = (P_{w_j} P)^* P_{w_j} P = P P_{w_j} P_{w_j} P = P P_{w_j} P.$$

Still, this is not sufficient, since we need to ensure that $\sum_{j=1}^m M_j = \mathbb{1}_{H^{\otimes n}}$, hence the definition is

$$M_j = A^{-1/2} P P_{w_j} P A^{-1/2} = (P_{w_j} P A^{-1/2})^* P_{w_j} P A^{-1/2},$$

where

$$A = \sum_{j=1}^m P P_{w_j} P,$$

and the inverse $A^{-1} = A^{-1} \mathbb{1}_{A>0}$ is actually a pseudo-inverse instead, defined to be 0 on the kernel of A. We also introduce $M_0 = \mathbb{1}_{A=0}$ as the orthogonal projection on the kernel of A, so that $\sum_{j=0}^m M_j = \mathbb{1}_{H^{\otimes n}}$ is satisfied.

After some estimations that we do not report here for brevity, see [Hol19, section 5.6] for more details, one obtains the upper bound

$$\bar{p}_e(W, M) \leq \frac{1}{m} \sum_{j=1}^m 4\mathrm{tr}[\Phi_{w_j}(1 - P)] + 4\mathrm{tr}[\Phi_{w_j}(1 - P_{w_j})] + \sum_{i \neq j} \mathrm{tr}[P \Phi_{w_j} P P_{w_i}].$$

Let us see how to conclude from this inequality. Assume that each w_j, $j = 1, \ldots, m$ is sampled as i.i.d. words with letter distributions given by p (each independently of each other), so that for every sampled letter, we have

$$\mathbb{E}\left[\Phi_x\right] = \sum_{x \in \mathcal{X}} p(x) \Phi_x = \rho_H.$$

Then, taking expectation in the inequality above and using independence, we obtain

$$\mathbb{E}\left[\bar{p}_e(W,M)\right] \leq 4\mathrm{tr}[\rho_H^{\otimes n}(1-P)] + 4\mathbb{E}\left[\mathrm{tr}[\Phi_w(1-P_w)]\right] + (m-1)\mathrm{tr}[P\rho_H^{\otimes n}P\mathbb{E}\left[P_w\right]].$$

For n large enough, we use the properties of projectors on typical subspaces, so that

$$\mathrm{tr}[\rho_H^{\otimes n}(1-P)] + \mathbb{E}\left[\mathrm{tr}[\Phi_w(1-P_w)]\right] \leq 2\varepsilon,$$

and, by definition of typical subspace,

$$P\rho_H^{\otimes n}P \leq 2^{-nS(H)_{\rho_H}+n\delta}\mathbb{1}_{H^{\otimes n}},$$

so that

$$\begin{aligned}\mathrm{tr}[P\rho_H^{\otimes n}P\mathbb{E}\left[P_w\right]] &\leq 2^{-nS(H)_{\rho_H}+n\delta}\mathbb{E}\left[\mathrm{tr}[P_w]\right] \\ &\leq 2^{-nS(H)_{\rho_H}+nS(H|\mathbb{C}^{\mathcal{X}})_{\rho}+2\delta n} \\ &= 2^{-nI(H;\mathbb{C}^{\mathcal{X}})_{\rho}+2\delta n}.\end{aligned}$$

Choosing p such that

$$I(H;\mathbb{C}^{\mathcal{X}})_{\rho} = \mathcal{C}_{\chi}(\Phi),$$

we obtain that, for every $r < \mathcal{C}_{\chi}(\Phi)$, letting $m = 2^{nr}$, we obtain

$$\mathbb{E}\left[\bar{p}_e(W,M)\right] \leq 8\varepsilon + 2^{n(r-\mathcal{C}_{\chi}(\Phi))+2\delta n}$$

which is infinitesimal, provided that δ is chosen small enough and then $\varepsilon \to 0$.

7.3. **Exercises.**

Exercise 7.1. Following the suggestion, complete the proof of (7.2).

Exercise 7.2. Prove the asymptotic equipartition properties.

References

[AF01] R. Alicki and M. Fannes. "Quantum dynamical systems". In: (2001) (cit. on p. 2).

[Agr13] J. Agredo. "A Wasserstein-type distance to measure deviation from equilibrium of quantum Markov semigroups". In: *Open Systems & Information Dynamics* 20.02 (2013), p. 1350009 (cit. on p. 42).

[BEŻ22] R. Bistroń, M. Eckstein, and K. Życzkowski. "Monotonicity of the quantum 2-Wasserstein distance". In: *arXiv preprint arXiv:2204.07405* (2022) (cit. on p. 43).

[BŻ17] I. Bengtsson and K. Życzkowski. *Geometry of quantum states: an introduction to quantum entanglement*. Cambridge university press, 2017 (cit. on p. 2).

[Car10] E. Carlen. "Trace inequalities and quantum entropy: an introductory course". In: *Entropy and the quantum* 529 (2010), pp. 73–140 (cit. on p. 31).

[Car22] E. A. Carlen. "On some convexity and monotonicity inequalities of Elliott Lieb". In: *arXiv preprint arXiv:2202.03591* (2022) (cit. on p. 32).

[CL92] A. Connes and J. Lott. "The metric aspect of noncommutative geometry". In: *New symmetry principles in quantum field theory*. Springer, 1992, pp. 53–93 (cit. on pp. 42, 44).

[CM14] E. A. Carlen and J. Maas. "An analog of the 2-Wasserstein metric in non-commutative probability under which the fermionic Fokker–Planck equation is gradient flow for the entropy". In: *Communications in mathematical physics* 331.3 (2014), pp. 887–926 (cit. on p. 42).

[De +21] G. De Palma, M. Marvian, D. Trevisan, and S. Lloyd. "The quantum Wasserstein distance of order 1". In: *IEEE Transactions on Information Theory* 67.10 (2021), pp. 6627–6643 (cit. on pp. 42, 44, 55).

[DT21] G. De Palma and D. Trevisan. "Quantum optimal transport with quantum channels". In: *Annales Henri Poincaré*. Vol. 22. 10. Springer. 2021, pp. 3199–3234 (cit. on p. 43).

[FS17] H. Fawzi and J. Saunderson. "Lieb's concavity theorem, matrix geometric means, and semidefinite optimization". In: *Linear Algebra and its Applications* 513 (2017), pp. 240–263 (cit. on p. 31).

[GMP16] F. Golse, C. Mouhot, and T. Paul. "On the mean field and classical limits of quantum mechanics". In: *Communications in Mathematical Physics* 343.1 (2016), pp. 165–205 (cit. on pp. 42, 43).

[Gos21] M. A. de Gosson. "Quantum Harmonic Analysis". In: *Quantum Harmonic Analysis*. De Gruyter, 2021 (cit. on pp. 15, 42).

[Hol19] A. S. Holevo. "Quantum systems, channels, information". In: *Quantum Systems, Channels, Information*. de Gruyter, 2019 (cit. on pp. 2, 40, 65).

[KW98] Z. Karol and S. Wojeciech. "The Monge distance between quantum states". In: (1998) (cit. on p. 42).

[Mey95] P. A. Meyer. *Quantum probability for probabilists*. Springer Science & Business Media, 1995 (cit. on p. 2).

[Mor19] V. Moretti. *Fundamental Mathematical Structures of Quantum Theory*. Springer, 2019 (cit. on p. 2).

[Naa13] P. Naaijkens. *Quantum spin systems on infinite lattices*. Springer, 2013 (cit. on p. 2).

[NC02] M. A. Nielsen and I. Chuang. *Quantum computation and quantum information*. 2002 (cit. on p. 2).

[PC+19] G. Peyré, M. Cuturi, et al. "Computational optimal transport: With applications to data science". In: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607 (cit. on p. 42).

[PL86] M. D. Plummer and L. Lovász. *Matching theory*. Elsevier, 1986 (cit. on p. 42).

[Sha48] C. E. Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3 (1948), pp. 379–423 (cit. on pp. 2, 46, 56).

[Vil09] C. Villani. *Optimal transport: old and new*. Vol. 338. Springer, 2009 (cit. on pp. 41, 42).

[Wil11] M. M. Wilde. "From classical to quantum Shannon theory". In: *arXiv preprint arXiv:1106.1445* (2011) (cit. on p. 2).

D.T.: Dipartimento di Matematica, Università degli Studi di Pisa, 56125 Pisa, Italy
Email address: dario.trevisan@unipi.it