Entropy and Large Deviations
Part 2: Quantum Theory

Lecture Notes, Master's course, Fall 2025

Klaas Landsman

Institute for Mathematics, Astrophysics, and Particle Physics
Radboud Center for Natural Philosophy
Radboud University, Nijmegen, The Netherlands
landsman@math.ru.nl

Contents

1 The probabilistic structure of classical physics

As a goal in itself, and also in preparation for quantum mechanics (which is a probabilistic theory by its very nature), in this section we discuss the probabilistic structure of classical physics. In Newtoninan physics, one assumes that the state of the system is exactly known; if M is the phase space of the system, henceforth called X (which is the usual notation in probability theory, whereas M is the usual notation for manifolds), then an 'exact' state is just a point $x \in X$. We will soon formalize this, but the idea is that if x is known, then everything can be predicted with certainty, at least in principle (some would say: everything is determined by x), like the values of all observables $f : X \to \mathbb{R}$ (which are given by $f(x)$) and the future (or even past) time evolution of the system. However, if $X = \mathbb{R}^{6N}$ with $N \sim 10^{23}$, then it is not feasible to assume that $x \in X$ is known exactly. This led to (classical) *statistical mechanics*, which was developed in the 19th century mainly by Maxwell, Boltzmann, and Gibbs. Furthermore, even if X is relatively small, the existence of *deterministic chaos* (first studied by Poincaré and others in connection with the stability of the solar system, in which context it was probably familiar already to Newton) indicated that in many dynamical systems x should really be known with almost infinite precision in order to make accurate predictions, which is unfeasible in practice.

Both cases suggested replacing 'exact' or 'pure' states $x \in X$ by *probability measures* on X, which incorporate the confession that although we do not know the exact state, we have at least some insight into the likelihood what it is.

As a warm-up we first assume that X is a ***finite set***, playing the role of the configuration space of some physical system, or, equivalently (as we shall see), of its pure state space. In general, X should be thought of as the phase space, but as long as X is discrete, the phase space coincides with the (intuitively more appealing) configuration space. Finite sets are not at all boring. For example, if X is supposed to describe the possible configurations of N bits (numbered $0,\dots,N-1$), then $X = \underline{2}^{\underline{N}}$. Here $\underline{N} = \{0,1,\dots,N-1\}$ (so that, in particular, $\underline{2} = \{0,1\}$), and, for arbitrary sets Y and Z, the set Y^Z consists of all functions $x : Z \to Y$. Of course, by binary coding the set $\underline{2}^{\underline{N}}$ may be identified with the set $\{0,1,\dots,2^N-1\}$, i.e., for $x \in \underline{2}^{\underline{N}}$ the binary number $x(N-1)\cdots x(0)$ is identified with its decimal counterpart $\sum_{k=0}^{N-1} x(k)2^k$. S

In any case, as we will see, X defines both the states and the observables, from the totality of each of which X may in turn be reconstructed. Furthermore, the states and the observables define each other even without relying on the underlying set X.

Definition 1.1 *Recall that the power set $\mathscr{P}(X)$ of X is the set of all subsets of X.*

1. *An* **event** *is a subset $U \subseteq X$, i.e., $U \in \mathscr{P}(X)$.*
2. *A* **probability distribution** *on X is a function $p : X \to [0,1]$ such that $\sum_x p(x) = 1$.*
3. *A* **probability measure** *on X is a function $\mu : \mathscr{P}(X) \to [0,1]$ such that $\mu(X) = 1$ and $\mu(U \cup V) = \mu(U) + \mu(V)$ whenever $U \cap V = \emptyset$.*
4. *A* **random variable** *on X is a function $f : X \to \mathbb{R}$.*
5. *The* **spectrum** *of a random variable f is the subset $\sigma(f) = \{f(x) \mid x \in X\}$ of $\mathbb{R}$.*

Probability distributions p and probability measures μ on finite sets come down to the same thing, as the former gives rise to the latter by

$$\mu(U) = \sum_{x \in U} p(x), \tag{1.1}$$

whilst *vice versa* one has

$$p(x) = \mu(\{x\}). \tag{1.2}$$

It is a matter of convenience which one is used.

A special class of probability measures stands out: each $y \in X$ defines a probability distribution p_y by $p_y(x) = \delta_{xy}$, or explicitly,

$$p_y(x) = 1 \text{ if } x = y; \tag{1.3}$$
$$p_y(x) = 0 \text{ if } x \neq y. \tag{1.4}$$

Equivalently, for the corresponding probability measure one has

$$\mu_y(U) = 1 \text{ if } y \in U; \tag{1.5}$$
$$\mu_y(U) = 0 \text{ if } y \notin U. \tag{1.6}$$

The single most important construction in probability theory, then, is as follows.

Theorem 1.2 *A probability measure μ on X and a random variable $f : X \to \mathbb{R}$ jointly yield a probability measure μ_f on the spectrum $\sigma(f)$ by means of*

$$\mu_f(\Delta) = \mu(f \in \Delta), \tag{1.7}$$

where $\Delta \subseteq \sigma(f)$ and $f \in \Delta$ denotes the event $\{x \in X \mid f(x) \in \Delta\}$ in X.

In terms of the corresponding probability distribution p on X, one has

$$p_f(\lambda) = \sum_{x \in X | f(x) = \lambda} p(x). \tag{1.8}$$

Instead of $f \in \Delta$, the notation $f^{-1}(\Delta)$ is often used. The proof is an exercise.

Given a finite set X, we may form the set $C(X)$ of all complex-valued functions on X, enriched with the structure of a complex vector space under pointwise operations:

$$(\lambda \cdot f)(x) = \lambda f(x) \ (\lambda \in \mathbb{C}); \tag{1.9}$$
$$(f + g)(x) = f(x) + g(x). \tag{1.10}$$

We use the notation $C(X)$ with some foresight, anticipating the case where X is no longer finite, but in any case, since for the moment it is, every function is continuous. Moreover, the vector space structure on $C(X)$ may be extended to that of a commutative algebra (where, by convention, all our algebras are associative and are defined over the complex scalars) by defining multiplication pointwisely, too:

$$(f \cdot g)(x) = f(x)g(x). \tag{1.11}$$

Note that this algebra has a unit 1_X, i.e., the function identically equal to 1. We also equip $C(X)$ with an ***involution***, which on an arbitrary (not necessarily commutative) algebra A is defined as an anti-linear anti-homomorphism, i.e., a real-linear map $* : A \to A$ (written $a \mapsto a^*$) that satisfies $(\lambda a)^* = \overline{\lambda} a^*$ and $(ab)^* = b^* a^*$. In our case $A = C(X)$, which is commutative, the latter property simply becomes $(fg)^* = f^* g^*$. In any case, we define this involution by pointwise complex conjugation, i.e.,

$$f^*(x) = \overline{f(x)}. \tag{1.12}$$

Finally, $C(X)$ has a natural ***norm***

$$\|f\|_\infty = \sup_{x \in X} \{|f(x)|\}. \tag{1.13}$$

These structures turn $C(X)$ into a ***commutative C*-algebra***.

Definition 1.3 *A C*-algebra is an associative algebra (over $\mathbb{C}$) equipped with an involution as well as a norm in which A is complete (i.e., a Banach space), such that algebra, involution, and norm are related by the axioms*

$$\|ab\| \leq \|a\|\,\|b\|; \tag{1.14}$$

$$\|a^*a\| = \|a\|^2. \tag{1.15}$$

If X is compact (but not necessarily finite), then $C(X)$ is a commutative C*-algebra. if X is locally compact but not compact, one should take the space $C_0(X)$ of all continuous functions $f : X \to \mathbb{C}$ that vanish at infinity (i.e., for any $\varepsilon > 0$ the set $\{x \in X \mid |f(x)| \geq \varepsilon\}$ is compact). It is of fundamental importance that $C(X)$ and $C_0(X)$ are *commutative*. The elements of $C_0(X)$ are called ***observables***. We already noted that $C(X)$ has a unit (as an algebra), namely the function 1_X; this is still the case if X is compact, but $C_0(X)$ has no unit.

Definition 1.4 *A* state *on a C*-algebra A with unit is a linear map $\omega : A \to \mathbb{C}$ that is* positive*, i.e.,*

$$\omega(a^*a) \geq 0 \tag{1.16}$$

for each $a \in A$, as well as normalized *in that*

$$\omega(1_A) = 1. \tag{1.17}$$

If A has no unit, like $C_0(X)$, then (1.17) should be replaced the condition

$$\|\omega\| = 1, \tag{1.18}$$

where $\|\cdot\|$ is the usual norm on the Banach dual A^*, i.e.,

$$\|\omega\| = \sup\{|\omega(a)|, a \in A, \|a\| \leq 1\|. \tag{1.19}$$

In fact, it can be shown that if (1.16) holds, then ω is bounded, and if in addition A has a unit 1_A, then ω satisfies $\|\omega\| = \omega(1_A)$. Therefore, in the presence of a unit and condition (1.16), the normalization conditions (1.18) and (1.19) are equivalent. A special case of this which is within our reach is:

Proposition 1.5 *Let X be a compact Hausdorff space. If a linear map $\varphi : C(X) \to \mathbb{C}$ is positive, then it is bounded, with norm $\|\varphi\| = \varphi(1_X)$.*

The proof is an exercise.

If we specialize Definition 1.4 to the case $A = C(X)$, where X is finite or compact, we note that if $f = a^*a$, then $f(x) = |a(x)|^2$, so that $f(x) \geq 0$ for each x. Conversely, if $f(x) \geq 0$ for each x, then we have $f = a^*a$ for $a = \sqrt{f}$. Hence we have:

Lemma 1.6 *A state on $C(X)$ is a complex-linear map $\omega : C(X) \to \mathbb{C}$ that satisfies:*

1. *$\omega(f) \geq 0$ for each $f \geq 0$ (i.e., $f(x) \geq 0$ for each x);*
2. *$\omega(1_X) = 1$.*

Proposition 1.7 *For finite X there is a bijective correspondence between states ω on $C(X)$ and probability measures μ on X, given by*

$$\omega(f) = \sum_{x \in X} p(x) f(x) = \sum_{\lambda \in \sigma(f)} \mu(f = \lambda) \cdot \lambda; \tag{1.20}$$

$$\mu(U) = \omega(1_U), \tag{1.21}$$

where $p(x)$ is given by (1.2) *and 1_U is the characteristic function of U (defined by $1_U(x) = 1$ if $x \in U$ and $1_U(x) = 0$ if $x \notin U$).*

The proof is an exercise. The ***state space*** $S(C(X))$ of the algebra $C(X)$ is the set of all states on $C(X)$; by the theorem, this is essentially the same as the set $Pr(X)$ of all probability distributions p on X. These are examples of ***compact convex sets***.

Definition 1.8 *A subset K of a real vector space V is called* **convex** *if whenever $v,w \in K$ and $t \in (0,1)$, one has $tv+(1-t)w \in K$. This is equivalent with the following property: if $t_1,\ldots,t_n$ are numbers in $[0,1]$ such that $\sum_i t_i = 1$ and if $v_1,\ldots,v_n$ are in K, then $\sum_{i=1}^n t_i \cdot v_i$ is in K.*

Thus any linear subspace of V is trivially convex, but the interesting convex sets are *compact* in the topology inherited from V (provided it has one, always assumed Hausdorff). We will usually have $V = W^*$, the space of linear functionals from some other (finite-dimensional) real vector space W to $\mathbb{R}$, and the topology on V is the so-called w^*-topology, defined by saying that $v_n \to v$ iff $v_n(w) \to v(w)$ for each $w \in W$. Then K is compact in V iff each infinite sequence in K has a convergent subsequence. If V is finite-dimensional, as is the case in this section, then compact just means closed and bounded.

Definition 1.9 *The* **(extreme) boundary** *$\partial_e K$ of a convex set K consists of all $v \in K$ satisfying the condition:*

$$\text{if } v = tw + (1-t)x \text{ for certain } w,x \in K \text{ and } t \in (0,1), \text{ then } v = w = x.$$

Elements $v \in \partial_e K$ of the boundary are called **extremal points** *of K.*

The main example of interest to us is:

Theorem 1.10 *For any locally compact Hausdorff space X, the boundary $\partial_e Pr(X)$ of the convex set $Pr(X)$ of all probability measures on X is isomorphic to X through $\delta_x \leftrightarrow x$, where $\delta_x : C_0(X) \to \mathbb{C}$ is the evaluation map*

$$\delta_x(f) = f(x). \tag{1.22}$$

The proof is an exercise. The probability distribution p corresponding to the state δ_x is δ_x.

In finite dimension, there are two kinds of compact convex sets: smooth ones, like the (closed) unit disc in $\mathbb{R}^2$ or the (closed) unit ball in $\mathbb{R}^3$ (which we will encounter in quantum mechanics), and ***convex polytopes***, which by definition are convex hulls of finitely many points (that is, the smallest convex sets containing these points). Examples of convex polytopes are ***regular polyhedra***. These were classified (up to affine isomorphism, i.e., bijections preserving convex sums) by Schläfli in 1852, who showed that the only possibilities are:

- The *simplices* $\Delta_n = \{x \in \mathbb{R}^{n+1} \mid x_i \geq 0, \sum_i x_i = 1\}$, $n \geq 1$;
- The *cubes* $Q_n = \{x \in \mathbb{R}^n \mid -1 \leq x_i \leq 1\}$, $n > 1$;
- The *cross-polytopes* $O_n = \{x \in \mathbb{R}^n \mid \sum_i |x_i| \leq 1\}$, $n > 1$;
- The countably many *regular polygons* in $\mathbb{R}^2$ (which include Q_2, O_2, Δ_2);
- The five *platonic solids* in $\mathbb{R}^3$ (which include Q_3, O_3, Δ_3);
- The six *regular polychora* in $\mathbb{R}^4$ (which include Q_4, O_4, Δ_4).

Here Δ_n is affinely homeomorphic to the convex hull of $n+1$ *linearly independent* points, and this property uniquely defines it (up to affine isomorphism). It is almost tautological that the simplex Δ_n is the set $Pr(X_{n+1})$ of all probability distributions on a set X_{n+1} of cardinality $n+1$. In finite dimension, Theorem 1.10 may be rewritten as follows:

Theorem 1.11 *The boundary of the n-dimensional simplex Δ_n is given by*

$$\partial_e \Delta_n = \{\vec{e}_1, \ldots . \vec{e}_{n+1}\}, \tag{1.23}$$

where $(\vec{e}_1, \ldots . \vec{e}_{n+1})$ is the standard basis of $\mathbb{R}^{n+1}$ (i.e., $\vec{e}_1 = (1,0,\ldots,0)$, etc.).

It follows that $|\partial_e \Delta_n| = n+1$, i.e., the boundary of Δ_n has $n+1$ points. This is another way to single out the simplices among all regular polyhedra. The simplest example is $\Delta_1 \cong [0,1]$, so that $\partial_e \Delta_1 \cong \{0,1\}$. Note that $\partial_e(0,1) = \emptyset$, so that the boundary of a convex set may well be empty (another example is the *open* disc). This cannot happen if K is compact (in finite dimension this was proved by Caratheodory, whereas in general it follows from the Krein–Milman Theorem of functional analysis).

2 Finite-dimensional Hilbert space

The quantum analogue of a finite set X (in its role as a phase space in classical mechanics) is the finite-dimensional Hilbert space $\ell^2(X)$, by which we mean the vector space of functions $\psi : X \to \mathbb{C}$, equipped with the inner product

$$\langle \psi, \varphi \rangle = \sum_{x \in X} \overline{\psi(x)} \phi(x). \tag{2.1}$$

For finite X we have $X \cong \underline{n} = \{1, 2, \ldots, n\}$ as sets, where n is the cardinality of X, inducing the unitary isomorphism $\ell^2(\underline{n}) \cong \mathbb{C}^n$ of Hilbert space through the map $\psi \mapsto (\psi(1), \ldots, \psi(n))$, where $\mathbb{C}^n$ has the standard inner product

$$\langle w, z \rangle = \sum_i \overline{w}_i z_i. \tag{2.2}$$

In particular, $\delta_k \in \ell^2(\underline{n})$ is mapped to the k'th standard basis vector $\upsilon_k \equiv |k\rangle$ of $\mathbb{C}^n$.

If H is finite-dimensional, we may therefore assume that $H = \mathbb{C}^n$ and that $L(H)$, i.e., the algebra of all bounded linear maps $a : H \to H$, is just the algebra $M_n(\mathbb{C})$ of $n \times n$ matrices (if $\dim(H) < \infty$, linear maps are automatically continuous, and we often refrain from making the subtle difference between linear maps $a : \mathbb{C}^n \to \mathbb{C}^n$ and the matrices representing such maps, once a basis of $\mathbb{C}^n$ has been chosen). In any case, $L(H)$ or $M_n(\mathbb{C})$ is the quantum analogue of the algebra $C(X)$ in the previous section. Like $C(X)$, it is a C*-algebra (with unit): the involution on $M_n(\mathbb{C})$ is given by hermitian conjugation, i.e.,

$$(a^*)_{ij} = \overline{a_{ji}}, \tag{2.3}$$

and, more abstractly, the involution on $L(H)$ is the map $a \mapsto a^*$, where a^* is the unique operator such that

$$\langle a^* \varphi, \psi \rangle = \langle \varphi, a\psi \rangle, \tag{2.4}$$

for each $\varphi, \psi \in H$. The unit is simply the unit operator $1 \equiv 1_H$. i.e., $1_H(\psi) = \psi$. Finally, it goes without saying that the algebraic structure on $M_n(\mathbb{C})$ (or $L(H)$) is given by matrix (or operator) multiplication and addition, that is,

$$(\lambda \cdot a)\psi = \lambda(a\psi); \tag{2.5}$$

$$(a+b)\psi = a\psi + b\psi; \tag{2.6}$$

$$(ab)\psi = a(b\psi), \tag{2.7}$$

The ***spectrum*** $\sigma(a)$ of $a \in L(H)$ consists of all eigenvalues of a, i.e., $\lambda \in \sigma(a)$ iff there exists $\psi \neq 0$ such that $a\psi = \lambda\psi$.

The key to the probabilistic setting of quantum mechanics is given by the following quantum counterpart of of a classical probability measure.

Definition 2.1 *Let H be a finite-dimensional Hilbert space. A* **density operator** *is a positive operator ρ on H such that*

$$\mathrm{Tr}(\rho) = 1. \tag{2.8}$$

The set of all density operators on H is called $D(H)$. It is easily seen to be convex.

We recall the definition of the trace, and, afterwards, of positivity.

Lemma 2.2 *If (υ_i) and (υ'_i) are bases of H, then for any operator $a : H \to H$,*

$$\sum_i \langle \upsilon_i, a\upsilon_i \rangle = \sum_i \langle \upsilon'_i, a\upsilon'_i \rangle.$$

This lemma (which you can prove for yourself) allows us to define the ***trace*** of a by

$$\mathrm{Tr}(a) = \sum_i \langle v_i, a v_i \rangle, \tag{2.9}$$

where (v_i) is any basis of H. We obtain

$$\mathrm{Tr}(ab) = \sum_{i,j} \langle v_i, a v_j \rangle \langle v_j, b v_i \rangle = \sum_{i,j} \langle v_i, b v_j \rangle \langle v_j, a v_i \rangle = \mathrm{Tr}(ba). \tag{2.10}$$

If u is ***unitary*** (in that $uu^* = u^*u = 1$), then from either Lemma 2.2 or (2.10),

$$\mathrm{Tr}(uau^*) = \mathrm{Tr}(a). \tag{2.11}$$

Finally, if $a^* = a$, then we may use the spectral theorem

$$a = \sum_{\lambda \in \sigma(a)} \lambda \cdot e_\lambda; \tag{2.12}$$

$$1_H = \sum_{\lambda \in \sigma(a)} e_\lambda, \tag{2.13}$$

where

$$H_\lambda = \{ \psi \in H \mid a\psi = \lambda \psi \}, \tag{2.14}$$

is the eigenspace for $\lambda \in \sigma(a)$ and e_λ is the unique projection $H \to H$ with image H_λ. We may also write the spectral resolution of $a^* = a$ as

$$a = \sum_{i=1}^{\dim(H)} \lambda_i |v_i\rangle\langle v_i|; \tag{2.15}$$

$$1_H = \sum_{i=1}^{\dim(H)} |v_i\rangle\langle v_i|, \tag{2.16}$$

where λ_i is the eigenvalue corresponding to the eigenvector v_i (i.e., $av_i = \lambda_i v_i$), and the v_i form a basis of H consisting of eigenvectors of a. Taking the trace over the basis in (2.15) then yields

$$\mathrm{Tr}(a) = \sum_{i=1}^{\dim(H)} \lambda_i = \sum_{\lambda \in \sigma(a)} m_\lambda \cdot \lambda, \tag{2.17}$$

where $m_\lambda = \dim(H_\lambda)$ is the multiplicity of λ.

We say that an operator $a : H \to H$ is ***positive*** if $\langle \psi, a\psi \rangle \geq 0$ for arbitrary $\psi \in H$, in which case we write $a \geq 0$. Without proof we mention some facts about positivity:

Proposition 2.3 *The following condition on an operator $a : H \to H$ are equivalent:*

1. $\langle \psi, a\psi \rangle \geq 0$ *for arbitrary* $\psi \in H$;
2. $a^* = a$ *and* $\sigma(a) \subset \mathbb{R}^+$;
3. $a = c^2$ *for some hermitian operator* c;
4. $a = b^*b$ *for some operator* b.

Being positive, a density operator ρ is hermitian, so by (2.16), we have

$$\rho = \sum_i p_i |v_i\rangle\langle v_i|,\ p_i \geq 0,\ \sum_i p_i = 1, \tag{2.18}$$

where the (v_i) form an orthonormal set in H and $|v_i\rangle\langle v_i|$ is the (orthogonal) projection on the one-dimensional subspace $\mathbb{C}\cdot v_i$. Conversely, an operator of the form (2.18) is a density operator. A special class of density operators stands out:

- Each *unit vector* $\psi \in H$ defines a density operator

$$e_\psi = |\psi\rangle\langle\psi|, \tag{2.19}$$

  i.e., the (orthogonal) projection on the one-dimensional subspace $\mathbb{C}\cdot\psi$. A basis of eigenvectors of e_ψ consists of $v_0 = \psi$ itself, supplemented by any basis $(v_1, \ldots, v_{\dim(H)-1})$ of the orthogonal complement of $\mathbb{C}\cdot\psi$. The corresponding probabilities are evidently $p_0 = 1$ and $p_i = 0$ for all $i > 0$.

It makes good sense to copy Definition 1.6, *mutatis mutandis*:

Definition 2.4 *A* **state** *on $L(H)$ is a complex-linear map $\omega : L(H) \to \mathbb{C}$ satisfying:*

1. *$\omega(a) \geq 0$ for each positive $a \in L(H)$, i.e., for each $a \geq 0$ (***positivity***);*
2. *$\omega(1) = 1$ (***normalization***).*

Despite its easy proof, the following result is of fundamental importance.

Theorem 2.5 *If H is finite-dimensional, there is a bijective correspondence between states ω on $L(H)$ and density operators ρ on H, given by*

$$\omega(a) = \mathrm{Tr}(\rho a). \tag{2.20}$$

The proof is an exercise; for finite-dimensional H the identification (2.20) even works for any linear map $\omega : L(H) \to \mathbb{C}$, matching it with some $\rho \in L(H)$.

Definition 2.6 *The* **state space** *$S(L(H))$ is the set of all states $\omega : L(H) \to \mathbb{C}$, seen as a subspace of $L(H)^*$ (in the w^*-topology).*

The quantum analogue of Proposition 1.7 is as follows.

Corollary 2.7 *Let H be a finite-dimensional Hilbert space. The state space $S(L(H))$ is isomorphic as a compact convex set to the set $D(H)$ of density matrices on H.*

The case $H = \mathbb{C}^2$ provides a beautiful illustration of this theorem (see exercise).

Proposition 2.8 *The state space $S(M_2(\mathbb{C}))$ of the 2×2 matrices is isomorphic (as a compact convex set) to the closed unit ball $B^3 = \{(x,y,z) \in \mathbb{R}^3 \mid x^2+y^2+z^2 \leq 1\}$.*

On this isomorphism, the (extreme) boundary (cf. Definition 1.9)

$$\partial B^3 = S^2 = \{(x,y,z) \in \mathbb{R}^3 \mid x^2+y^2+z^2 = 1\} \tag{2.21}$$

corresponds to the set of all density matrices $\rho = e_\psi$, where $\psi \in \mathbb{C}^2$ with $\|\psi\| = 1$ (these are exactly the one-dimensional projections on $\mathbb{C}^2$).

The proof is an exercise; you may use the fact that any hermitian 2×2 matrix may be parametrized by $(t,x,y,z)\in\mathbb{R}^4$ as

$$\rho(t,x,y,z)=\tfrac{1}{2}\begin{pmatrix} t+z & x-iy \\ x+iy & t-z \end{pmatrix}. \tag{2.22}$$

In this example, the pure state space of $L(H)$ is by no means empty, and we will now see that also in general, the special density operators e_ψ in (2.19) to some extent play the role of the points $x\in X$. Let $\mathscr{P}(H)$ be the set of all projections on H, that is,

$$\mathscr{P}(H)=\{e\in L(H)\mid e^2=e^*=e\}, \tag{2.23}$$

and let $\mathscr{P}_1(H)\subset\mathscr{P}(H)$ the the subset of all *one-dimensional* projections, so we have $e\in\mathscr{P}_1(H)$ iff $e=e_\psi$, see (2.19), where $\psi\in H$ is a unit vector. This also means that $e\in\mathscr{P}_1(H)$ iff $e\in\mathscr{P}(H)$ and $\dim(eH)=\mathrm{Tr}\,(e)=1$.

Proposition 2.9 *A density operator ρ is an extremal point of the convex set $D(H)$ of all density operators on H iff $\rho=e_\psi$ for some unit vector $\psi\in H$.*

The proof is an exercise. Consequently,

$$P(L(H))=\mathscr{P}_1(H). \tag{2.24}$$

Denoting the state ω defined by the density operator $\rho=e_\psi$ via (2.20) by ω_ψ,

$$\omega_\psi=\langle\psi,a\psi\rangle; \tag{2.25}$$

to see this, take the trace over some basis that contains ψ.

3 The Born measure and the Born rule of quantum mechanics

The ***Born rule*** provides a link between the mathematical formalism of quantum theory and experiment, and as such is almost single-handedly responsible for practically all predictions of quantum physics. On a par with the Heisenberg uncertainty relations the Born rule is often seen as a turning point where indeterminism entered fundamental physics. For these two reasons, its importance cannot be overestimated.

A simple version of the Born rule was first stated by Max Born (1882-1970) in the context of scattering theory, following a slightly earlier paper in which he famously omitted the absolute value squared signs (though he corrected this is a footnote added in proof). The modern formulation below is due to von Neumann.

We first construct the Born measure, which is a successful attempt to adapt Theorem 1.2 to quantum mechanics. We initially assume that H is finite-dimensional. For given $a=a^*\in L(H)$ and $\Delta\subset\sigma(a)$, we use the notation

$$e_\Delta \;=\; \sum_{\lambda\in\Delta} e_\lambda, \tag{3.1}$$

where e_λ is the projection onto the eigenspace $H_\lambda=\{\psi\in H\mid a\psi=\lambda\psi\}$, see (2.14).

Theorem 3.1 *A density operator ρ on H and a hermitian operator $a:H\to H$ jointly yield a probability distribution p_a on the spectrum $\sigma(a)$ by*

$$p_a(\lambda)=\mathrm{Tr}\,(\rho e_\lambda). \tag{3.2}$$

The associated probability measure μ_a, called the **Born measure** *associated to the observable a and the density operator ρ, is given at $\Delta \subseteq \sigma(a)$ by (cf.* (3.1)*)*

$$\mu_a(\Delta) = \mathrm{Tr}(\rho e_\Delta). \tag{3.3}$$

In terms of the state ω associated to ρ, cf. (2.20)*, we simply have*

$$p_a(\lambda) = \omega(e_\lambda); \tag{3.4}$$
$$\mu_a(\Delta) = \omega(e_\Delta). \tag{3.5}$$

Note that this relationship between p_a and μ_a is the same as the one in (1.1) - (1.2): from the point of view of Theorem (1.7), we have

$$\mu_a(\Delta) = \omega(e_\Delta) = \omega\left(\sum_{\lambda\in\Delta} e_\lambda\right) = \sum_{\lambda\in\Delta} \omega(e_\lambda) = \sum_{\lambda\in\Delta} p_a(\lambda). \tag{3.6}$$

It is an exercise to verify that p_a is indeed a probability distribution on $\sigma(a)$, or, equivalently, that μ_a is a probability measure on $\sigma(a)$. Expectation values may then be rewritten as follows:

$$\omega(a) = \sum_{\lambda\in\sigma(a)} \lambda \cdot p_a(\lambda). \tag{3.7}$$

If $\rho = e_\psi$, writing p_a^ψ for the associated probability, (3.2) yields

$$p_a^\psi(\lambda) = \langle \psi, e_\lambda \psi\rangle = \|e_\lambda \psi\|^2. \tag{3.8}$$

If in addition $\lambda \in \sigma(a)$ is non-degenerate, so that $e_\lambda = |\upsilon_\lambda\rangle\langle\upsilon_\lambda|$ for some unit vector υ_λ with $a\upsilon_\lambda = \lambda\upsilon_\lambda$, then the Born rule (3.3) assumes its original form

$$p_a^\psi(\lambda) = |\langle \psi, \upsilon_\lambda\rangle|^2. \tag{3.9}$$

What does the Born measure mean physically? First, quantum mechanics stipulates that measurements of some observable a always have outcomes in its spectrum $\sigma(a)$. Following up on this, the ***Born rule states that:***

If an observable a is measured in a state ω, then the probability that the outcome lies in $\Delta \subset \sigma(a)$ equals $\mu_a(\Delta)$, where μ_a is the Born measure defined by a and ω. In particular, the probability that the outcome is $\lambda \in \sigma(a)$ is equal to $p_a(\lambda)$.

4 Tensor products

In the context of quantum mechanics, the tensor product of Hilbert spaces replaces the cartesian product of sets. Notation: for finite A the Hilbert space $\ell^2(A)$ consists of all functions $f : A \to \mathbb{C}$ with inner product $\langle f, g\rangle = \sum_{a\in A} \overline{f(a)} g(a)$. Likewise $\ell^2(B)$ etc. Clearly, any finite-dimensional Hilbert space is isomorphic (that is, unitarily equivalent) to $\ell^2(A)$ for some finite ste A, whose cardinality equals the dimension of $\ell^2(A)$: indeed, the functions $|a\rangle \equiv \delta_a$, where $a \in A$, form an orthonormal basis, where $\delta_a(a') = \delta_{aa'}$ as usual. In this case, we define the tensor product as

$$\ell^2(A) \otimes \ell^2(B) := \ell^2(A \times B). \tag{4.1}$$

For $f \in \ell^2(A)$ and $g \in \ell^2(B)$ we also define $f \otimes g \in \ell^2(A) \otimes \ell^2(B)$ to be the function

$$f \otimes g(a,b) = f(a)g(b). \tag{4.2}$$

As a special case, we have $\delta_a \otimes \delta_b$, which we also write as $|a\rangle|b\rangle$ or $|ab\rangle$, which, as $a \in A$ and $b \in B$, obviously forms a basis of $\ell^2(A) \otimes \ell^2(B)$. More generally, from (4.2) we infer that the map

$$p : \ell^2(A) \times \ell^2(B) \to \ell^2(A) \otimes \ell^2(B); \qquad p(f,g) = f \otimes g, \tag{4.3}$$

is bilinear (that is, linear in each of its two arguments). In particular,

$$(f_1 + f_2) \otimes g = f_1 \otimes g + f_2 \otimes g; \tag{4.4}$$

$$f \otimes (g_1 + g_2) = f \otimes g_1 + f \otimes g_2; \tag{4.5}$$

$$(\lambda f) \otimes g = f \otimes (\lambda g) = \lambda (f \otimes g), \tag{4.6}$$

where $\lambda \in \mathbb{C}$. General elements of $\ell^2(A) \otimes \ell^2(B)$, i.e., arbitrary functions of (a,b), are then given by linear combinations of the 'elementary tensors' $f \otimes g$, since every function of (a,b) is a finite sum of products like fg with $f \in \ell^2(A)$ and $g \in \ell^2(B)$. If some finite-dimensional Hilbert spaces H_A and H_B are not given as $\ell^2(A)$ and $\ell^2(B)$, but are assumed to haves bases $(|a\rangle)_{a \in A}$ and $(|b\rangle)_{b \in B}$, respectively, then we may define $H_A \otimes H_B$ as the Hilbert space with basis $(|ab\rangle)_{a \in A, b \in B}$, so that

$$\langle ab, a'b' \rangle = \delta_{aa'} \delta_{bb'}. \tag{4.7}$$

By construction, we then have

$$\dim(H_A \otimes H_B) = \dim(H_A) \cdot \dim(H_B). \tag{4.8}$$

For given $f \in H_A$ as $f = \sum_a f_a |a\rangle$ and $g \in H_B$ as $g = \sum_b g_b |b\rangle$, we then define $f \otimes g \in H_A \otimes H_B$ as

$$f \otimes g := \sum_{a,b} f_a g_b |ab\rangle, \tag{4.9}$$

from which one can easily check the rules (4.4) - (4.6). Once again, $H_A \otimes H_B$ consists of finite linear combinations of the elementary tensors $f \otimes g$, with inner product:[1]

$$\left\langle \sum_i \lambda_i f_i' \otimes g_i', \sum_j \mu_j f_j \otimes g_j \right\rangle = \sum_{i,j} \overline{\lambda}_i \mu_j \langle f_i', f_j \rangle_{H_A} \cdot \langle g_i', g_j \rangle_{H_B}. \tag{4.10}$$

The 'ultimate' but very abstract definition of a tensor product of vector spaces (of which Hilbert spaces are of course a special case with more structure) is given by the following proposition:

Proposition 4.1 *Let H_A and H_B be finite-dimensional (complex) vector spaces. There is a vector space $H_A \otimes H_B$ and a bilinear map $p : H_A \times H_B \to H_A \otimes H_B$, such that for any vector space H_C and any bilinear map $\beta : H_A \times H_B \to H_C$, there is a unique linear map $\beta' : H_A \otimes H_B \to H_C$ such that*

$$\beta = \beta' \circ p. \tag{4.11}$$

In other words, the following diagram commutes:

$$\begin{array}{ccc} H_A \times H_B & \xrightarrow{\ p\ } & H_A \otimes H_B \\ & \searrow^{\forall \beta} & \downarrow \exists ! \beta' \\ & & H_C \end{array} \tag{4.12}$$

[1] In fact, since the representation of some element Ψ of $H_A \otimes H_B$ as $\Psi = \sum_i \lambda_i f_i \otimes g_i$ is not unique one should verify that this is well defined. This can be done most easily be relating H_A to $\ell^2(A)$, etc. See the lecture notes *Introduction to Mathematical Physics* from 2018 for a direct proof.

This universal property implies that $H_A \otimes H_B$ is unique up to isomorphism (exercise). In our examples, the map p is just given by $p(f,g) = f \otimes g$, cf. (4.3), so that, given this map p as part of the axiomatic setting, we may define the inner product in $H_A \otimes H_B$ by linear extension of

$$\langle p(f',g'), p(f,g)\rangle_{H_A \otimes H_B} := \langle f', f\rangle_{H_A} \langle g', g\rangle_{H_B}. \tag{4.13}$$

It is then easy to show that if $(|a\rangle)_{a\in A}$ and $(|b\rangle)_{b\in B}$ are bases for H_A and H_B, respectively, then

$$p(|a\rangle, |b\rangle) \equiv |a\rangle \otimes |b\rangle \equiv |ab\rangle, \tag{4.14}$$

as (a,b) run through $A \times B$, is a basis of $H_A \otimes H_B$.

We now turn to operators on tensor product Hilbert spaces. For $T_A \in L(H_A)$ and $S_B \in L(H_B)$ we define the operator $T_A \otimes S_B$ on H_{AB} by linear extension of the rule

$$T_A \otimes S_B(\psi_A \otimes \varphi_B) := T_A \psi_A \otimes S_B \varphi_B. \tag{4.15}$$

It should be clear from the example $H_A = \ell^2(A)$ etc. that any operator on $H_A \otimes H_B \equiv H_{AB}$ is a finite linear combination of such $T_A \otimes S_B$, so that we may, literally or symbolically, write

$$L(H_A \otimes H_B) \cong L(H_A) \otimes L(H_B), \tag{4.16}$$

where the right-hand side should now be defined in the abstract way above, cf. Proposition 4.1.

We have linear isometric injections

$$L(H_A) \hookrightarrow L(H_{AB}); \tag{4.17}$$
$$T_A \mapsto T_A \otimes 1_{H_B}; \tag{4.18}$$
$$L(H_B) \hookrightarrow L(H_{AB}); \tag{4.19}$$
$$S_B \mapsto 1_{H_A} \otimes S_B, \tag{4.20}$$

which play a major role in our next topic about Hilbert space tensor products, namely the ***partial trace*** of the physicists. Recall that $D(H) \subset L(H)$ is the (convex) set of density matrices on H.

Theorem 4.2 *Let H_A and H_B be Hilbert spaces, with $H_{AB} \equiv H_A \otimes H_B$, Then there is an affine map*

$$D(H_{AB}) \to D(H_A); \tag{4.21}$$
$$\rho_{AB} \mapsto \rho_A, \tag{4.22}$$

which is completely characterized by any one the equivalent properties

$$(\rho_A \otimes \rho_B)_A = \mathrm{Tr}_{H_B}(\rho_B) \cdot \rho_A; \qquad (\rho_A \in D(H_A), \rho_B \in D(H_B)); \tag{4.23}$$
$$\langle \psi_A, \rho_A \varphi_A\rangle_{H_A} = \sum_b \langle \psi_A \otimes u_b, \rho_{AB}(\varphi_A \otimes u_b)\rangle_{H_{AB}}; \qquad (\psi_A, \varphi_A \in H_A); \tag{4.24}$$
$$\mathrm{Tr}_{H_A}(\rho_A T_A) = \mathrm{Tr}_{H_{AB}}(\rho_{AB} \cdot T_A \otimes 1_{H_B}); \qquad (T_A \in D(H_A), \tag{4.25}$$

where (u_b) is an arbitrary orthonormal basis of H_B. Mutatis mutandis*, we have a map*

$$D(H_{AB}) \to D(H_B); \tag{4.26}$$
$$\rho_{AB} \mapsto \rho_B. \tag{4.27}$$

Here ρ_A and ρ_B are called *reduced density matrices*. Because of (4.24), they are often written as

$$\rho_A = \mathrm{Tr}_{H_B}(\rho_{AB}); \qquad \rho_B = \mathrm{Tr}_{H_A}(\rho_{AB}). \tag{4.28}$$

The significance and structure of (4.25) comes out if we use (2.20). Denoting the (convex) set of states on $L(H)$ by $S(L(H))$, so that $\omega \in S(L(H))$ iff it takes the form $\omega(T) = \mathrm{Tr}(\rho T)$ for some $\rho \in D(H)$, Theorem 4.2 produces maps $S(L(H_{AB})) \to S(L(H_A))$, written $\omega_{AB} \mapsto \omega_A$, defined by the counterparts of (4.25), namely

$$\omega_A(T_A) = \omega_{AB}(T_A \otimes 1_{H_B}). \tag{4.29}$$

Thus ω_A is simply the restriction of ω_{AB} to $L(H_A) \subset (L(H_{AB})$, cf. (4.17) - (4.18), and similarly, ω_B defining $\rho_B$is the restriction of ω_{AB} to $L(H_B) \subset L(H_{AB})$, cf. (4.19) - (4.20).

An interesting case is where some unit vector $\psi_{AB} \in H_{AB}$ defines a pure state on $L(H_{AB})$ via

$$\rho_{AB} = |\psi_{AB}\rangle\langle\psi_{AB}|. \tag{4.30}$$

Now there are two different cases. If $\psi_{AB} = \psi_A \otimes \psi_B$ for unit vectors $\psi_A \in H_A$ and $\psi_B \in H_B$, then

$$\rho_A = |\psi_A\rangle\langle\psi_A|; \qquad \rho_B = |\psi_B\rangle\langle\psi_B|, \tag{4.31}$$

so that the reduced density matrices are again pure. Otherwise, we use the *Schmidt decomposition*:

Proposition 4.3 *Any unit vector $\psi_{AB} \in H_{AB}$ (where $H_{AB} \equiv H_A \otimes H_B$) may be written as*

$$\psi_{AB} = \sum_{i=1}^{d} \lambda_i e_{a(i)} \otimes u_{b(i)}, \tag{4.32}$$

where $\lambda_i \in \mathbb{C}$ with $\sum_i |\lambda_i|^2 = 1$, and (e_a) and (u_b) are bases of H_A and H_B, respectively.

Clearly, $d \leq \dim(H_A)$ and $d \leq \dim(H_B)$. To see the point, note that by definition of the tensor product one always has bases (e'_a) and (u'_b) such that $\psi_{AB} = \sum_{a,b} \lambda_{ab} e'_a \otimes u'_b$ is a *double* sum; the Schmidt decomposition (4.32) however is a *single* sum. From this, one easily infers (exercise) that

$$\rho_A = \sum_i |\lambda_i|^2 |e_{a(i)}\rangle\langle e_{a(i)}|; \qquad \rho_B = \sum_i |\lambda_i|^2 |u_{b(i)}\rangle\langle u_{a(i)}|. \tag{4.33}$$

Unless (4.32) has just one term, a pure state therefore reduces to mixed ones. This is impossible in classical probability theory, where pure states on $C(A \times B)$ correspond to point measures $\delta_{(a,b)} \in \mathrm{Prob}(A \times B)$, which reduce to point measures $\delta_a \in \mathrm{Prob}(A)$ and $\delta_b \in \mathrm{Prob}(B)$ (see exercise 8).

For example, let $H_A = H_B = \mathbb{C}^2$ with the usual basis (e_1, e_2), and take the ***Bell state***

$$\psi_{AB} = \frac{1}{\sqrt{2}}(e_1 \otimes e_2 - e_2 \otimes e_1) \in \mathbb{C}^2 \otimes \mathbb{C}^2. \tag{4.34}$$

If we start from (4.30), then ρ_A and ρ_B are (even maximally) mixed:

$$\rho_A = \tfrac{1}{2} \cdot 1_{H_A} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}; \tag{4.35}$$

$$\rho_B = \tfrac{1}{2} \cdot 1_{H_B} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}. \tag{4.36}$$

The state (4.34) is one of four Bell states, usually written as follows (or some variation thereof):

$$\Phi_\pm := \frac{1}{\sqrt{2}}(|00\rangle \pm |11\rangle); \qquad \Psi_\pm := \frac{1}{\sqrt{2}}(|01\rangle \pm |10\rangle. \tag{4.37}$$

These have the same properties, in that they all reduce to (4.35) - (4.36); see exercise.

Proposition 4.3 has an easy but spectacular converse, called *purification*:

Proposition 4.4 *Any density matrix $\rho_A \in D(H_A)$ is the reduced density matrix of a pure state, in the sense that there exists a Hilbert space H_B and a unit vector $\psi_{AB} \in H_{AB}$ with associated pure density matrix* (4.30)*, such that ρ_A is given by* (4.28)*, i.e., arises from* (4.22)*.*

The proof is easy: take the spectral resolution (2.18), which we rewrite as

$$\rho_A = \sum_{a \in A} p_a |a\rangle\langle a|, \tag{4.38}$$

where $(|a\rangle)_{a\in A}$ is some basis of H_A and $p \in \mathrm{Prob}(A)$. Take $H_B = \ell^2(A)$ with basis $(\delta_a)_{a\in A}$, and

$$\psi_{AB} = \sum_{a \in A} \sqrt{p_a}\, |a\rangle \otimes \delta_a. \tag{4.39}$$

The rest is an exercise. A neater construction is as follows. For any finite-dimensional Hilbert space H the space $L(H)$ is itself a Hilbert space, with (so-called *Hilbert–Schmidt*) inner product

$$\langle S, T \rangle := \mathrm{Tr}\,(S^* T). \tag{4.40}$$

The map $\psi \otimes \varphi \mapsto |\psi\rangle\langle\varphi|$ extends by linearity to a unitary $U : H \otimes \overline{H} \xrightarrow{\cong} L(H)$. Thus $L(H_A) \cong H_{AB}$ with $H_B = \overline{H}$. Take $\phi_{AB} := U^* \sqrt{\rho_A} \in H_{AB}$ and $\rho_{AB} = |\phi_{AB}\rangle\langle\phi_{AB}|$. Then $\mathrm{Tr}_{H_B} \rho_{AB} = \rho_A$ (exercise).

Clearly, the unit vector ψ_{AB}, with associated density matrix (4.30), is not the only purification of ρ_A. For example, one may also take a third Hilbert space H_C, replace H_B above by H_{BC}, and replace (4.39) by $\psi_{AB} \otimes \varphi_C$, where φ_C is any unit vector in H_C. Here is the general case:

Proposition 4.5 *Let unit vectors $\psi_{AB} \in H_{AB}$ and $\varphi \in H_{AC}$ both be purifications of $\rho_A \in D(H_A)$. Wlog, assume* $\dim(H_B) \leq \dim(H_C)$*. Then there is an isometry $V_{B\to C} : H_B \to H_C$ such that*

$$(\mathrm{id}_A \otimes V_{B\to C})\, \psi_{AB} = \varphi_{AC}. \tag{4.41}$$

Recall that an *isometry* $V : H_B \to H_C$ is a linear map such that $\langle V\psi_B, V\varphi_B\rangle_{H_C} = \langle \psi_B, \varphi_B\rangle_{H_B}$ for all $\psi_B, \varphi_B \in H_B$ (this is the Hilbert space analogue of an injective map between sets). It follows that $V^*V : H_B \to H_B$ is the identity and $VV^* : H_C \to H_C$ is the projection onto the image of V. More generally, a *partial isometry* $V : H_B \to H_C$ is a linear map such that V^*V is a projection on H_B and VV^* is a projection on H_C; this is interesting also if $H_C = H_B$.

Exercises

1. Show that $A \otimes B$, defined as in Proposition 4.1, is unique up to isomorphisms.
2. Prove the equivalence between the three conditions (4.23), (4.24), and (4.25).
3. Prove Proposition 4.3.
4. Prove (4.33).
5. Show that each of the four Bell states reduces to the matrices in (4.35) - (4.36) on both sides.
6. Show that (4.39) and (4.30) return (4.38) via (4.22). Also check the 'neater construction'.
7. Prove Proposition 4.5.
8. The partial trace is the quantum counterpart of taking marginals classically. Explain this. Hint: if $AB = A \times B$, then $p_{AB} \in \mathrm{Prob}(AB)$ gives marginals $p_A \in \mathrm{Prob}(A)$ and $p_B \in \mathrm{Prob}(B)$:

$$p_A(a) := \sum_{b\in B} p_{AB}(a,b); \qquad\qquad p_B(b) := \sum_{a\in A} p_{AB}(a,b). \tag{4.42}$$

   Also show that, as claimed in the main text, point measures $p_{AB} = \delta_{(a,b)} \in \mathrm{Prob}(A \times B)$ thus reduce to point measures $p_A = \delta_a \in \mathrm{Prob}(A)$ and $p_B = \delta_b \in \mathrm{Prob}(B)$, respectively.

5 Quantum entropy

In 1927, von Neumann presciently defined the quantum entropy of a state $\rho \in D(H)$ by

$$S(\rho) := -\mathrm{Tr}\,(\rho \log \rho), \tag{5.1}$$

where the logarithm and the entire expression $\rho \log \rho$ are defined by the spectral calculus: in finite-dimensional Hilbert spaces, which we use throughout this section, this simply means that if some operator has a spectral resolution $T = \sum_{\lambda \in \sigma(T)} \lambda \cdot e_\lambda$, and $f : \sigma(T) \to \mathbb{C}$ is well defined, then

$$f(T) = \sum_{\lambda \in \sigma(T)} f(\lambda) \cdot e_\lambda. \tag{5.2}$$

For example, because of (2.18) or (4.38) our density matrix ρ has its spectrum contained in $[0,1]$ and if we agree that $f(\lambda) = -\lambda \log \lambda$ vanishes at $\lambda = 0$ (either by continuity or by definition, then

$$\rho \log \rho = \sum_{a \in A} p_a \log p_a |a\rangle\langle a|. \tag{5.3}$$

Taking the trace in (5.1) over the basis (v_i) we obtain, writing $H(p)$ for the Shannon entropy $S(p)$,

$$S(\rho) = -\sum_{a \in A} p_a \log p_a = H(p). \tag{5.4}$$

According to the first two exercises below, the quantum entropy $S(\rho)$ seems to closely resemble its classical counterpart $H(p)$. But there is a BIG difference if we combine systems. Take any of the Bell states (4.37), for example Ψ_-, with ensuing pure state $\rho_{AB} = |\Psi_-\rangle\langle\Psi_-| \in D(\mathbb{C}^2 \otimes \mathbb{C}^2)$. Its entropy is zero (see exercise 1a). By (4.35) - (4.36) we have $S(\rho_A) = S(\rho_B) = \log 2$.

Why is this strange? Using the notation in (4.42), it can be shown that

$$H(p_{AB}) = H(p_A) + H(p_B|p_A); \qquad H(p_B|p_A) := -\sum_{a \in A, b \in B} p_{AB}(a,b) \log(p_{AB}(b|a)), \tag{5.5}$$

where $p_{AB}(b|a) = p_{AB}(a,b)/p_A(a)$, assuming $p_A(a) > 0$ for all $a \in A$ for simplicity. Consequently,

$$H(p_{AB}) \geq H(p_A); \qquad\qquad H(p_{AB}) \geq H(p_B). \tag{5.6}$$

So if $H(p_{AB}) = 0$, which is the case iff p_{AB} is a point measure $\delta_{(a,b)}$, then $H(p_A) = H(p_B) = 0$ by these inequalities, which is indeed the case since $p_A = \delta_a$ and $p_B = \delta_b$. But we have just seen that in quantum mechanics one may have $S(\rho_{AB}) = 0$ and yet $S(\rho_A) = S(\rho_B) > 0$, so that inequalities like $S(\rho_{AB}) \geq S(\rho_A)$ may be *wrong*. This is strange, because the inequalities (5.6) are expected: coding (a,b) should take more bits than coding a or b (on average), and hence by Shannon's noiseless coding theorem (5.6) should hold. Or: p_A is a coarse-graining of p_{AB}, so the average surprise in revealing $a \in A$ should be less than the average surprise in learning $(a,b) \in A \times B$.

One often finds an apparently more general (but actually equivalent) version of this: combine random variables $X \in A$ (that is, $X : \Omega \to A$ where $\mathbb{P} \in \mathrm{Prob}(\Omega)$ defines $P_X \in \mathrm{Prob}(A)$ via $P_X = \mathbb{P}X^{-1}$) and $Y \in B$ (where $Y : \Omega \to B$ and $P_Y = \mathbb{P}Y^{-1} \in \mathrm{Prob}(B)$) into $(X,Y) \in A \times B$, with joint law (i.e. probability distribution) $P_{X,Y}$, and associated Shannon entropies $H(X) \equiv H(P_X)$, $H(Y) \equiv H(P_Y)$, and $H(X,Y) \equiv H(P_{X,Y}$. Define a conditional entropy by

$$H(Y|X) := -\sum_{a \in A} P(X=a) \sum_{b \in B} P(Y=b|X=a) \log P(Y=b|X=a). \tag{5.7}$$

Then $H(X,Y) = H(X) + H(Y|X) \geq H(X)$, and likewise $H(X,Y) = H(Y) + H(X|Y) \geq H(Y)$. This reduces to our earlier analysis by taking $p_A = P_X$, $p_B = P_Y$, and $p_{AB} = P_{X,Y}$.

Apparently this is no longer the case in quantum mechanics! Here is the general situation:

Proposition 5.1 *Let $H_{AB} = H_A \otimes H_B$, with pure density matrix $\rho_{AB} \in D(H_{AB})$ given by* (4.30) *Then $S(\rho_{AB}) = 0$. For the corresponding reduced density matrices ρ_A and ρ_B (cf. Theorem 4.2) we have:*

1. *If $\psi_{AB} = \psi_A \otimes \psi_B$ for unit vectors $\psi_A \in H_A$ and $\psi_B \in H_B$, then*

$$S(\rho_A) = S(\rho_B) = 0. \tag{5.8}$$

2. *If ψ_{AB} does not consist of a single elementary tensor (and is called* correlated*, then*

$$S(\rho_A) = S(\rho_B) > 0. \tag{5.9}$$

This immediately follows from the Schmidt decomposition (4.32), which gives (4.33) and hence

$$S(\rho_A) = S(\rho_B) = H(|\lambda|^2), \tag{5.10}$$

where $|\lambda|^2$ is meant to be the probability distribution $i \mapsto p(i) = |\lambda_i|^2$ on the set $\{1,\ldots,d\}$.

Let $\rho_{AB} \in D(H_{AB})$ with reduced density matrices $\rho_A \in D(H_A)$ and $\rho_B \in D(H_B)$. Then

$$S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B), \tag{5.11}$$

with equality iff $\rho_{AB} = \rho_A \otimes \rho_B$ (this is called *subadditivity* of the entropy).[2] See exercise 3.

We now define the *relative quantum entropy* (or *quantum Kullback–Leibler divergence*) by

$$S(\rho\|\sigma) := \mathrm{Tr}\,(\rho(\log\rho - \log\sigma)) \quad \text{if } \ker(\sigma) \subseteq \ker(\rho); \tag{5.12}$$

$$S(\rho\|\sigma) := \infty \quad \text{otherwise}, \tag{5.13}$$

where $\rho, \sigma \in D(H)$. In the first line we define $\rho \log \sigma = 0$ on $\ker(\sigma)$, like $\rho \log \rho = 0$ on $\ker(\rho)$. Thus $\ker(\sigma) \subseteq \ker(\rho)$ is the quantum analogue of the condition $q \ll p$ in the definition of $D(p\|q)$.

Some properties of $D(p\|g$ generalize to $S(\rho\|\sigma)$; others don't. From the first category,

$$S(\rho\|1_H/\dim(H)) = -S(\rho) + \log(\dim(H)). \tag{5.14}$$

Indeed, the density matrix $1_H/\dim(H)$ plays the role of the flat probability distribution f. Similarly, the Gibbs inequality $D(p\|q) \geq 0$ is now replaced by the *Klein inequality* (cf. exercise 3)

$$S(\rho\|\sigma) \geq 0. \tag{5.15}$$

Again as in the classical case (where we did not mention this), eq. (5.15) may be sharpened to

$$S(\rho\|\sigma) \geq \tfrac{1}{2}\|\rho - \sigma\|_1^2, \tag{5.16}$$

where the *trace norm* on $L(H)$ is $\|T\|_1 := \mathrm{Tr}\,(\sqrt{T^*T})$. The main result in this category is:

Theorem 5.2 *The following properties are all true and are equivalent to each other:*

1. joint convexitiy: $S(t\rho_1 + (1-t)\rho_2\|t\sigma_1 + (1-t)\sigma_2) \leq tS(\rho_1\|\sigma_1) + (1-t)S(\rho_2\|\sigma_2)$.
2. monotonicity under state reduction: $S(\rho_A\|\sigma_A) \leq S(\rho_{AB}\|\sigma_{AB})$*, where* $\rho_A = \mathrm{Tr}_{H_B}\rho_{AB}$*, etc.*
3. monotonicity under quantum channels: $S(\Phi(\rho)\|\Phi(\sigma)) \leq S(\rho\|\sigma)$*, where* $\rho, \sigma \in D(H)$ *and* $\Phi : L(H) \to L(H)$ *is a completely positive and trace-preserving map (see next section).*

[2] This may be expanded to the *Araki–Lieb inequality* $|S(\rho_A) - S(\rho_B)| \leq S(\rho_{AB}) \leq S(\rho_A) + S(\rho_B)$.

Exercises

1. Show that $0 \leq S(\rho) \leq \log(\dim(H))$, and also that:
   (a) $S(\rho) = 0$ iff ρ is a pure state (so $\rho = |\psi\rangle\langle\psi|$ for some $\psi \in H$ with $\|\psi\| = 1$);
   (b) $S(\rho) = \log(\dim(H))$ iff $\rho = 1_H/\dim(H)$, the "most impure" state in $D(H)$.

2. Here is an interesting generalization of (5.4). Let density matrices $\rho_1, \ldots, \rho_n$ all in $D(H)$ have disjoint supports, where $\mathrm{supp}(\rho_i) = \ker(\rho_i)^\perp$, and let $p \in \mathrm{Prob}(\{1, \ldots, n\})$. Show that

$$S\left(\sum_{i=1}^{n} p_i \rho_i\right) = \sum_i p_i S(\rho_i) + H(p). \tag{5.17}$$

3. (a) Prove the following inequality, for self-adjoint $S, T \in L(H)$ with $\ker(T) \subseteq \ker(S)$:[3]

$$\mathrm{Tr}\,(S\log S - S\log T) \geq \mathrm{Tr}\,(S-T),. \tag{5.18}$$

   with equality iff $S = T$ (this is often called Klein's inequality, too).
   (b) Prove (5.15) from this inequality.
   (c) Prove (5.11) either from the same inequality or from (5.15).
   (d) Again from (5.18), prove that S is concave: for $t \in (0,1)$ and $\rho, \sigma \in D(H)$ we have

$$S(t\rho + (1-t)\sigma) \geq tS(\rho) + (1-t)S(\sigma). \tag{5.19}$$

   (e) Use (5.17) and $\mathrm{Tr}\,(S(\log(S+T) - \log S)) \geq 0$ for all $S \geq 0$, $T \geq 0$, to strengthen (5.19):

$$tS(\rho) + (1-t)S(\sigma) \leq S(t\rho + (1-t)\sigma) \leq tS(\rho) + (1-t)S(\sigma) + H(t), \tag{5.20}$$

   where $H(t) := -t\log t - (1-t)\log(1-t)$.

4. Prove (5.5).

5. Prove that for any $\rho_A, \sigma_A \in D(H_A)$ and $\rho_B, \sigma_B \in D(H_B)$ we have

$$S(\rho_A \otimes \rho_B \| \sigma_A \otimes \sigma_B) = S(\rho_A \| \sigma_A) + S(\rho_B \| \sigma_B). \tag{5.21}$$

6 Intermezzo: classical and quantum channels

Channels play a crucial role in information theory. In the classical case, the idea is that Alice sends Bob messages along a possibly noisy channel. If Alice sends the letter $a \in A$, then ideally Bob receives it as such. In practice, Bob has his own alphabet B and for each $a \in A$ he receives $b \in B$ with probability $p(b|a)$ (the ideal case would be $B = A$ and $p(b|a) = \delta_{ab}$). Thus a noisy channel is just a map $\mathcal{N} : A \to \mathrm{Prob}(B)$, where we write $(\mathcal{N}(a))(b) = \mathcal{N}_{ab} = p(b|a)$, such that

$$\mathcal{N}_{ab} \geq 0 \;\; (a \in A, b \in B); \qquad \sum_{b \in B} \mathcal{N}_{ab} = 1 \;\; (a \in A). \tag{6.1}$$

In the spirit of our earlier reformulation of probability distributions $p_A \in \mathrm{Prob}(A)$ in terms of states

$$\omega_A : C(A) \to \mathbb{C}; \qquad \omega_A(f_A) = \sum_{a \in A} p_A(a) f(a), \tag{6.2}$$

[3] In which case we defin $S\log T = 0$ on $\ker(T)$, much as $S\log S = 0$ on $\ker(S)$.

we may also regard $\mathscr{N}$ as a map (using the same symbol)

$$\mathscr{N} : C(A) \to C(B); \qquad \mathscr{N}(f)(b) = \sum_{a \in A} \mathscr{N}_{ab} f(a). \tag{6.3}$$

It follows from (6.3) the first part of (6.1) that $\mathscr{N}$ is *positive*, in the sense that if $f \geq 0$ (pointwise), then $\mathscr{N}(f) \geq 0$ (pointwise). In addition, (6.3) and the second part of (6.1) imply that f is *trace preserving*: simply defining $\mathrm{Tr}_A(f) = \sum_{a \in A} f(a)$ for $f \in C(A)$, we see that

$$\mathrm{Tr}_B(\mathscr{N}(f)) = \mathrm{Tr}_A(f). \tag{6.4}$$

Definition 6.1 *A* classical channel *is a trace-preserving positive linear map* $\mathscr{N} : C(A) \to C(B)$.

A classical channel has the virtue, for example, of restricting to an affine map from $\mathrm{Prob}(A) \subset C(A)$ to $\mathrm{Prob}(B) \subset C(B)$. Conversely, one may see (6.3) as as a linear extension of such a map.

It is therefore tempting to define a quantum channel as a positive and trace-preserving linear map between $L(H_A)$ and $L(H_B)$; for one thing, this restricts to an affine map from $D(H_A) \subset L(H_A)$ to $D(H_B) \subset L(H_B)$. But these conditions are not enough for building a successful theory of noisy asymptotic quantum communication (or, as we shall see later, quantum hypothesis testing). Noise comes from coupling Alice and Bob to some environment E, so that H_A and H_B are replaced by H_{AE} and H_{BE}, respectively (where $H_{AE} = H_A \otimes H_E$, etc.). The initial channel $\Phi : L(H_A) \to L(H_B)$ from Alice to Bob must then be extended to $\Phi \otimes \mathrm{id}_E : L(H_{AE}) \to L(H_{BE})$, defined by linear extension of $\Phi \otimes \mathrm{id}_E(T_A \otimes S_E) = \Phi(T_A) \otimes S_E$. Unexpectedly, if Φ is positive (and linear), then $\Phi \otimes \mathrm{id}_E$ may not be! See exercise. This cannot happen in the classical case: if $\mathscr{N} : C(A) \to C(B)$ is positive (and linear), and we couple A and B to E, then so is the canonically extended map

$$\mathscr{N} \times \mathrm{id}_E : C(A \times E) \to C(B \times E); \qquad \mathscr{N} \times \mathrm{id}_E(f)(b,e) = \sum_{a \in A} \mathscr{N}_{ab} f(a,e). \tag{6.5}$$

Apart from this, preserving the trace is automatic both in the classical and the quantum case.

To remedy this, we call a linear map $\Phi : L(H_A) \to L(H_B)$ *completely positive* or *CP* if for any finite-dimensional Hilbert space H_E the extension $\Phi \otimes \mathrm{id}_E : L(H_{AE}) \to L(H_{BE})$ is positive.

Definition 6.2 *A quantum channel* $\Phi : L(H_A) \to L(H_B)$ *is a CP trace-preserving linear map.*

The study of such maps is a huge field in functional analysis. Here are some examples:[4]

- *Unitary evolution:* $\Phi : L(H_A) \to L(H_B)$ is $\Phi(\rho_A) = U \rho_A U^*$, for some unitary $U : H_A \to H_B$.
- *Isometric embedding:* $\Phi : L(H_A) \to L(H_B)$ is $\Phi(\rho_A) = V \rho_A V^*$, for an isometry $V : H_A \to H_B$.
- *Conditional expectation:* To define a map $\Phi : L(H_B) \to L(H_A)$ in the opposite direction, let $H_A \subset H_B$ and identify $L(H_A)$ as a subspace of $L(H_B)$ by putting $T_A \psi_B = 0$ for all $\psi_B \in H_A^\perp$ (whilst T_A is 'itself" on $\psi_B \in H_A$). Using the projection $e : H_B \to H_B$, put

$$\Phi(\rho_B) = e \rho_B e. \tag{6.6}$$

- *Reduction:* $\Phi : L(H_{AB}) \to L(H_A)$ is now $\Phi(\rho_{AB}) = \rho_A \equiv \mathrm{Tr}_{H_B}(\rho_{AB})$, cf. Theorem 4.2.
- *Extension:* $\Phi : L(H_A) \to L(H_{AB})$ is given by $\Phi(\rho_A) = \rho_A \otimes \rho_B$, for some $\rho_B \in D(H_B)$.

[4] The adjoint $W^* : H_B \to H_A$ of $W : H_A \to H_B$ is the unique operator such that $\langle \psi_A, W^* \varphi_B \rangle_{H_A} = \langle W \psi_A, \varphi_B \rangle_{H_B}$ for all $\psi_A \in H_A$ and $\varphi_B \in H_B$. W is an isometry if $W^* W = 1_{H_A}$, and it is unitary if in addition $W W^* = 1_{H_B}$. In finite dimension (but only then!), if $H_B = H_A$ (or just $\dim(H_B) = \dim(H_A)$), then an isometry is automatically unitary.

- *Kraus maps:* To define $\Phi : L(H_A) \to L(H_B)$, take finitely many linear maps $X_i : H_A \to H_B$ satisfying $\sum_i X_i^* X_i = 1_{H_A}$ (which guarantees that Φ is trace preserving), and define

$$\Phi(\rho_A) = \sum_i X_i \rho_A X_i^*, \tag{6.7}$$

- *Stinespring maps:* To construct $\Phi : L(H_A) \to L(H_B)$ in a different way, take some "ancillary" Hilbert space H_C and an isometry $V : H_A \to H_{BC}$ (that is, $V^*V = 1_{H_A}$), and define Φ by

$$\Phi(\rho_A) = \mathrm{Tr}_{H_C}(V \rho_A V^*). \tag{6.8}$$

- *Lindblad maps:* To define $\Phi : L(H_A) \to L(H_B)$ we now take *two* ancillary Hilbert spaces H_C and H_D, a unit vector $\psi_D \in H_D$, and a unitary $U : H_{AD} \to H_{BC}$, to construct

$$\Phi(\rho_A) = \mathrm{Tr}_{H_C}(U(\rho_A \otimes |\psi_D\rangle\langle\psi_D|)U^*). \tag{6.9}$$

We omit the proof that these maps indeed satisfy the conditions in Definition 6.2, and also that every quantum channel can be put into any of the forms (6.7) - (6.9).[5] We also state without (the nontrivial) proof that the composition of CP maps is again CP; the same is true more trivially for TP maps, and hence, as expected, quantum channels may be composed into quantum channels.

But at least we now understand part 3 of Theorem 5.2! Here is a classical analogue:

Proposition 6.3 *For any classical channel $\mathcal{N} : C(A) \to C(B)$ and any $p, q \in \mathrm{Prob}(A)$ we have*

$$D(\mathcal{N}(p) \| D(\mathcal{N}(q))) \leq D(p\|q). \tag{6.10}$$

Exercises

1. Find an example of a positive map $\Phi : L(H_A) \to L(H_B)$ that fails to be completely positive.
2. Put the examples of Extension and Reduction into the Kraus form (6.7).
3. Show that any Kraus map is a quantum channel. Which property is lost if $\sum_i X_i^* X_i \neq 1_{H_A}$?
4. Prove Proposition 6.3.
5. Prove Theorem 5.2 on the basis of M.B. Ruskai's paper from 2002 (see Brightspace):
   (a) Take Lieb's inequality (17) as given without proof (all proofs of Theorem 5.2 need something like that), but verify it is an equality for positive real numbers R, S, T.
   (b) Explain her proof of (32) in section IV.A. This proves part 1 of Theorem 5.2.
   (c) Derive part 3 from part 2 using the Stinespring form (6.8) of Φ.
   (d) Derive part 1 from part 2 by explaining her argument at the beginning of section V.B.

   Any other way of proving Theorem 5.2 you present is also fine as long as it is clear!
6. Prove *strong subadditivity* of the von Neumann entropy S from Theorem 5.2.2, i.e.,

$$S(\rho_{ABC}) + S(\rho_B) \leq S(\rho_{AB}) + S(\rho_{BC}), \tag{6.11}$$

   where $\rho_{ABC} \in D(H_{ABC})$ is given and the other density matrices follow by partial tracing.

[5] See Theorem 5.3 and Lemma 5.5, as well as Exercise 5.5. for (6.8), in the Mastermath *Lecture Notes on Quantum Information Theory* by M. Walter and M. Ozols.

7 Noisy data compression: Classical theory

The oldest (and still most important) application of the von Neumann quantum entropy $S(\rho)$ and the relative quantum entropy $S(\rho\|\sigma)$ are to noisy data compression and quantum hypothesis testing, respectively. The first is an extension (also due to Shannon) of the noiseless data compression theory we discussed in Part 1, which by itself has no corresponding quantum theory since the latter always carries intrinsic noise. The second application is the one closest to Sanov's theorem (whose direct quantum counterpart is controversial and is still the subject of active research).

We start with classical encoding. Alice (the sender) holds a finite set or alphabet A and wants to send encoded messages to Bob (the receiver). We still work with a memoryless source, so that as before A carries a probability distribution $p \in \mathrm{Prob}(A)$, expanded to $p^N \in \mathrm{Prob}(A^N)$. Then:

- The noiseless case in part 1 was based on a map $C_N : A^N \to 2^*$ (where N is fixed) providing binary codewords $C_N(\sigma)$ of *variable* length for each message $\sigma \in A^N$ of *fixed* length. This map was assumed to be a prefix code and hence was uniquely decodable, so that Bob's decoding did not play a role. This is what made the procedure *noiseless*. Shannon's noiseless coding theorem (see Theorem 3.3 in part 1) stated that the optimal average length of Alice's codewords per symbol (i.e. $\times 1/N$) is approximately $S_2(p)$ (which we now call $H_2(p)$).

- The noisy case works with codewords of fixed length ℓ_N (again for given N), based on an encoding map $E_N : A^N \to 2^{\ell_N}$. Bob's decoding now enters explicitly via a map $D_N : 2^{\ell_N} \to A^N$, so that we obtain a combined map $D_N \circ E_N : A^N \to A^N$. Define the *fault probability*

$$\mathscr{F}_N(\ell_N, E_N, D_N) := p^N(D_N \circ E_N \neq \mathrm{id}) = p^N(\{\sigma \in A^N \mid D_N \circ E_N(\sigma) \neq \sigma\}), \tag{7.1}$$

  This is obviously the probability that Bob's decoding of Alice's codeword for σ is wrong. The goal is then to find, for some give error tolerance $\delta \in (0,1)$, the smallest possible integer ℓ_N so that maps E_N and D_N exist for which the fault probability remains acceptable, in that

$$\mathscr{F}_N(\ell_N, E_N, D_N) \leq \delta. \tag{7.2}$$

The answer to this optimization problem is *Shannon's noisy coding theorem*. We recall that for a finite probability space (A, p), the corresponding memoryless source (A^N, p^N) has entropy

$$H_2(p^N) = NH_2(p) = -N \sum_{a \in A} p(a) \log_2 p(a). \tag{7.3}$$

Theorem 7.1 *Given (A, p) and $\delta \in (0,1)$, for all except finitely many $N \in \mathbb{N}$ we have that:*[6]

- *If $\ell_N > H_2(p^N)$, then* (7.2) *can be achieved for some coding scheme $A^N \xrightarrow{E_N} 2^{\ell_N} \xrightarrow{D_N} A^N$.*
- *If $\ell_N < H_2(p^N)$, then* (7.2) *fails for all coding schemes (E_N, D_N) of this kind.*[7]

Since δ is arbitrary, it follows that one may find (E_N, D_N) with $\ell_N > H_2(p^N)$ for all N, such that

$$\lim_{N \to \infty} \mathscr{F}_N(\ell_N, E_N, D_N) = 0. \tag{7.4}$$

Compare with Shannon's noiseless coding theorem, according to which optimal noiseless codes (i.e., $\delta = 0$) with *varying* length ℓ_N of codewords (for given N), have *average* length $\langle \ell_N \rangle \approx H_2(p^N)$ as $N \to \infty$. In the noisy case, then, this is the optimal *fixed* length of codewords (for given N).

[6] This is sometimes stated in terms of the *rate* $R = \ell_N/N$, which is independent of Ni.e., if $R > H_2(p)$, then ....

[7] Equivalently: for all $\delta \in (0,1)$, there exists $N_0 \in \mathbb{N}$ such that for all $N > N_0$ and all $\ell_N > NS_2(p)$ there are (E_N, D_N) such that (7.2) holds; whereas for any N_0 there is $N > N_0$ such that $\ell_N < H_2(p^N)$, then (7.2) fails for all (E_N, D_N).

But if the codewords of the schemes $A^N \stackrel{E_N}{\to} 2^{\ell_N} \stackrel{D_N}{\to} A^N$ (almost) all have length $\ell_N < H_2(p^N)$, then

$$\lim_{N\to\infty} \mathscr{F}_N(\ell_N, E_N, D_N) = 1. \tag{7.5}$$

We now sketch a proof of Theorem 7.1, based on the AEP, which we now recall (cf. Part 1, §2). We switch from δ to ε, since δ is already in use, cf. (7.2). For any $\varepsilon > 0$, define $T^E_{N,\varepsilon}(p) \subset A^N$ by

$$T^E_{N,\varepsilon}(p) := \{\sigma \in A^N \mid 2^{-N(H_2(p)+\varepsilon)} \leq p^N(\sigma) \leq 2^{-N(H_2(p)-\varepsilon)}\}. \tag{7.6}$$

Then one has, from (2.41) and the proof of (2.40) from (2.48) in part 1, respectively,

$$(1-\varepsilon)2^{N(H_2(p)-\varepsilon)} \leq |T^E_{N,\varepsilon}(p)| \leq 2^{N(H_2(p)+\varepsilon)}; \tag{7.7}$$

$$p^N(T^E_{N,\varepsilon}(p)) \geq 1 - \frac{\sigma^2}{N\varepsilon^2}, \tag{7.8}$$

where in (7.7) the upper bound is true for any N, whereas the lower bound holds for sufficiently large N, and σ in (7.8) is a constant independent of ε and N (it does depend on p).

We start with the "successful" case $\ell_N > H_2(p^N)$, which in terms of the *rate* $R = \ell_N/N$ is the same as $R > H_2(p)$. We may now take $\varepsilon = \frac{1}{2}(R - H_2(p)) > 0$, in which case

$$N(H_2(p)+\varepsilon) = \tfrac{1}{2}(H_2(p^N) + \ell_N) < \ell_N, \tag{7.9}$$

so that it follows from (7.7) that $|T^E_{N,\varepsilon}(p)| < 2^{\ell_N}$. Therefore, Alice is able to define her encoding map $E_N : A^N \to 2^{\ell_N}$ such that it is injective on $T^E_{N,\varepsilon}(p) \subset A^N$, so that Bob can trivially construct his decoding map $D_N : 2^{\ell_N} \to A^N$ in such a way that $D_N \circ E_N(\sigma) = \sigma$ for all $\sigma \in T^E_{N,\varepsilon}(p)$, namely by sending $E_N(\sigma)$ to σ, which is unique by injectivity of E_N on $T^E_{N,\varepsilon}(p)$. Thus for all $N \geq \sigma^2/\varepsilon^2\delta$,

$$\mathscr{F}_N(\ell_N, E_N, D_N) = p^N(\{\sigma \in A^N \mid D_N \circ E_N(\sigma) \neq \sigma\}) \leq p^N(\sigma \notin A^N) \leq \frac{\sigma^2}{N\varepsilon^2} \leq \delta, \tag{7.10}$$

where we used (7.8). This proves the first case. The proof for $\ell_N < H_2(p^N)$ is an exercise. □

Preparing for quantum theory, we reformulate this in terms of classical channels. In (6.1) etc. we replace both A and B by A^N, so that initially we have a map $\mathscr{N}^{(N)} : A^N \to \mathrm{Prob}(A^N)$ defined by

$$\mathscr{N}^{(N)}_{ab} = \delta_{D_N(E_N(a))b}, \tag{7.11}$$

with ensuing classical channel $\mathscr{N}^{(N)} : C(A^N) \to C(A^N)$ as in (6.3), abusing notation. Restricting the channel $\mathscr{N}^{(N)}$ to $\mathrm{Prob}(A^N) \subset C(A^N)$, which indeed is its real home, you may check that

$$\mathscr{N}^{(N)}(p^N) = p^N \circ (D_N \circ E_N)^{-1}. \tag{7.12}$$

Now add an "environment" E (not to be confused with the map E_N!), and extend $\mathscr{N}^{(N)}$ to a map

$$\mathscr{N}^{(N)}_E \equiv \mathscr{N}^{(N)} \times \mathrm{id}_E : C(A^N \times E) \to C(A^N \times E), \tag{7.13}$$

as in (6.5). Finally, introduce the *trace norm* on any function space $C(X)$ for finite X by

$$\|f\|_1 := \sum_{x\in X} |f(x)|; \qquad \Rightarrow \qquad \|f-g\|_1 = \sum_{x\in X} |f(x) - g(x)|. \tag{7.14}$$

Proposition 7.2 *For any coding scheme* (ℓ_N, E_N, D_N) *and* $\delta \in (0,1)$ *the bound* (7.2) *holds iff*

$$\tfrac{1}{2}\|\mathscr{N}^{(N)}_E(p_{A^NE}) - p_{A^NE}\|_1 \leq \delta, \tag{7.15}$$

for any probability distribution $p_{A^NE} \in \mathrm{Prob}(A^N \times E)$ *whose marginal* $p_{A^N} \in \mathrm{Prob}(A^N)$ *equals* p^N.

Exercises

1. Prove the case $\ell_N < H_2(p^N)$ of Theorem 7.1, this time using the *lower* bound in (7.7).
2. Verify (7.12).
3. Prove Proposition 7.2. Hint: see Lemma 7.5 in the Mastermath *Lecture Notes on Quantum Information Theory* by M. Walter and M. Ozols (available on Brightspace).

8 Noisy data compression: Quantum theory

We now turn to Schumacher's quantum version of Shannon's noisy coding Theorem 7.1. First, it is natural to replace the classical coding scheme $A^N \stackrel{E_N}{\to} 2^{\ell_N} \stackrel{D_N}{\to} A^N$ by a Hilbert space analogue:

$$L(H_A^N) \stackrel{\mathscr{E}_N}{\longrightarrow} L(2^{\ell_N}) \stackrel{\mathscr{D}_N}{\longrightarrow} L(H_A^N), \tag{8.1}$$

where $\mathscr{E}_N : L(H_A^N) \to L(2^{\ell_N})$ and $\mathscr{D}_N : L(2^{\ell_N}) \to L(H_A^N)$ are quantum channels, that is, CP and TP linear maps. Hence we obtain a quantum channel $\Phi_N : L(H_A^N) \to L(H_A^N)$, where

$$\Phi_N := \mathscr{D}_N \circ \mathscr{E}_N. \tag{8.2}$$

Furthermore, there is a clean quantum analogue of the "classical" trace norm (7.14), namely

$$\|T\|_1 := \mathrm{Tr}\,(|T|); \qquad |T| := \sqrt{T^*T}; \qquad \|T-S\|_1 = \mathrm{Tr}\,(|T-S|), \tag{8.3}$$

where $T \in L(H)$; since $T^*T \geq 0$ the absolute value $|T|$ may be defined by the functional calculus. It turns out that proofs are easier if instead of the trace norm (8.3) one uses the *fidelity distance*,[8]

$$F_1(\rho,\sigma) := 1 - F(\rho,\sigma); \qquad F(\rho,\sigma) := \|\sqrt{\rho}\sqrt{\sigma}\|_1. \tag{8.4}$$

This is only defined for $\rho, \sigma \in D(H)$, but in that case it may replace the trace norm because of

$$1 - \sqrt{1 - (\tfrac{1}{2}\|\rho-\sigma\|_1)^2} \leq F_1(\rho,\sigma) \leq \tfrac{1}{2}\|\rho-\sigma\|_1. \tag{8.5}$$

In particular, if $\frac{1}{2}\|\rho-\sigma\|_1 \leq \delta$, then $F_1(\rho,\sigma) \leq \delta$. Moreover, one has $F_1(\rho,\sigma) = F_1(\sigma,\rho)$, and

$$0 \leq F_1(\rho,\sigma) \leq 1 \qquad \text{with} \qquad F_1(\rho,\sigma) = 0 \ \text{ iff } \ \rho = \sigma; \tag{8.6}$$

$$F(|\psi\rangle\langle\psi|,\sigma) = \sqrt{\langle\psi,\sigma\psi\rangle} \qquad \Rightarrow \qquad F_(|\psi\rangle\langle\psi|,|\varphi\rangle\langle\varphi|) = |\langle\varphi,\psi\rangle|. \tag{8.7}$$

A much more difficult result, which we merely state without proof, is *Uhlmann's theorem*:[9]

$$F_1(\rho_A,\sigma_A) = \inf\{F_1(\rho_{AB},\sigma_{AB})\}; \qquad F(\rho_A,\sigma_A) = \sup\{F(\rho_{AB},\sigma_{AB})\}, \tag{8.8}$$

where inf and sup are taken over all purifications ρ_{AB} of ρ_A and σ_{AB} of σ_A, respectively.

Now recall our general ideology of replacing (A,p) by (H_A,ρ_A) and (A^N,p^N) by (H_A^N,ρ_A^N), where $H_A^N \equiv H_A^{\otimes N}$ is the N-fold tensor product of H_A with itself, and likewise $\rho_A^N \equiv \rho_A^{\otimes N} \in D(H_A^N)$.

Replacing $\frac{1}{2}\|\rho-\sigma\|_1$ by $F_1(\rho,\sigma)$, Proposition 7.2 and (7.12) then suggest that for given (H_A,ρ_A), N, and fault tolerance $\delta \in (0,1)$, Alice and Bob's (de)coding task is to find the smallest (i.e. most efficient) value of ℓ_N for which there are quantum channels (8.1) such that one has

$$F_1((\mathscr{D}_N \circ \mathscr{E}_N) \otimes \mathrm{id}_E(\rho_{A^N E}), \rho_{A^N E}) \leq \delta, \tag{8.9}$$

[8] $F(\rho,\sigma)$ (or sometimes its square!) is called the *fidelity*.

[9] For a proof see the Mastermath *Lecture Notes on Quantum Information Theory*, Theorem 4.13.

for all H_E and all extensions $\rho_{A^N E} \in D(H_{A^N E})$ of $\rho_A^N \in D(H_A^N)$. It will be useful to define the *fidelity distance of a quantum channel* $\Phi_C : L(H_C) \to L(H_C)$ relative to a density matrix $\rho_C \in D(H_C)$ by

$$\mathscr{F}_1(\Phi_C, \rho_C) := \sup\{F_1(\Phi_C \otimes \mathrm{id}_E(\rho_{CE}), \rho_{CE})\}, \tag{8.10}$$

where the supremum is taken over all purifications ρ_{CE} of ρ_C, cf. (8.8). Taking $H_C = H_A^N$, $\Phi_C = \Phi_N$ as defined in (8.2), and $\rho_C = \rho_A^N$ in (8.10), the condition (8.9) is obviously equivalent to

$$\mathscr{F}_1(\mathscr{D}_N \circ \mathscr{E}_N, \rho_A^N) \leq \delta. \tag{8.11}$$

The reason this slight reformulation of (8.9) is useful lies in the following result:

Proposition 8.1 *The supremum in* (8.10) *is reached for an arbitrary purification* ρ_{CE} *of* ρ_C*, i.e.,*

$$\rho_{CE} = |\psi_{CE}\rangle\langle\psi_{CE}|. \tag{8.12}$$

This implies that the supremum is the same for any such purification. We postpone the proof of this proposition, whose use is twofold. First, via the proof of Lemma 8.3 below it provides a key step in the computation of (8.11), and hence in the proof of Theorem 8.2 below. Second, it helps to understand what it all means. The above formulation of the quantum coding task was found by analogy with the classical (noisy) case, but perhaps there is a different, more quantum-mechanical way of arriving at the same formulation? There is!

- Classically, Alice sends N letters a from her alphabet A to Bob, distributed by $p \in \mathrm{Prob}(A)$.
- Quantum-mechanically, if her source is H_A she sends pure states

  $$\rho_a = |\psi_a\rangle\langle\psi_a|, \tag{8.13}$$

  which as we have seen are the closest counterparts to $a \in A$, with probability p_a, where the unit vectors $\psi_a \in H_A$ need not be orthogonal.[10] Her associated density matrix is the mixture

  $$\rho_A := \sum_{a \in A} p_a \rho_a. \tag{8.14}$$

Let $\rho_{AD} = |\psi_{AD}\rangle\langle\psi_{AD}|$ be some purification of ρ_A. Now Bob only has observables in $L(H_A)$ at his disposal, more precisely, operators $T_A \otimes 1_{H_D} \subset L(H_{AD})$ where $T_A \in L(H_A)$. Since

$$\mathrm{Tr}_{H_{AD}}(\rho_{AD}(T_A \otimes 1_{H_D})) = \mathrm{Tr}_{H_A}(\rho_A T_A), \tag{8.15}$$

see Theorem 4.2 (in other words, ρ_{AD} reduces to ρ_A), the receiver Bob cannot distinguish between:

1. Alice sending him pure states ρ_a (or, as physicists would say, ψ_a) with probability p_a;
2. Alice and Doris (who holds H_D) sending him the pure state ρ_{AD} (or ψ_{AD}) every time.

Hence the communication protocol between Alice and Bob should not distinguish these scenarios:

- In particular, their problem of finding optimal numbers ℓ_N and quantum channels $\mathscr{E}_N$, $\mathscr{D}_N$ only depends on ρ_A in (8.14) and not on the p_a and ρ_a separately. And:
- It should not matter which purification ρ_{AD} of ρ_A Alice and Doris share.

[10] Hence (8.14) in general is different from the spectral resolution (4.38) of ρ_A, in which the unit vectors $|a\rangle$ are orthogonal, so that the p_a's are also different from those in (8.14). We do not bother to change the notation, though.

For example, take the canonical purification

$$\psi_{AD} = \sum_{a\in A} \sqrt{p_a}\, \psi_a \otimes \delta_a, \tag{8.16}$$

which lies in $H_A \otimes \ell^2(A)$, so that $H_D = \ell^2(A)$. Then $\psi_{AD}^N \in H_A^N \otimes \ell^2(A^N)$ is given by

$$\psi_{AD}^N = \sum_{\sigma\in A^N} \sqrt{p^N(\sigma)}\psi_{\sigma_0} \otimes \cdots \otimes \psi_{\sigma_{N-1}} \otimes \delta_\sigma, \tag{8.17}$$

from which it follows (exercise) that $\rho_{A^N E} = |\psi_{AD}^N\rangle\langle\psi_{AD}^N|$ is a purification of

$$\rho_A^N = \sum_{\sigma\in A^N} p^N(\sigma)\rho_{\sigma_0} \otimes \cdots \otimes \rho_{\sigma_{N-1}} \in H_A^N, \tag{8.18}$$

where $H_E = H_D^N = \ell^2(A)^N = \ell^2(A^N)$, noting that $H_{A^N E} = H_A^N \otimes H_E = H_A^N \otimes H_D^N \cong H_{AD}^N$. That is,

$$\mathrm{Tr}_{H_E}(|\psi_{AD}^N\rangle\langle\psi_{AD}^N|) = \rho_A^N. \tag{8.19}$$

It follows from Proposition 8.1 that $\rho_{A^N E}$ may be used to compute the error bound (8.10) - (8.11).

We are now ready to state and prove a satisfactory quantum version of Shannon's noisy coding theorem 7.1, due to Schumacher (1995). Like Theorem 7.1, this may be stated either in terms of either ℓ_N and $S(\rho_A^N)$ or $R = \ell_N/N$ and $S(\rho_A)$, as we shall do here for a change; the relationship is

$$S(\rho_A^N) = NS(\rho_A), \tag{8.20}$$

as follows from the comment below (5.11). Also, instead of using (5.1) we switch to base 2:

$$S_2(\rho) = -\mathrm{Tr}\,(\rho \log_2 \rho). \tag{8.21}$$

Theorem 8.2 *Given (H_A, ρ_A) and $\delta \in (0,1)$, for all except finitely many $N \in \mathbb{N}$ we have that:*

- *If $R > S_2(\rho_A)$, then* (8.11) *can be achieved for some coding scheme* (8.1)*.*
- *If $R < S_2(\rho_A)$, then* (8.11) *fails for all coding schemes of this kind.*

This implies the obvious quantum analogues of (7.4) and (7.5), too.

Proof. We use the spectral resolution $\rho_A = \sum_{a\in A} p_a |a\rangle\langle a|$, see (4.38), so that

$$S_2(\rho_A) = H_2(p), \tag{8.22}$$

see (5.4).[11] Regarding the vectors $|a\rangle$ as basis vectors $\delta_a \in \ell^2(A)$, we henceforth identify H_A with $\ell^2(A)$ and, using (4.1), identify H_A^N with $\ell^2(A^N)$, with basis vectors written as $|\sigma\rangle \equiv \delta_\sigma$. Then

$$\rho_A^N = \sum_{\sigma\in A^N} p^N(\sigma)|\sigma\rangle\langle\sigma|. \tag{8.23}$$

As in the classical case, we now use the AEP, recall (7.6) - (7.8). The subset $T_{N,\varepsilon}^E(p) \subset A^N$ has an obvious counterpart in the linear subspace $H_{N,\varepsilon}^E(\rho_A) \subset H_A^N$, defined by

$$H_{N,\varepsilon}^E(\rho_A) := \mathrm{span}\{|\sigma\rangle, \sigma \in T_{N,\varepsilon}^E(p)\} = \Pi_{N,\varepsilon}^E(\rho_A) H_A^N; \tag{8.24}$$

$$\Pi_{N,\varepsilon}^E(\rho_A) := \sum_{\sigma\in T_{N,\varepsilon}^E(p)} |\sigma\rangle\langle\sigma|, \tag{8.25}$$

[11] This is not the same p as in (8.14) except if the ψ_a form a basis of H_A and then may be identified with the $|a\rangle$.

where we also introduced the associated projection $\Pi^E_{N,\varepsilon}(\rho_A)$. From (7.7) one derives

$$\mathrm{Tr}_{H_A^N}(\Pi^E_{N,\varepsilon}(\rho_A)) = \dim(H^E_{N,\varepsilon}(\rho_A)) \leq 2^{N(S_2(\rho_A)+\varepsilon)}. \tag{8.26}$$

In the first case $R > S_2(\rho_A)$, we again chose $\varepsilon = \frac{1}{2}(R - S_2(\rho_A))$, so that

$$\dim(H^E_{N,\varepsilon}(\rho_A)) \leq 2^{N(S_2(\rho_A)+\varepsilon)} = 2^{N(R-\varepsilon)} \leq 2^{\ell_N} = \dim(\mathbb{C}^{2\ell_N}),. \tag{8.27}$$

where $\ell_N = \lfloor RN \rfloor$. Hence there is a partial isometry $V : H_N^A \to \mathbb{C}^{2\ell_N}$ (see the end of §4) such that $V^*V : H_A^N \to H_A^N$ equals the projection $\Pi^E_{N,\varepsilon}(\rho_A)$ and $VV^* : \mathbb{C}^{2\ell_N} \to \mathbb{C}^{2\ell_N}$ is the identity on VH_A^N and zero on its orthogonal complement $(VH_A^N)^\perp$.

We are now in a position to define Alice's and Bobs' quantum channels

$$\mathscr{E}_N : L(H_A^N) \to L(\mathbb{C}^{2\ell_N}); \qquad \mathscr{D}_N : L(\mathbb{C}^{2\ell_N}) \to L(H_A^N). \tag{8.28}$$

The main ingredient of $\mathscr{E}_N$ is the Hilbert space counterpart of the classical injective map from $T^E_{N,\varepsilon}(p)$ to 2^{ℓ_N}, namely $\mathscr{E}'_N(\rho) = V\rho V^*$, but although CP (why?) this map is not TP (why not?). To make it TP, pick arbitrary density matrices $\rho_1 \in D(\mathbb{C}^{2\ell_N})$ and $\rho_2 \in D(H_A^N)$, and define

$$\mathscr{E}_N(\rho) := V\rho V^* + \mathrm{Tr}_{H_A^N}((1_{H_N^A} - V^*V)\rho)\rho_1; \tag{8.29}$$

$$\mathscr{D}_N(\sigma) := V^*\sigma V + \mathrm{Tr}_{\mathbb{C}^{2\ell_N}}((1_{\mathbb{C}^{2\ell_N}} - VV^*)\sigma)\rho_2. \tag{8.30}$$

These maps are obviously TP (check!), and it can be shown from the property

$$\mathrm{Tr}_{H_A^N}((1_{H_N^A} - V^*V)\rho) = \mathrm{Tr}_{H_A^N}(\sqrt{1_{H_N^A} - V^*V}\,\rho\,\sqrt{1_{H_N^A} - V^*V})), \tag{8.31}$$

and likewise the trace in (8.30), that they are also CP, and hence are quantum channels. Note that

$$\mathscr{D}_N \circ \mathscr{E}_N(\rho) = V^*V\rho V^*V + \cdots, \tag{8.32}$$

which assumes the Kraus form (6.7), with $X_1 = V^*V = \Pi^E_{N,\varepsilon}(\rho_A)$. This motivates the need for:

Lemma 8.3 *Let a quantum channel $\Phi_C : L(H_C) \to L(H_C)$ have a Kraus representation* (6.7)*. Then*

$$\mathscr{F}_1(\Phi_C, \rho_C) = 1 - \sqrt{\sum_{i=1}^d |\mathrm{Tr}_{H_C}(X_i\rho_C)|^2}, \tag{8.33}$$

where the fidelity distance $\mathscr{F}_1(\Phi_C, \rho_C)$ is defined by (8.10) *via* (8.4)*.*

Before we prove this lemma, we note that (8.33) immediately implies that for any $i = 1, \ldots, d$,

$$\mathscr{F}_1(\Phi_C, \rho_C) \leq 1 - |\mathrm{Tr}_{H_C}(X_i\rho_C)|. \tag{8.34}$$

We apply Lemma 8.3 and (8.34) to $\Phi_C = \mathscr{D}_N \circ \mathscr{E}_N$ and $H_C = H_A^N$, taking $i = 1$; Recall that

$$X_1 = X_1^* = \Pi^E_{N,\varepsilon}(\rho_A). \tag{8.35}$$

To get an estimate of the right-hand side of (8.34), as exercise it follows from (7.8) that

$$\mathrm{Tr}_{H_A^N}(\rho_A^N \Pi^E_{N,\varepsilon}(\rho_A)) \geq 1 - \frac{\sigma^2}{N\varepsilon^2}. \tag{8.36}$$

Eqs. (8.34) and (8.36) give (8.11) for all $N \geq \sigma^2/\varepsilon^2\delta$, just as in the classical case.

This proves Theorem 8.2 for the case $R > S_2(\rho_A)$. The opposite case $R < S_2(\rho_A)$ follows by similar tricks from the lower bound in (7.7), but since this (negative) case is less important we refer to Theorem 7.10 in the Mastermath *Lecture Notes on Quantum Information Theory*. □

Proof of Lemma 8.3. Let $\Phi_C(\rho) = \sum_i X_i \rho X_i^*$ for some $X_i : H_C \to H_C$. We now invoke Proposition 8.1 to compute the channel capacity (8.10). Using (8.12) and (8.7), we have

$$\begin{aligned}\mathscr{F}_1(\Phi_C, \rho_C) &= F(\Phi_C \otimes \mathrm{id}_E(\rho_{CE}), \rho_{CE})^2 = \langle \psi_{CE}, \Phi_C \otimes \mathrm{id}_E(\rho_{CE})\psi_{CE}\rangle \\ &= \sum_i |\langle \psi_{CE}, X_i \otimes 1_{H_E}\psi_{CE}\rangle|^2 = \sum_i |\mathrm{Tr}_{H_{CE}}(X_i \otimes 1_{H_E}\rho_{CE})|^2 \\ &= \sum_i |\mathrm{Tr}_{H_C}(X_i\rho_C)|^2. \end{aligned} \tag{8.37}$$

Proof of Proposition 8.1. Using F instead of F_1, see (8.4), we have to prove that the infimum in

$$\mathscr{F}(\Phi_C, \rho_C) := \inf\{F(\Phi_C \otimes \mathrm{id}_E(\rho_{CE}), \rho_{CE})\} \tag{8.38}$$

is attained by an arbitrary purification (8.12) of ρ_C. This relies on two properties of F:

1. $F(\rho_{CD}, \sigma_{CD}) \leq F(\rho_C, \sigma_C)$ for any extensions ρ_{CD} and σ_{CD} of ρ_C and σ_C, respectively.
2. $F(V\rho_C V^*, V\sigma_C V^*) = F(\rho_C, \sigma_C)$ for any isometry $V : H_C \to H_D$.

By definition of an infimum, for any $\varepsilon > 0$ there exists some H_D and $\rho_{CD} \in D(H_{CD})$ for which

$$|\mathscr{F}(\Phi_C, \rho_C) - F(\Phi_C \otimes \mathrm{id}_D(\rho_{CD}), \rho_{CD})| < \varepsilon. \tag{8.39}$$

Take a purification ρ_{CDB} of ρ_{CD}. By property 1 of F, eq. (8.39) also holds for ρ_{CDB} instead of ρ_{CD}. Since ρ_{CD} reduces to ρ_C, so does ρ_{CDB} (via ρ_{CD}), so ρ_{CDB} is also a purification of ρ_C. Now take an arbitrary purification ρ_{CE} of ρ_C. By Proposition 4.5 the purifications ρ_{CDB} and ρ_{CE} of ρ_C are related by an isometry. Hence by property 2 of F, eq. (8.39) also holds for ρ_{CB} instead of ρ_{CD}. Now ρ_{CE} is independent of ε, so letting $\varepsilon \to 0$ finishes the proof. □

Exercises

1. Prove the upper bound in (8.5), as well as (8.6) and (8.7).
2. Prove (8.19).
3. Provide the details of the computation in the proof of Lemma 8.3.
4. Prove the two properties of F in the proof of Proposition 8.1.
5. Derive (8.26) and (8.36). Hint: See Lemma 7.12 in the Mastermath *Lecture Notes on Quantum Information Theory*.

9 Classical hypothesis testing

We only consider the simplest case of hypothesis testing that falls into the context of this course: based on randomly drawing letters from our finite alphabet A, we try to determine the probability distribution $p \in \mathrm{Prob}(A)$ supposed to control the sampling. And within this simple setting we only consider the case where just two hypotheses are considered, say $p = p_0$ and $p = p_1$. Throughout this section we assume for simplicity that $p_i(a) > 0$ for each a and $i = 0, 1$. Let H_i be the hypothesis that p_i is true ($i = 0, 1$). If we only draw once and get $a \in A$, it seems reasonable to:[12]

- accept H_0 (i.e., $p = p_0$) if $p_0(a) > p_1(a)$;
- accept H_1 (i.e., $p = p_1$) if $p_0(a) \leq p_1(a)$;

In other words, H_0 is accepted if $a \in T^*$ and H_1 is accepted if $a \notin T^*$, where

$$T^* = \{a \in A \mid p_0(a) > p_1(a)\}. \tag{9.1}$$

More generally, a *test* of $\mathsf{H}_0, \mathsf{H}_1$ consists of a subset $T \subset A$ such that if $a \in T$ is drawn we accept H_0, whereas $a \notin T$ yields H_1. What is it, then, that makes the test $T = T^*$ stand out? There are many ways to define what it means for a test to be optimal, usually involving costs and risks associated with accepting or rejecting H_0 and H_1. The simplest possibility involves:[13]

- $\alpha(T) = p_0(T^c)$, the *false alarm* probability that H_1 is accepted on the basis of the observation a (i.e., one draws $a \in T^c$) although in fact H_0 is true (which explains the p_0 in α).
- $\beta(T) = p_1(T)$, the *miss* probability that H_0 is accepted from a although H_1 is true.

If H_0 and H_1 are treated symmetrically one may simply minimize the total error $\alpha(T) + \beta(T)$.

Lemma 9.1 *For any test $T \subset A$ one has $\alpha(T) + \beta(T) \geq \alpha(T^*) + \beta(T^*)$, or, in other words,*

$$\gamma := \inf_{T \subset A} \{\alpha(T) + \beta(T)\} = \alpha(T^*) + \beta(T^*). \tag{9.2}$$

Therefore, if $\alpha(T) \leq \alpha(T^*)$, then $\beta(T) \geq \beta(T^*)$, and if $\beta(T) \leq \beta(T^*)$, then $\alpha(T) \geq \alpha(T^*)$.

This is the simplest case of the so-called *Neyman–Pearson lemma*.[14] The test T^* is not a unique minimizer of γ; for example, one could equally well take $\tilde{T}_* = \{a \in A \mid p_0(a) \geq p_1(a)\}$.

What happens if there are *priors* π_i for H_i, that is, π_i is the probability that H_i is true before any observation is made? Then we want to minimize $\pi_0\alpha(T) + \pi_1\beta(T)$;[15] above we implicitly had $\pi_0 = \pi_1 = \frac{1}{2}$. The optimal test follows from Lemma 9.1, in which we just replace p_i by $\pi_i p_i$:

Proposition 9.2 *Let $t > 0$ and define $T^*(t) := \{a \in A \mid p_0(a) > t\,p_1(a)\}$. For any test $T \subset A$,*

$$\inf_{T \subset A} \{\alpha(T) + t\beta(T)\} = \alpha(T^*(t)) + t\beta(T^*(t)). \tag{9.3}$$

If $\alpha(T) \leq \alpha(T^*(t))$, then $\beta(T) \geq \beta(T^*(t))$, and if $\beta(T) \leq \beta(T^*(t))$, then $\alpha(T) \geq \alpha(T^*(t))$.

[12] One may accept either hypothesis if $p_0(a) = p_1(a)$; we choose to accept H_1 in that case. Also, $B^c = A \backslash B$.

[13] The terminology comes from situations where H_0 states the absense of some threat and H_1 its presence.

[14] The actual lemma incorporates more tests: a Neyman–Pearson test T is seen as a map $\mathsf{T} : A \to \mathrm{Prob}\{0, 1\}$ that gives the probability $\mathsf{T}(a)$ of accepting H_0 or H_1. Our setting corresponds to $\mathsf{T}(a)(0) = 1$ if $a \in T^c$ and $\mathsf{T}(a)(1) = 1$ if $a \in T$. Lemma 9.1 remains true also in that case, so that the optimal test is of this binary form despite the generality. See for example H.V. Poor, *An Introduction to Signal Detection and Estimation*, Second Edition (Springer, 1994), §II.D.

[15] We assume that $\pi_i > 0$ for $i = 0, 1$.

This answers our question, since it follows that $T^*(\pi_1/\pi_0)$ minimizes $\pi_0\alpha(T)+\pi_1\beta(T)$, i.e.,

$$\inf_{T\subset A}\{\pi_0\alpha(T)+\pi_1\beta(T)\}=\pi_0\alpha(T^*(\pi_1/\pi_0))+\pi_1\beta(T^*(\pi_1/\pi_0)). \tag{9.4}$$

Though trivial, it will be crucial that $T^*(t)$ is the same as the *log-likelihood test* $\mathsf{LL}>\log t$, where

$$\mathsf{LL}(a)=\log\left(\frac{p_0(a)}{p_1(a)}\right). \tag{9.5}$$

Any test becomes more reliable if we replace a single drawing $a\in A$ by N such drawings $\sigma\in A^N$, assumed i.i.d. and hence governed by p_i^N if H_i is true (we keep H_0 and H_1 as stated, i.e., about p_0 and p_1). Given some test $T_N\subset A^N$, the interpretation of this evidence is then as follows:

- If $\sigma\in T_N$ we accept H_0, with false alarm error $\alpha_N(T_N)=p_0^N(T_N^c)$;
- If $\sigma\in T_N^c$ we accept H_1, with miss error $\beta_N(T_N)=p_1^N(T_N)$.

Taking $\pi_0=\pi_1=\frac{1}{2}$ for simplicity, we are interested in the asymptotic behaviour as $N\to\infty$ of

$$\gamma_N:=\inf_{T_N\subset A^N}\{\alpha_N(T_N)+\beta_N(T_N)\}. \tag{9.6}$$

This involves a new entropy-like function called the *Chernoff information*,[16] which is defined by

$$C(p_0,p_1):=-\log\left(\inf_{\lambda\in(0,1)}Z_\lambda(p_0,p_1)\right);\qquad Z_\lambda(p_0,p_1):=\sum_{a\in A}p_0(a)^{1-\lambda}p_1(a)^\lambda \tag{9.7}$$

This is usually seen as a particular distance between p and q in $\mathrm{Prob}(A)$: $C(p,q)$ is symmetric (unlike $D(p\|q)$) and can be shown to be convex in each of its arguments. This may trivially be rewritten in terms of the relative *Rényi entropy* $R_\lambda(p_0,p_1)$, as

$$C(p_0,p_1)=-\inf_{\lambda\in(0,1)}R_\lambda(p_0,p_1);\qquad R_\lambda(p_0,p_1):=\log\left(\sum_{a\in A}p_0(a)^{1-\lambda}p_1(a)^\lambda\right). \tag{9.8}$$

Theorem 9.3 (Chernoff) *With $\alpha_N(T)=p_0^N(T_N^c)$, $\beta_N(T)=p_1^N(T_N)$, and γ_N as in* (9.6)*, we have*

$$\lim_{N\to\infty}\frac{1}{N}\log\gamma_N=-C(p_0,p_1). \tag{9.9}$$

In fact, this result for $\pi_0=\pi_1=\frac{1}{2}$ is also true for general values of (π_0,π_1); the inequality in (9.14) below gets an extra term $u_N=\log(t)/N$ on the right, which vanishes as $N\to\infty$.

Proof. Mathematically, $\sigma\in A^N$ is a single drawing from A^N. Hence Lemma 9.1 applies with $(A,p)\rightsquigarrow(A^N,p^{(N)})$ and $p_i\rightsquigarrow p_i^N$, where now $H_i^{(N)}$ states that $p^{(N)}=p_i^N$, $i=1,2$, but this is the same as $p=p_i$, so $H_i^{(N)}=H_i$. Hence an optimal test attaining the infimum in (9.6) is given by

$$T_N^*\equiv T_N^*(0)=(\mathsf{LL}_N>0), \tag{9.10}$$

where the log-likelihood and associated *Neyman–Pearson tests* $T_N^*(u_N)$ for $u_N\in\mathbb{R}$ are given by

$$\mathsf{LL}_N(\sigma)=\frac{1}{N}\log\left(\frac{p_0^N(\sigma)}{p_1^N(\sigma)}\right);\quad T_N^*(u_N):=(\mathsf{LL}_N>u_N)=\left\{\sigma\in A^N\mid\mathsf{LL}_N(\sigma)>u_N\right\}. \tag{9.11}$$

[16] The expression $R_\lambda(p_0,p_1)=\log(Z_\lambda(p_0,p_1))$ is called the relative *Rényi entropy*.

Now recall the formulae for the empirical measure $L_N : A^N \to \mathrm{Prob}(A)$ and the relative entropy:

$$L_N(\sigma) = \frac{1}{N}\sum_{n=0}^{N-1} \delta_{\sigma_n}; \qquad\qquad D(p\|q) = \sum_{a\in A} p(a)(\log p(a) - \log q(a)), \tag{9.12}$$

respectively (the latter for $p \ll q$, which is all we need). From these, it is an exercise to derive

$$\mathsf{LL}_N(\sigma) = D(L_N(\sigma)\|p_1) - D(L_N(\sigma)\|p_0). \tag{9.13}$$

Hence

$$T_N^* = \{\sigma \in A^N \mid D(L_N(\sigma)\|p_1) > D(L_N(\sigma)\|p_0)\}. \tag{9.14}$$

Defining

$$B_* := \{p \in \mathrm{Prob}(A) \mid D(p\|p_1) > D(p\|p_0)\}; \tag{9.15}$$
$$B_*^c = \{p \in \mathrm{Prob}(A) \mid D(p\|p_1) \leq D(p\|p_0)\}, \tag{9.16}$$

it follows that $\sigma \in T_N^*$ iff $L_N(\sigma) \in B_*$, and likewise for their complements. The asymptotics of

$$\alpha_N(T_N^*) = p_0^N(\sigma \notin T_N^*) = p_0^N(L_N(\sigma) \in B_*^c); \tag{9.17}$$
$$\beta_N(T_N^*) = p_1^N(\sigma \in T_N^*) = p_1^N(L_N(\sigma) \in B_*), \tag{9.18}$$

now follow from Sanov's theorem, which for our neat sets B_* and B_*^c applies in its simplest form

$$\lim_{N\to\infty} \frac{1}{N} \log p_0^N(L_N \in B_*^c) = -\inf\{D(p\|p_0) \mid p \in B_*^c\}; \tag{9.19}$$
$$\lim_{N\to\infty} \frac{1}{N} \log p_1^N(L_N \in B_*) = -\inf\{D(p\|p_1) \mid p \in B_*\}. \tag{9.20}$$

In (9.19) the infimum of $D(p\|p_0)$ must be computed over B_*^c, where $D(p\|p_1) \leq D(p\|p_0)$. It is an exercise to show that this is the same as its infimum over the boundary ∂B_*^c, which is given by

$$\partial B_*^c = \partial B_* = \{p \in \mathrm{Prob}(A) \mid D(p\|p_1) = D(p\|p_0)\}. \tag{9.21}$$

A similar argument applies to (9.20). Minimizing $D(p\|p_0)$ and $D(p\|p_1)$ over ∂B_* can be done either directly with Lagrange mutlipliers, or using Proposition 5.3 in Part 1, where the "energy" function $E : A \to \mathbb{R}$ in the setup of Cramér's theorem is now given by LL as in (9.5), so that the constraint $\langle E \rangle_p = x$ in (5.3), with $x = 0$, now becomes $D(p\|p_1) - D(p\|p_0) = 0$. Either way, let

$$p_\lambda(a) := \frac{p_0(a)^{1-\lambda} p_1(a)^\lambda}{Z_\lambda(p_0, p_1)}, \tag{9.22}$$

where $\lambda \in [0,1]$, define $p_\lambda \in \mathrm{Prob}(A)$. Our notation is consistent, since putting $\lambda = 0$ in (9.22) reproduces the original p_0, and likewise for p_1. Indeed, $\lambda \mapsto p_\lambda$ is a curve from p_0 to p_1. This curve intersects ∂B_* in a unique point, i.e., there is a unique $\overline{\lambda} \in (0,1)$ for which $p_{\overline{\lambda}} \in \partial B_*$, i.e.,

$$D(p_{\overline{\lambda}}\|p_1) = D(p_{\overline{\lambda}}\|p_0). \tag{9.23}$$

Then $p_{\overline{\lambda}}$ minimizes both $p \mapsto D(p\|p_0)$ and $p \mapsto D(p\|p_1)$ over ∂B_* (exercise). Moreover, the very same value of $\overline{\lambda}$ turns out to minimize the function $\lambda \mapsto Z_\lambda(p_0, p_1)$, at which value we have

$$D(p_{\overline{\lambda}}\|p_1) = D(p_{\overline{\lambda}}\|p_0) = -\log Z_{\overline{\lambda}}(p_0, p_1). \tag{9.24}$$

Theorem 9.3 now follows from eqs. (9.19) - (9.20) and (9.24). □

Although in view of (9.24) our new Chernoff information is closely related to the familiar relative entropy D, a different setup of the errors actually leads to D itself. Namely, we define

$$\beta_N^\varepsilon := \inf_{T_N \subset A^N} \{\beta_N(T_N) \mid \alpha_N(T_N) < \varepsilon\}; \qquad (0 < \varepsilon < 1). \tag{9.25}$$

Here we no longer treat H_0 and H_1 symmetrically, but fix an error bound α_N and make the best of β_N given this bound. Here, as before, the idea is that one may make β_N as small as one likes by decreasing T_N (up to $\beta_N(T_N) = 0$ for $T_N = \emptyset$), but this increases α_N (up to $\alpha_N = 1$). But if we put an upper bound on α_N, then β_N should have a positive lower bound. Here is its asymptotic value:

Theorem 9.4 (Stein) *With $\alpha_N(T_N) = p_0^N(T_N^c)$, $\beta_N(T_N) = p_1^N(T_N)$, and β_N^ε as in* (9.25)*, we have*

$$\lim_{N\to\infty} \frac{1}{N} \log \beta_N^\varepsilon = -D(p_0\|p_1). \tag{9.26}$$

Note that the rate function is independent of $\varepsilon \in (0,1)$. The proof is quite different from Theorem 9.3, since H_0 and H_1 now enter asymmetrically (though Proposition 9.2 will appear at the end).

Proof. An upper bound for (9.26) comes from an AEP for the *relative* entropy, cf. (9.11). Recall

$$T_{N,\varepsilon}^E(p) := \{\sigma \in A^N \mid e^{-N(H(p)+\varepsilon)} \le p^N(\sigma) \le e^{-N(H(p)-\varepsilon)}\}, \tag{9.27}$$

which for any $\varepsilon > 0$ carried most of the probability of within A^N relative to p^N: if similar to (9.11) we define $\ell_N(\sigma) = N^{-1} \log p^N(\sigma)$, we have $\ell_N(\sigma) \to -H(p)$ in probability w.r.t. p, that is,

$$\lim_{N\to\infty} p^N(|\ell_N + H(p)| > \varepsilon) = 0, \tag{9.28}$$

for all $\varepsilon > 0$, cf. (7.8), where the plus sign should be read as two minus signs! Similarly, one has $\mathsf{LL}_N \to D(p_0\|p_1)$ with respect to p_0 (and $\mathsf{LL}_N \to -D(p_0\|p_1)$ w.r.t. p_1), i.e., for all $\varepsilon > 0$,

$$\lim_{N\to\infty} p_0^N(|\mathsf{LL}_N - D(p_0\|p_1)| > \varepsilon) = 0. \tag{9.29}$$

This follows from the weak law of large numbers for $E = \mathsf{LL}$, as mentioned after (9.21). Defining

$$\begin{aligned} T_{N,\varepsilon}(p_0,p_1) &:= \{\sigma \in A^N \mid p_1^N(\sigma) e^{N(D(p_0\|p_1)-\varepsilon)} \le p_0^N(\sigma) \le p_1^N(\sigma) e^{N(D(p_0\|p_1)+\varepsilon)}\} \\ &= \{\sigma \in A^N \mid |\mathsf{LL}_N(\sigma) - D(p_0\|p_1)| < \varepsilon\}, \end{aligned} \tag{9.30}$$

it trivially follows from (9.29) that for given $\varepsilon > 0$ and sufficiently large $N = N(\varepsilon)$ we have

$$p_0^N(T_{N,\varepsilon}(p_0,p_1)) > 1 - \varepsilon. \tag{9.31}$$

One also has bounds similar to (7.7) for the AEP, viz.

$$(1-\varepsilon) e^{-N(D(p_0\|p_1)+\varepsilon)} < p_1^N(T_{N,\varepsilon}(p_0,p_1)) < e^{-N(D(p_0\|p_1)-\varepsilon)}. \tag{9.32}$$

Hence for $T_N = T_{N,\varepsilon}(p_0,p_1)$, from (9.31) we have

$$\alpha_N(T_{N,\varepsilon}(p_0,p_1)) = p_0^N(T_{N,\varepsilon}(p_0,p_1)^c) < \varepsilon. \tag{9.33}$$

The upper bound in (9.32) gives

$$\beta_N(T_{N,\varepsilon}(p_0,p_1)) < e^{-N(D(p_0\|p_1)-\varepsilon)}. \tag{9.34}$$

If we replace ε by $0 < \varepsilon' < \varepsilon$, eq. (9.33) gives

$$\alpha_N(T_{N,\varepsilon'}(p_0,p_1)) < \varepsilon' < \varepsilon, \tag{9.35}$$

so that in view of the definition of an infimum, the inequality (9.34) with $\varepsilon \rightsquigarrow \varepsilon'$ gives

$$\beta_N^\varepsilon \leq \frac{1}{N}\log\beta_N(T_{N,\varepsilon'}(p_0,p_1)) < -D(p_0\|p_1)+\varepsilon'. \tag{9.36}$$

Letting $\varepsilon' \to 0$ therefore gives the upper bound

$$\limsup_{N\to\infty}\frac{1}{N}\log\beta_N^\varepsilon \leq -D(p_0\|p_1). \tag{9.37}$$

To get a lower bound we need a different argument.[17] First, for given $\varepsilon \in (0,1)$ and test T_N with $\alpha_N(T_N) < \varepsilon$, and sufficiently large N, there exists a Neyman–Pearson test $T_N^*(u_N)$ for which

$$0 < \alpha_N(T_N) < \alpha_N(T_N^*(u_N)) < \varepsilon. \tag{9.38}$$

Indeed, we have $\alpha_N(T_N^*(u_N)) = p_0^N(\mathsf{LL}_N \leq u_N)$, in which $(\mathsf{LL}_N \leq u_N)$ is the set of all $\sigma \in A^N$ for which $p_0^N(\sigma) \leq e^{u_N}p_1^N(\sigma)$. For sufficiently small $u_N \to -\infty$ this inequality cannot be satisfied, so that $\alpha_N(T_N^*(u_N)) \to 0$. For sufficiently large $u_N \to \infty$ the inequality is always satisfied, so that $\alpha_N(T_N^*(u_N)) \to 1$. In this way, we can steer $\alpha_N(T_N^*(u_N))$ so as to meet (9.38). Replacing $T_N^* \equiv T_N^*(0)$ by $T_N^*(u_N)$ in the proof of Theorem 9.3 gives the asymptotics of $\beta_N(T_N^*(u_N))$. With

$$B_*(u_N) = \{p \in \mathrm{Prob}(A) \mid D(p\|p_1) > D(p\|p_0)+u_N\}; \tag{9.39}$$

$$\partial B_*(u_N) = \{p \in \mathrm{Prob}(A) \mid D(p\|p_1) = D(p\|p_0)+u_N\}, \tag{9.40}$$

cf. (9.15), and $D(\partial B_*(u_N)\|p_1) = \inf\{D(p\|p_1) \mid p \in \partial B_*(u_N)\}$ as usual, we obtain

$$\beta_N(T_N^*(u_N)) = p_1^N(T_N^*(u_N)) \approx e^{-ND(\partial B_*(u_N)\|p_1)}. \tag{9.41}$$

As in the proof of Theorem 9.3,

$$D(\partial B_*(u_N)\|p_1) = D(p_{\tilde\lambda}\|p_1), \tag{9.42}$$

where this time $\tilde\lambda \in (0,1)$ is the unique value for which $D(p_{\tilde\lambda}\|p_1) = D(p_{\tilde\lambda}\|p_0)+u_N$. Now

$$0 \leq D(p_{\tilde\lambda}\|p_1) \leq D(p_0\|p_1), \tag{9.43}$$

as follows from (9.22); the function $\lambda \mapsto D(p_\lambda\|p_1)$ decreases monotonically from $D(p_0\|p_1)$ to $D(p_1\|p_1) = 0$ as λ increases from $\lambda = 0$ to $\lambda = 1$. Hence (9.41) and (9.42) give

$$\liminf_{N\to\infty}\frac{1}{N}\log\beta_N(T_N^*(u_N)) \geq -D(p_0\|p_1). \tag{9.44}$$

Finally, we invoke Proposition 9.2 with $A \rightsquigarrow A^N$ and $p_i \rightsquigarrow p_i^N$, to the effect that given (9.38) we must have $\beta_N(T_N) \geq \beta_N(T_N^*(u_N))$. This strengthens (9.44) to

$$\liminf_{N\to\infty}\frac{1}{N}\log\beta_N^\varepsilon \geq -D(p_0\|p_1). \tag{9.45}$$

Combining this with (9.37) gives (9.26). □

[17] The proof in Cover & Thomas, *Elements of Information Theory*, Second Edition (2006), §11.8, which we have followed up to this point, seems incomplete about the lower bound. It is based on the (correct) fact that if $T_N \subset A^N$ also satisfies $p_0^N(T_N) > 1-\varepsilon$, then provided $D(p_0\|p_1) < \infty$, we also have $p_1^N(T_N) > (1-2\varepsilon)e^{-N(D(p_0\|p_1)+\varepsilon)}$. But this only proves that $\liminf_{N\to\infty}\frac{1}{N}\log\beta_N^\varepsilon \geq -D(p_0\|p_1)-\varepsilon$, where ε is the one in β_N^ε and is not easily removed. See also J.A. Bucklew, *Large Deviation Techniques, Indecision, Simulation, and Estimation* (Wiley), §VI.A.

Exercises

1. Prove Lemma 9.1 (i.e., show that $\alpha(T)+\beta(T) \geq \alpha(T^*)+\beta(T^*)$ for any $T \subset A$).

2. This exercise provides some details of the proof of Theorem 9.3. Assume (9.22).

   (a) Prove (9.13)

   (b) Prove the claim before (9.21). Hint: suppose that $q \in \mathrm{Prob}(A)$ satisfies

   $$D(q\|p_1) \leq D(q\|p_0) \leq D(p\|p_0)$$

   for all $p \in B_*^c$, in which case q attains the desired infimum of $D(p\|p_0)$ over B_c^*. Now consider the line $q_\lambda := (1-\lambda)p_0 + \lambda q$ from p_0 to q, where $\lambda \in [0,1]$. Show that

   $$\frac{d}{d\lambda} D(q_\lambda\|p_0) > 0.$$

   Using the convexity of B_* and B_*^c and the fact that ∂B_c is a hyperplane in $\mathrm{Prob}(A)$, show that the line $\lambda \mapsto q_\lambda$ starts in B_*, then hits ∂B_c, and then enters B_c^*. And finish!

   (c) Show (e.g. using Lagrange multipliers) that there is a unique $\overline{\lambda} \in (0,1)$ such that (9.23) holds, and that $p_{\overline{\lambda}}$ then minimizes both $p \mapsto D(p\|p_0)$ and $p \mapsto D(p\|p_1)$ over ∂B_*.

   (d) Show that $\lambda \in (0,1)$ satisfies (9.23) iff it minimizes $\lambda \mapsto Z_\lambda(p_0,p_1)$.

   (e) Prove (9.24).

3. Prove that $R_0'(p_0,p_1) = -D(p_0\|p_1)$ and $R_1'(p_0,p_1) = D(p_1\|p_0)$, where the prime is $d/d\lambda$.

4. Prove (9.29) and the upper bound in (9.32).

5. Prove (9.43).

10 Quantum hypothesis testing

The classical state discrimination problem of the previous section, namely the question if some unknown $p \in \mathrm{Prob}(A)$ is given by either $p = p_0$ or $p = p_1$, is easily adapted to the quantum setting: we now have a (finite-dimensional) Hilbert space H and ask if an unknown density matrix $\rho \in D(H)$ is given by either $\rho = \rho_0$ or $\rho = \rho_1$. A test,[18] corresponding to a subspace $T \subset A$ in the classical case, is now a (closed) linear subspace $T \subset H$, or, equivalently, the corresponding projection e_T (that is, $e_T\psi = \psi$ iff $\psi \in T$ and $e_T\psi = 0$ iff $\psi \in T^\perp$). We accept H_0, stating that $\rho = \rho_0$, iff a measurement of the operator e_T has outcome 1, and accept H_1 if the outcome is 0; more neatly, we accept H_i iff the outcome of measuring $e_{T^\perp} = 1_H - e_T$ is $i \in \{0,1\}$. Informally,[19] but analogously to the classical setting, this is like randomly drawing some unit vector $\psi \in H$ according to the unknown distribution ρ, and accepting H_0 if $\psi \in T$ and accepting H_1 if $\psi \in T^\perp$. The first case corresponds to the outcome $e_T = 1$, whereas the second is $e_T = 0$ (more generally, the outcome of a measurement of a self-adjoint operator $S \in L(H)$ is one of its eigenvalues λ).[20]

[18] Except for replacing A by H and p by ρ we try to keep the same notation as in the classical case.

[19] In Hilbert space the *orthogonal complement* $T^\perp := \{\psi \in H \mid \langle\psi,\varphi\rangle = 0 \forall \varphi \in T\}$ of some closed linear subspace $H_T \subset H$ plays the role of the set-theoretic complement $B^c = A\backslash B$ of a subset $B \subset A$ in the classical setting. If e_T is the projection onto T, then $1_H - e_T$ is the projection onto $T^\perp$. A big difference between classical and quantum is that whereas $a \in A$ lies in either B or in B^c, a vector $\psi \in H$ need not lie in either T or $T^\perp$ (explain)! The classical analogue of the projection e_T is the indicator (or characteristic) function 1_T on A, which also assumes the values $0,1$ only.

[20] Indeed, the eigenvalues of the projection e_T are 0 and 1, with corresponding eigenspaces $H_0 = T^\perp$ and $H_1 = T$.

Hence by the Born rule of quantum theory false alarm and miss probabilities are given by

$$\alpha(T) = \mathrm{Tr}\,(\rho_0(1_H - e_T)); \qquad \beta(T) = \mathrm{Tr}\,(\rho_1 e_T); \tag{10.1}$$

indeed, the first expression gives the probability that a "draw" ψ from H according to the true state ρ_0 nonetheless lies in $T^\perp$ (which is the outcome confirming H_1), whereas the second is the the probability that we draw some $\psi \in T$ (which confirms H_0) although the true state is ρ_1.

It seems natural to assume that the quantum analogue of a Neyman–Pearson test (9.1), which singles out the region in A where $p_0 > p_1$, is given by the linear subspace of H on which $\rho_0 > \rho_1$, or $\rho_0 - \rho_1 > 0$. Technically, for any self-adjoint $S = S^* \in L(H)$, we define $[S > 0] \subset H$ as the linear span of all $\psi \in H$ for which $S\psi = \lambda\psi$ for some $\lambda > 0$; this is the direct sum of all eigenspaces H_λ of S for which $\lambda > 0$. Its orthogonal complement is $[S > 0]^\perp = [S \leq 0]$, in obvious notation.[21] Thus $[\rho_0 > \rho_1]$ is the linear subspace of H on which $\rho_0 > \rho_1$ in the above sense. This suggestion is confirmed by the following "quantum Neyman–Pearson lemma":[22]

Lemma 10.1 (Holevo, Helstrom) *The total error $\alpha(T) + \beta(T)$ is minimized by the test $\rho_0 > \rho_1$:*

$$\gamma := \inf_{T \subset H} \{\alpha(T) + \beta(T)\} = \alpha(T^*) + \beta(T^*); \qquad T^* = [\rho_0 > \rho_1]. \tag{10.2}$$

As in the classical case, we improve the precision of a test by independently repeating it N times. This corresponds to replacing (H, ρ) by (H^N, ρ^N), and $T \subset H$ by $T_N \subset H^N$, and may again mathematically be seen as a single test, analogous to drawing $\sigma \in A^N$ controlled by $p^N \in \mathrm{Prob}(A^N)$. This leads to satisfactory quantum analogues of Theorem 9.3 and Theorem 9.4. The first involves the obvious quantum analogue of the relative Rényi entropy (9.8), defined by

$$Q(\rho_0, \rho_1) := -\inf_{\lambda \in (0,1)} R_\lambda(\rho_0, \rho_1); \qquad R_\lambda(\rho_0, \rho_1) := \log\left(\mathrm{Tr}\,(\rho_0^{1-\lambda}\rho_1^\lambda)\right), \tag{10.3}$$

where the powers ρ_i^s are defined by the functional calculus, that is, if ρ is given by (2.18), then

$$\rho^s = \sum_i p_i^s |v_i\rangle\langle v_i|. \tag{10.4}$$

Theorem 10.2 *For any $\rho_0, \rho_1 \in D(H)$, the asymptotics of the optimal total error*

$$\gamma_N = \inf_{T_N \subset H^N} \{\mathrm{Tr}\,(\rho_0^N(1_{H^N} - e_{T_N})) + \mathrm{Tr}\,(\rho_1^N e_{T_N})\}, \tag{10.5}$$

i.e. the sum of the false alarm for $\mathsf{H}_0 : \rho = \rho_0$ and the miss error for $\mathsf{H}_1 : \rho = \rho_1$, is given by

$$\lim_{N\to\infty} \frac{1}{N} \log \gamma_N = -Q(\rho_0, \rho_1). \tag{10.6}$$

[21] For any subset $\Delta \subset \sigma(S)$ of the spectrum $\sigma(S)$ of S, which in the finite-dimensional situation considered here is just the set of its eigenvalues, $[S \in \Delta]$ is defined as the direct sum of all eigenspaces H_λ with $\lambda \in \Delta$.

[22] As in the classical case, this is actually a special case of a more general result that concerns more general tests, namely maps $\mathsf{t} : \Omega \to [0, 1_H]$, where $[0, 1_H] \subset L(H)$ consists of all self-adjoint operators whose spectrum lies in $[0, 1]$. One can measure such maps t in a state $\rho \in D(H)$, with outcomes $\omega \in \Omega$, so that according to quantum mechanics the probability of finding ω equals $\mathrm{Tr}\,(\rho\mathsf{t}(\omega))$. Finding ω assigns this probability to (say) H_0. Our simplified setting corresponds to $\Omega = \{0, 1\}$ and the subset of $[0, 1_H]$ consisting of projections. As in the classical case, even in this more general setting the optimal test is in fact a projection. See M. Hayashi, *Quantum Information Theory*, Second Edition (Springer, 2017), §3.2. Our treatment is mainly based on Chapter 3 of this book, simplified as appropriate.

Proof. Unlike the classical case of Chernoff's theorem (but like Stein's lemma), the proof consists of separate arguments for upper and lower bounds on the left-hand side of (10.6).

The upper bound follows from (10.2), which gives an optimal test T_N^*, so that

$$\gamma_N = \mathrm{Tr}\left(\rho_0^N(1_{H^N} - e_{T_N^*})\right) + \mathrm{Tr}\left(\rho_1^N e_{T_N^*}\right); \qquad T_N^* := [\rho_0^N > \rho_1^N], \tag{10.7}$$

and the following pretty nontrivial inequality: for all positive operators $A, B \geq 0$ and $\lambda \in [0,1]$,

$$\mathrm{Tr}\,(Ae_{[A\leq B]}) + \mathrm{Tr}\,(Be_{[A>B]}) \leq \mathrm{Tr}\,(A^{1-\lambda}B^{\lambda}), \tag{10.8}$$

which applies because $e_{[A\leq B]} = 1_H - e_{[A>B]}$. From this, it is a simple exercise to show that

$$\limsup_{N\to\infty} \frac{1}{N} \log \gamma_N \leq -Q(\rho_0, \rho_1). \tag{10.9}$$

The crucial counterpart of (10.8) providing a lower bound on γ_N is the following inequality:

$$\inf_{T\subset H} \{\mathrm{Tr}\,(\rho_0(1_H - e_T)) + \mathrm{Tr}\,(\rho_1 T)\} \geq \tfrac{1}{2} \sum_{a\in A, b\in B} \min\{p_0(a), p_1(b)\} |\langle \upsilon_a, \upsilon'_b\rangle|^2, \tag{10.10}$$

where $A = B = \{1, \ldots, \dim(H)\}$, and the spectral resolutions of ρ_0 and ρ_1 are assumed to be

$$\rho_0 = \sum_{a=1}^{d} p_0(a) |\upsilon_a\rangle\langle\upsilon_a|; \qquad \rho_1 = \sum_{b=1}^{d} p_1(b) |\upsilon'_b\rangle\langle\upsilon'_b|. \tag{10.11}$$

The inequality (10.10) can be proved, as an exercise, without (10.2). Now introduce probability distributions $P_0(\rho_0, \rho_1)$ and $P_1(\rho_0, \rho_1)$ on $A \times B$, where ρ_0 and ρ_1 are labels not arguments, by

$$P_0(a, b|\rho_0, \rho_1) := p_0(a) |\langle \upsilon_a, \upsilon'_b\rangle|^2; \qquad P_1(a, b|\rho_0, \rho_1) := p_1(b) |\langle \upsilon_a, \upsilon'_b\rangle|^2. \tag{10.12}$$

Then the marginal of P_0 on A equals p_0, whilst the marginal of P_1 on B equals p_1. Moreover,

$$D(P_0 \| P_1) = S(\rho_0 \| \rho_1); \tag{10.13}$$

$$R_\lambda(P_0, P_1) = R_\lambda(\rho_0, \rho_1). \tag{10.14}$$

cf. (9.8), (10.3), and (5.12). The point is that we may now rewrite (10.10) as

$$\gamma \geq \tfrac{1}{2}(P_0(P_0 \leq P_1) + P_1(P_0 > P_1)), \tag{10.15}$$

where $P_0 \leq P_1$ is the set $\{(a,b) \in A \times B \mid P_0(a,b) \leq P_1(a,b)\}$, etc. Similarly, the spectral resolutions

$$\rho_0^N = \sum_{\sigma\in A^N} p_0^N(\sigma) |\upsilon_\sigma\rangle\langle\upsilon_\sigma|; \qquad \rho_1^N = \sum_{\tau\in A^N}^{d} p_1^N(\tau) |\upsilon'_\tau\rangle\langle\upsilon'_\tau|, \tag{10.16}$$

where $\upsilon_\sigma = \upsilon_{\sigma_0} \otimes \cdots \otimes \upsilon_{\sigma_{N-1}}$ are eigenvectors of ρ_0^N with eigenvalues $p_0^N(\sigma)$, and likewise for ρ_1, give probability distributions $P_0(\rho_0^N, \rho_1^N)$ and $P_0(\rho_0^N, \rho_1^N)$ on $A^N \times A^N = (A \times A)^N$ as in (10.12):

$$P_0(\sigma, \tau|\rho_0^N, \rho_1^N) := p_0^N(\sigma) |\langle \upsilon_\sigma, \upsilon'_\tau\rangle|^2; \qquad P_1(\sigma, \tau|\rho_0^N, \rho_1^N) := p_1^N(\tau) |\langle \upsilon_\sigma, \upsilon'_\tau\rangle|^2. \tag{10.17}$$

It is an easy exercise to show that

$$P_0(\rho_0^N, \rho_1^N) = P_0(\rho_0, \rho_1)^N; \qquad P_1(\rho_0^N, \rho_1^N) = P_1(\rho_0, \rho_1)^N. \tag{10.18}$$

The arguments from (10.10) onwards may now be repeated for $H \rightsquigarrow H^N$ and $\rho_i \rightsquigarrow \rho_i^N$.

$$\gamma_N \geq \tfrac{1}{2}(P_0^N(P_0^N \leq P_1^N) + P_1^N(P_0^N > P_1^N)), \tag{10.19}$$

where now $P_0^N \leq P_1^N$ is the set $\{(\sigma,\tau) \in A^N \times B^N \mid P_0^N(\sigma,\tau) \leq P_1^N(\sigma,\tau)\}$, etc. The classical case (based on Sanov's theorem), with p_i replaced by P_i, combined with (10.14), then gives

$$\liminf_{N\to\infty} \frac{1}{N} \log \gamma_N \geq -C(P_0,P_1) = -Q(\rho_0,\rho_1). \tag{10.20}$$

Combining (10.20) and (10.9) gives (10.6). □

We now turn to a quantum version of Theorem 9.4 ("Stein's lemma"):

Theorem 10.3 (Hiai–Petz) *With* $\alpha_N(T_N) = \mathrm{Tr}\,(\rho_0^N(1_{H^N} - e_{T_N}))$, $\beta_N(T_N) = \mathrm{Tr}\,(\rho_1^N e_{T_N})$, *and*

$$\beta_N^\varepsilon := \inf_{T_N \subset H^N} \{\beta_N(T_N) \mid \alpha_N(T_N) < \varepsilon\}; \qquad (0 < \varepsilon < 1), \tag{10.21}$$

where $\rho_0, \rho_1 \in D(H)$ *are given, and* T_N *are linear subspaces of* H^N, *we have, cf.* (5.12),

$$\lim_{N\to\infty} \frac{1}{N} \log \beta_N^\varepsilon = -S(\rho_0\|\rho_1). \tag{10.22}$$

Though a clean analogue of the classical Stein lemma, this is more difficult to prove than any previous result, and so we just provide the barest of sketches of the main ideas.[23]

An upper bound on (10.22) can be proved by an AEP argument for the relative entropy similar to the proof of Theorem 9.4, brought into the Hilbert space setting as in the proof of Theorem 8.2, using the obvious versions of (8.24) - (8.25), with ensuing inequalities like (8.26) and (8.27). A different strategy, going back to Hiai and Petz,[24] is to prove a so-called "quantum Sanov theorem":

Lemma 10.4 *There exists a sequence* (T_N) *of linear subspaces of* H^N *(i.e.* $T_N \subset H^N$*) such that*

$$\lim_{N\to\infty} \alpha_N(T_N) = 0; \qquad \lim_{N\to\infty} \frac{1}{N} \log \beta_N(T_N) = -S(\rho_0\|\rho_1). \tag{10.23}$$

The reason for calling this a "quantum Sanov theorem" is that its classical analogue, namely the statement that for any $p_0, p_1 \in \mathrm{Prob}(A)$ there exists a sequence (T_N) of subsets of A^N such that

$$\lim_{N\to\infty} \alpha_N(T_N) = 0; \qquad \lim_{N\to\infty} \frac{1}{N} \log \beta_N(T_N) = -D(p_0\|p_1), \tag{10.24}$$

where $\alpha_N(T) = p_0^N(T_N^c)$ and $\beta_N(T) = p_1^N(T_N)$ as before, follows from Sanov's theorem. Anyway, since this sequence of tests eventually makes $\alpha_N(T_N)$ smaller than any $\varepsilon > 0$, this shows that

$$\limsup_{N\to\infty} \frac{1}{N} \log \beta_N^\varepsilon \leq -S(\rho_0\|\rho_1); \tag{10.25}$$

this is not yet an equality because (T_N) may not reach the infimum in (10.21) - (10.22).

One way to find a lower bound is a proof by contradiction from another lemma:[25]

[23] A complete proof may be found in Hayashi's book, Chapter 3.

[24] F. Hiai and D. Petz, The proper formula for relative entropy and its asymptotics in quantum probability, *Communications in Mathematical Physics* 143, 99–114 (1991). See alsoi M. Ohya & D. Petz, *Quantum Entropy and its Use* (Springer, 1994), Theorem 1.18; Hayashi's book already cited, Lemma 3.6; and I. Bjelaković *et al.*, A quantum version of Sanov's theorem, *Communications in Mathematical Physics* 260, 659–671 (2005).

[25] T. Ogawa & H. Nagaoka, Strong converse and Stein's lemma in quantum hypothesis testing, *IEEE Transactions on Information Theory* 46, 2428–2433 (2002). A nicer proof may be found in Hayashi's book, Lemma 3.7.

Lemma 10.5 (Ogawa & Nagaoka) *If* (T_N), $T_N \subset H^N$, *satisfies*

$$\lim \inf_{N\to\infty} \frac{1}{N} \log \beta_N(T_N) < -S(\rho_0 \| \rho_1), \tag{10.26}$$

then $\limsup_{N\to\infty} \alpha_N(T_N) = 1$.

Since $0 < \varepsilon < 1$ in (10.21) - (10.22), this at once gives

$$\lim \inf_{N\to\infty} \frac{1}{N} \log \beta_N(T_N) \geq -S(\rho_0 \| \rho_1). \tag{10.27}$$

Combined with the upper bound (10.25), this lower bound gives (10.22). □

Exercises

1. Prove (10.2).
2. Prove (10.9) from (10.7) and (10.8).
3. Try to prove (10.8).
4. Prove (10.10).
5. Show that P_0 and P_1 in (10.12) are indeed probability distributions on $\mathsf{d} \times \mathsf{d}$.
6. Prove (10.13) and (10.14).
7. Prove (10.18).
8. Based on Theorem 9.3, give the details of the classical derivation of the inequality in (10.20).