# Advanced Topics in Quantum Information Theory

Lecture notes for CS 798/QIC 890 Spring 2020

John Watrous

*School of Computer Science and Institute for Quantum Computing*
*University of Waterloo*

November 8, 2020

# Lectures

Lecture 1

# Conic Programming

The first topic we will discuss in the course is *conic programming*, which is a valuable tool for the study of quantum information. In particular, *semidefinite programs*, which are a specific type of conic program, have proven to be particularly useful in the theory of quantum information and computation—and you may already be familiar with some of their applications. There is, however, much to be gained in considering conic programs in greater generality. A few points in support of this claim follow.

1. Conic programming offers a formulation through which fundamental concepts in convex analysis may be conveniently expressed and analyzed. These fundamental concepts can offer valuable insights into semidefinite programs that might otherwise be easily obscured by technical details.

2. Certain key properties of semidefinite programs are, in fact, possessed by conic programs defined over a wide variety of convex cones. A noteworthy example is that *Slater's theorem*, which provides a simple-to-check condition for the critically important property of *strong duality* for many semidefinite programs, generalizes to conic programs.

3. While not all properties of semidefinite programs generalize to conic programs, an understanding of conic programming serves to illuminate the specific attributes of the cone of positive semidefinite operators that have allowed for these properties to hold in the semidefinite programming case.

4. The generality offered by conic programming will be of use in this course.

One must appreciate that semidefinite programs do have a very special property that contributes enormously to their utility, which is that one can generally solve a given semidefinite program with reasonable efficiency and precision using a (classical!) computer. Conic programs, in contrast, are in general hard to solve with a computer. For example, maximizing a linear function over the cone $\text{SepD}(\mathbb{C}^n : \mathbb{C}^n)$ of *bipartite separable density operators* with local dimension $n$ is an NP-hard optimization problem, even to approximate with a modest degree of precision.

Having sufficient motivation (I presume) for a study of conic programming, we will now proceed to such a study.

## 1.1 Preliminaries

This preliminary subsection defines various notions and discuss a few known facts (without proofs) that will be needed for a proper treatment of conic programming, mostly relating to convex analysis.

Let $\mathcal{V}$ be a finite-dimensional real inner product space, with the inner product of any two vectors $u, v \in \mathcal{V}$ being denoted $\langle u, v \rangle$. Note that this inner product is necessarily symmetric in its arguments, given that $\mathcal{V}$ is a vector space over the *real* numbers $\mathbb{R}$, which will be our ground field throughout this entire discussion of conic programming.

A subset $\mathcal{C} \subseteq \mathcal{V}$ is *convex* if, for all $u, v \in \mathcal{C}$ and $\lambda \in [0, 1]$, one has

$$\lambda u + (1 - \lambda) v \in \mathcal{C}. \tag{1.1}$$

A subset $\mathcal{K} \subseteq \mathcal{V}$ is a *cone* if, for all $u \in \mathcal{K}$ and $\lambda \geq 0$, one has $\lambda u \in \mathcal{K}$. We will be principally concerned with subsets having both properties simultaneously, which are aptly named *convex cones*. The letters $\mathcal{K}$ and $\mathcal{L}$ are typical names for convex cones, and we will often make the additional assumption that these cones are *closed* when we are discussing conic programs.

The Cartesian product of any two convex sets is convex. Explicitly, if $\mathcal{C}$ and $\mathcal{D}$ are convex, and $(v_0, w_0), (v_1, w_1) \in \mathcal{C} \times \mathcal{D}$ and $\lambda \in [0, 1]$, then

$$\lambda(v_0, w_0) + (1 - \lambda)(v_1, w_1) = (\lambda v_0 + (1 - \lambda) v_1, \lambda w_0 + (1 - \lambda) w_1) \in \mathcal{C} \times \mathcal{D}. \tag{1.2}$$

Similarly, the Cartesian product of any two cones is a cone: if $\mathcal{K}$ and $\mathcal{L}$ are cones, and $(u, v) \in \mathcal{K} \times \mathcal{L}$ and $\lambda \geq 0$, then

$$\lambda(u, v) = (\lambda u, \lambda v) \in \mathcal{K} \times \mathcal{L}. \tag{1.3}$$

The notion of *separating hyperplane* is fundamental within convex analysis. Here is one form of the *separating hyperplane theorem*, which establishes that for any two disjoint, nonempty, convex sets, there is a hyperplane that separates the two convex sets, with one lying within one of the two closed half-spaces defined by the hyperplane and the second set lying within the opposite closed half-space.

**Theorem 1.1** (Separating hyperplane theorem)**.** *Let $\mathcal{V}$ be a finite-dimensional real inner-product space and let $\mathcal{C}$ and $\mathcal{D}$ be nonempty, disjoint, convex subsets of $\mathcal{V}$. There exists a nonzero vector $w \in \mathcal{V}$ and a real number $\gamma \in \mathbb{R}$ such that*

$$\langle w, u \rangle \leq \gamma \leq \langle w, v \rangle \tag{1.4}$$

*for every $u \in \mathcal{C}$ and $v \in \mathcal{D}$. If either of $\mathcal{C}$ or $\mathcal{D}$ is a cone, then there must exist a nonzero vector $w \in \mathcal{V}$ as above for which the inequality (1.4) is true, for all $u \in \mathcal{C}$ and $v \in \mathcal{D}$, when $\gamma = 0$.*

Given any set $\mathcal{A} \subseteq \mathcal{V}$, one defines the *dual cone* to $\mathcal{A}$ as

$$\mathcal{A}^* = \bigl\{ v \in \mathcal{V} \,:\, \langle u, v \rangle \geq 0 \text{ for all } u \in \mathcal{A} \bigr\}. \tag{1.5}$$

The fact that $\mathcal{A}^*$ is indeed a cone, and is also closed and convex, irrespective of the choice of the set $\mathcal{A}$, can be verified. For any two cones $\mathcal{K}, \mathcal{L} \subseteq \mathcal{V}$, it is the case that

$$(\mathcal{K} \times \mathcal{L})^* = \mathcal{K}^* \times \mathcal{L}^*. \tag{1.6}$$

Finally, if $\mathcal{K}$ is a closed, convex cone, then $\mathcal{K}^{**} = \mathcal{K}$.

## 1.2 Definitions

Now suppose that a finite-dimensional real inner-product space $\mathcal{V}$ and a closed, convex cone $\mathcal{K} \subseteq \mathcal{V}$ have been fixed. In addition, let $\mathcal{W}$ be a finite-dimensional real inner product space, let $\phi : \mathcal{V} \rightarrow \mathcal{W}$ be a linear map, and let $a \in \mathcal{V}$ and $b \in \mathcal{W}$ be vectors. These choices of objects define a *conic program*, with which the following optimization problem is associated.

**Optimization Problem 1.2** (Standard Conic Program)

| *Primal problem* | | *Dual problem* | |
|---|---|---|---|
| maximize: | $\langle a, x \rangle$ | minimize: | $\langle b, y \rangle$ |
| subject to: | $\phi(x) = b$ | subject to: | $\phi^*(y) - a \in \mathcal{K}^*$ |
| | $x \in \mathcal{K}$ | | $y \in \mathcal{W}$ |

In the dual problem statement, $\phi^* : \mathcal{W} \rightarrow \mathcal{V}$ denotes the uniquely determined linear map, known as the *adjoint map* to $\phi$, that satisfies $\langle w, \phi(v) \rangle = \langle \phi^*(w), v \rangle$ for all $v \in \mathcal{V}$ and $w \in \mathcal{W}$.

## 1.3 Feasible solutions, optimal values, and weak duality

It is convenient to associate two sets of vectors with Optimization Problem 1.2:

$$\mathcal{A} = \bigl\{x \in \mathcal{K} \,:\, \phi(x) = b\bigr\} \quad \text{and} \quad \mathcal{B} = \bigl\{y \in \mathcal{W} \,:\, \phi^*(y) - a \in \mathcal{K}^*\bigr\} \tag{1.7}$$

are the sets of *primal feasible* and *dual feasible* vectors for that conic program. We also define the *primal optimal* and *dual optimal* values of this conic program as

$$\alpha = \sup_{x \in \mathcal{A}} \langle a, x\rangle \quad \text{and} \quad \beta = \inf_{y \in \mathcal{B}} \langle b, y\rangle, \tag{1.8}$$

respectively. These values may be finite or infinite, and by convention we define $\alpha = -\infty$ in case $\mathcal{A} = \varnothing$ and $\beta = \infty$ in case $\mathcal{B} = \varnothing$.

**Proposition 1.3** (Weak duality for conic programs)**.** *Let $\mathcal{V}$ and $\mathcal{W}$ be finite-dimensional real inner product spaces, let $\mathcal{K} \subseteq \mathcal{V}$ be a closed, convex cone, let $\phi : \mathcal{V} \rightarrow \mathcal{W}$ be a linear map, and let $a \in \mathcal{V}$ and $b \in \mathcal{W}$ be vectors. For $\alpha, \beta \in \mathbb{R} \cup \{-\infty, \infty\}$ as defined in* (1.8) *above, it is the case that $\alpha \leq \beta$.*

*Proof.* If either of the sets $\mathcal{A}$ and $\mathcal{B}$ defined in (1.7) are empty, then the proposition is vacuously true: either $-\infty \leq \beta$ or $\alpha \leq \infty$. It therefore suffices to consider the case in which $\mathcal{A}$ and $\mathcal{B}$ are nonempty.

Suppose $x \in \mathcal{A}$ and $y \in \mathcal{B}$ are chosen arbitrarily. The set $\mathcal{A}$ is a subset of $\mathcal{K}$, so $x \in \mathcal{K}$, and because $y \in \mathcal{B}$ it is the case that $\phi^*(y) - a \in \mathcal{K}^*$, and therefore

$$\langle \phi^*(y) - a, x\rangle \geq 0. \tag{1.9}$$

We may therefore observe the following inequality and chain of equalities:

$$\langle a, x\rangle \leq \langle \phi^*(y), x\rangle = \langle y, \phi(x)\rangle = \langle y, b\rangle = \langle b, y\rangle. \tag{1.10}$$

This inequality is maintained as one takes the supremum over all $x \in \mathcal{A}$ and infimum over $y \in \mathcal{B}$, and therefore $\alpha \leq \beta$, as required. □

## 1.4 Minimization and maximization

In some situations, it may be convenient or natural to take the primal problem to be a minimization rather than a maximization problem. The dual problem then becomes a maximization problem, as follows.

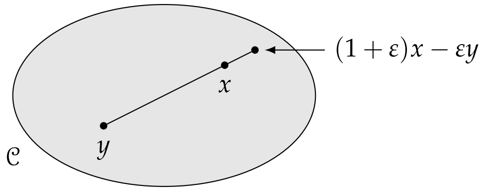

Figure 1.1: An illustration of the point $(1+\varepsilon)x - \varepsilon y$, as it relates to points $x$ and $y$ in a convex set $\mathcal{C}$.

**Optimization Problem 1.4**

| *Primal problem* | | *Dual problem* | |
|---|---|---|---|
| minimize: | $\langle a, x\rangle$ | maximize: | $\langle b, y\rangle$ |
| subject to: | $\phi(x) = b$ | subject to: | $a - \phi^*(y) \in \mathcal{K}^*$ |
| | $x \in \mathcal{K}$ | | $y \in \mathcal{W}$ |

Notice, in particular, that the dual constraint $a - \phi^*(y) \in \mathcal{K}^*$ has replaced the constraint $\phi^*(y) - a \in \mathcal{K}^*$ in Optimization Problem 1.2. The reason for this substitution is that Optimization Problem 1.4 is equivalent, up to a negation of sign, to an instance of Optimization Problem 1.2 in which $a$, $b$, and $\phi$ have been replaced with $-a$, $-b$, and $-\phi$, respectively.

With this equivalence in mind, we shall feel free to minimize or maximize in the primal problem of any conic program as we see fit—naturally selecting the corresponding dual problem formulation—but at the same time we shall lose no generality by adopting Optimization Problem 1.2 as the standard form for conic programs.

## 1.5 Slater's theorem

Now let us state and prove Slater's theorem. To do this, we must refer to the concept of the *relative interior* of a set $\mathcal{S} \subseteq \mathcal{V}$. This is the set denoted $\operatorname{relint}(\mathcal{S})$ that is obtained by taking the interior of the set $\mathcal{S}$, assuming that we have restricted our attention to the smallest affine subspace of $\mathcal{V}$ that contains $\mathcal{S}$.

The relative interior of a convex set $\mathcal{C}$ may be described in the following simple way:

$$\operatorname{relint}(\mathcal{C}) = \bigl\{x \in \mathcal{C} : (\forall y \in \mathcal{C})(\exists \varepsilon > 0)\bigl((1+\varepsilon)x - \varepsilon y \in \mathcal{C}\bigr)\bigr\}. \tag{1.11}$$

Figure 1.1 illustrates how the point $(1+\varepsilon)x - \varepsilon y$ relates to $x$ and $y$.

**Theorem 1.5** (Slater's theorem for conic programs)**.** *Let $\mathcal{V}$ and $\mathcal{W}$ be finite-dimensional real inner-product spaces, let $\mathcal{K} \subseteq \mathcal{V}$ be a closed, convex cone, let $\phi : \mathcal{V} \rightarrow \mathcal{W}$ be a linear map, and let $a \in \mathcal{V}$ and $b \in \mathcal{W}$ be vectors. With respect to the notations $\mathcal{A}$, $\mathcal{B}$, $\alpha$, and $\beta$ defined in Subsection 1.3, the following two statements are true.*

1. *If $\mathcal{B}$ is nonempty and there exists $x \in \operatorname{relint}(\mathcal{K})$ such that $\phi(x) = b$, then there must exist $y \in \mathcal{B}$ such that $\langle b, y \rangle = \alpha$.*
2. *If $\mathcal{A}$ is nonempty and there exists $y \in \mathcal{W}$ for which $\phi^*(y) - a \in \operatorname{relint}(\mathcal{K}^*)$, then there must exist $x \in \mathcal{A}$ such that $\langle a, x \rangle = \beta$.*

*Both statements imply the equality $\alpha = \beta$.*

*Proof.* We will prove just the first statement—the second statement can be proved through a similar technique, or one may conclude that the second statement is true given the first by formulating the dual problem of Optimization Problem 1.2 as the primal problem of a (different but equivalent) conic program. We will also make the simplifying assumptions that $\mathcal{V} = \operatorname{span}(\mathcal{K})$ and $\mathcal{W} = \operatorname{im}(\phi)$, both of which cause no loss of generality. Note that $\alpha$ and $\beta$ are both necessarily finite, as the assumptions of the first statement imply that $\mathcal{A}$ and $\mathcal{B}$ are nonempty.

Define two subsets of $\mathcal{W} \oplus \mathcal{V} \oplus \mathbb{R}$ as follows:

$$\begin{aligned} \mathcal{C} &= \bigl\{ (b - \phi(x), z, \langle a, x \rangle) \,:\, x, z \in \mathcal{V},\ x - z \in \mathcal{K} \bigr\}, \\ \mathcal{D} &= \bigl\{ (0, 0, \eta) \,:\, \eta > \alpha \bigr\}. \end{aligned} \tag{1.12}$$

Both of these sets are evidently convex, and they are disjoint by the definition of $\alpha$, so they are separated by a hyperplane. That is, there must exist a nonzero vector $(y, u, \lambda) \in \mathcal{W} \oplus \mathcal{V} \oplus \mathbb{R}$ such that

$$\bigl\langle (y, u, \lambda), (b - \phi(x), z, \langle a, x \rangle) \bigr\rangle \leq \bigl\langle (y, u, \lambda), (0, 0, \eta) \bigr\rangle, \tag{1.13}$$

or equivalently

$$\langle y, b - \phi(x) \rangle + \langle u, z \rangle + \lambda \langle a, x \rangle \leq \lambda \eta, \tag{1.14}$$

for all $x, z \in \mathcal{V}$ for which $x - z \in \mathcal{K}$ and all $\eta > \alpha$. Let us observe that there is no loss of generality in assuming $\lambda \in \{-1, 0, 1\}$, as the inequality (1.14) remains true when the vector $(y, u, \lambda)$ is rescaled (i.e., multiplied by any positive real number).

We will now draw several conclusions from the fact that (1.14) holds for all $x, z \in \mathcal{V}$ for which $x - z \in \mathcal{K}$ and all $\eta > \alpha$.

1. The inequality (1.14) must be true when $x = 0$ and $z = 0$, and therefore

   $$\langle y, b \rangle \leq \lambda \eta \tag{1.15}$$

   for all $\eta > \alpha$. This implies that $\lambda \geq 0$, for otherwise the right-hand side of the inequality tends to $-\infty$ as $\eta$ becomes large while the left-hand side remains fixed. Thus, $\lambda = -1$ is impossible, so $\lambda \in \{0, 1\}$.

2. For any choice of $x \in \mathcal{A}$ and $z \in -\mathcal{K}$, it is the case that $x - z \in \mathcal{K}$. Substituting these vectors into the inequality (1.14) and rearranging yields

   $$\langle u, z \rangle \leq \lambda(\eta - \langle a, x \rangle) \tag{1.16}$$

   for every $\eta > \alpha$. We conclude that $u \in \mathcal{K}^*$, for otherwise the left-hand side of the above inequality can be made to approach $\infty$ while the right-hand side remains bounded, for any fixed $\eta > \alpha$, through an appropriate selection of $z \in -\mathcal{K}$.

3. Assume toward contradiction that $\lambda = 0$. Fix any choice of $x \in \mathcal{A} \cap \text{relint}(\mathcal{K})$, which is possible by the assumption of the statement being proved. The inequality (1.14) simplifies to

   $$\langle u, z \rangle \leq 0 \tag{1.17}$$

   for every choice of $z \in \mathcal{V}$ for which $x - z \in \mathcal{K}$.

   Setting $z = x$, we have that $x - z \in \mathcal{K}$, and therefore

   $$\langle u, x \rangle \leq 0. \tag{1.18}$$

   As $x \in \mathcal{K}$ and $u \in \mathcal{K}^*$, we conclude that $\langle u, x \rangle = 0$.

   On the other hand, for an arbitrarily chosen vector $v \in \mathcal{K}$, there must exist $\varepsilon > 0$ such that

   $$x - \varepsilon(v - x) = (1 + \varepsilon)x - \varepsilon v \in \mathcal{K}, \tag{1.19}$$

   by virtue of the fact that $x$ is in the relative interior of $\mathcal{K}$. Setting $z = \varepsilon v - \varepsilon x$, one therefore has that $x - z \in \mathcal{K}$, and so

   $$\varepsilon\langle u, v \rangle = \varepsilon\langle u, v \rangle - \varepsilon\langle u, x \rangle = \langle u, z \rangle \leq 0. \tag{1.20}$$

   As it was was for $x$, we find that $\langle u, v \rangle = 0$. Seeing that this is true for all $v \in \mathcal{K}$, and recognizing that $u \in \mathcal{V} = \text{span}(\mathcal{K})$, we conclude that $u = 0$.

   But if $\lambda = 0$ and $u = 0$, then we may free the vector $x$ to range over all of $\mathcal{V}$ and set $z = x$ to conclude from (1.14) that

   $$\langle y, b - \phi(x) \rangle \leq 0 \tag{1.21}$$

   for every $x \in \mathcal{V}$. Bearing in mind the assumption that $\mathcal{W} = \text{im}(\phi)$, we conclude that $y = 0$.

   This, however, is a contradiction to the assumption that $(y, u, \lambda)$ is nonzero. One concludes that $\lambda = 1$.

The steps just describe have allowed us to conclude that there exist vectors $y \in \mathcal{W}$ and $u \in \mathcal{K}^*$ such that

$$\langle y, b - \phi(x) \rangle + \langle u, z \rangle \leq \eta - \langle a, x \rangle \tag{1.22}$$

for all $x, z \in \mathcal{V}$ for which $x - z \in \mathcal{K}$ and all $\eta > \alpha$. Setting $z = 0$ and flipping sign, we find that

$$\langle \phi^*(y) - a, x \rangle \geq \langle y, b \rangle - \eta \tag{1.23}$$

for all $x \in \mathcal{K}$ and $\eta > \alpha$. This implies

$$\phi^*(y) - a \in \mathcal{K}^*, \tag{1.24}$$

for otherwise the left-hand side of the previous inequality can be made to approach $-\infty$ while the right-hand side remains fixed. The vector $y$ is therefore a dual-feasible point: $y \in \mathcal{B}$. Finally, again considering the possibility that $x = 0$ and $z = 0$, we conclude that $\langle b, y \rangle \leq \eta$ for all $\eta > \alpha$. It is therefore the case that $\langle b, y \rangle \leq \alpha$, and hence $\langle b, y \rangle = \alpha$ by weak duality. This concludes the proof (of the first statement). □

## 1.6 Example: conic program for optimal measurements

In this section we will discuss an example of a conic program that is relevant to quantum information. We'll begin with a somewhat general example, or perhaps a category of examples, and then discuss a specific, concrete example.

Suppose $\mathcal{X}$ is a complex Euclidean space and $\mathcal{K} \subseteq \text{Herm}(\mathcal{X})$ is a closed, convex cone. Suppose further that $H_1, \ldots, H_n \in \text{Herm}(\mathcal{X})$, and consider this optimization problem.

$$\begin{aligned} \text{maximize:} \quad & \langle H_1, X_1 \rangle + \cdots + \langle H_n, X_n \rangle \\ \text{subject to:} \quad & X_1 + \cdots + X_n = \mathbb{1}_{\mathcal{X}} \\ & X_1, \ldots, X_n \in \mathcal{K} \end{aligned}$$

This is essentially an *optimal measurement* problem, where $X_1, \ldots, X_n$ represent measurement operators; these operators must sum to the identity as usual, but in place of the usual constraint on measurement operators being positive semidefinite, we are instead constraining them to the cone $\mathcal{K}$. By choosing $\mathcal{K}$ to be the cone of positive semidefinite operators $\text{Pos}(\mathcal{X})$, which is closed and convex, we naturally obtain the ordinary optimal measurement problem, but we can consider any closed, convex cone $\mathcal{K}$ we choose. For example, we may take $\mathcal{K}$ to be the cone of *separable operators* when $\mathcal{X} = \mathcal{Y} \otimes \mathcal{Z}$ is a bipartite tensor product space.

We can set this problem up as a conic program as follows. First, observe that $\mathcal{K}^n$ is a closed, convex cone. The problem may therefore be expressed as the primal problem of a conic program:

$$\begin{aligned}
\text{maximize:}\quad & \big\langle (H_1,\ldots,H_n),(X_1,\ldots,X_n)\big\rangle \\
\text{subject to:}\quad & \phi(X_1,\ldots,X_n) \overset{\diamond}{=} X_1+\cdots+X_n = \mathbb{1}_{\mathcal{X}} \\
& (X_1,\ldots,X_n)\in\mathcal{K}^n
\end{aligned}$$

The symbol $\overset{\diamond}{=}$ indicates that this is the definition of the function $\phi$, whereas the ordinary equal sign represents a constraint.

Here is the dual problem, as is dictated by Optimization Problem 1.2:

$$\begin{aligned}
\text{minimize:}\quad & \big\langle \mathbb{1}_{\mathcal{X}}, Y\big\rangle \\
\text{subject to:}\quad & \phi^*(Y)-(H_1,\ldots,H_n)\in\big(\mathcal{K}^n\big)^* \\
& Y\in\operatorname{Herm}(\mathcal{X})
\end{aligned}$$

We can simplify this by observing that $\phi^*(Y) = (Y,\ldots,Y)$ and $(\mathcal{K}^n)^* = (\mathcal{K}^*)^n$, and naturally replacing the inner-product with the identity operator as the trace in the objective function. The following problem is obtained.

$$\begin{aligned}
\text{minimize:}\quad & \operatorname{Tr}(Y) \\
\text{subject to:}\quad & Y-H_1\in\mathcal{K}^* \\
& \qquad\vdots \\
& Y-H_n\in\mathcal{K}^* \\
& Y\in\operatorname{Herm}(\mathcal{X})
\end{aligned}$$

In summary, the following conic program, expressed in a simplified form, has been obtained.

**Optimization Problem 1.6**

*Primal problem*

$$\begin{aligned}
\text{maximize:}\quad & \langle H_1,X_1\rangle+\cdots+\langle H_n,X_n\rangle \\
\text{subject to:}\quad & X_1+\cdots+X_n=\mathbb{1}_{\mathcal{X}} \\
& X_1,\ldots,X_n\in\mathcal{K}
\end{aligned}$$

*Dual problem*

$$\begin{aligned}
\text{minimize:}\quad & \operatorname{Tr}(Y) \\
\text{subject to:}\quad & Y-H_1\in\mathcal{K}^* \\
& \qquad\vdots \\
& Y-H_n\in\mathcal{K}^* \\
& Y\in\operatorname{Herm}(\mathcal{X})
\end{aligned}$$

Now let us consider a specific instance of this conic program. Define four pure states $|\psi_1\rangle, \ldots, |\psi_4\rangle \in \mathbb{C}^4 \otimes \mathbb{C}^4$ as follows:

$$\begin{aligned}
|\psi_1\rangle &= \frac{1}{2}\bigl(|0\rangle|0\rangle + |1\rangle|1\rangle + |2\rangle|2\rangle + |3\rangle|3\rangle\bigr) = \frac{1}{2}\operatorname{vec}(\mathbb{1} \otimes \mathbb{1}),\\
|\psi_2\rangle &= \frac{1}{2}\bigl(|0\rangle|3\rangle + |1\rangle|2\rangle + |2\rangle|1\rangle + |3\rangle|0\rangle\bigr) = \frac{1}{2}\operatorname{vec}(\sigma_x \otimes \sigma_x),\\
|\psi_3\rangle &= \frac{1}{2}\bigl(|0\rangle|3\rangle + |1\rangle|2\rangle - |2\rangle|1\rangle - |3\rangle|0\rangle\bigr) = \frac{i}{2}\operatorname{vec}(\sigma_y \otimes \sigma_x),\\
|\psi_4\rangle &= \frac{1}{2}\bigl(|0\rangle|1\rangle + |1\rangle|0\rangle - |2\rangle|3\rangle - |3\rangle|2\rangle\bigr) = \frac{1}{2}\operatorname{vec}(\sigma_z \otimes \sigma_x).
\end{aligned}$$

These states were identified by Yu, Duan, and Ying (2012), who proved that they cannot be perfectly discriminated by a PPT measurement. Cosentino (2013) subsequently showed that the optimal PPT discrimination probability is 7/8, by solving the associated semidefinite program.

We will prove that no *separable measurement* can discriminate these states with probability greater than 3/4, assuming that one of the four states is selected uniformly at random. This is easily achievable by coarse-graining a measurement with respect to the standard basis, so this is in fact the optimal probability to correctly discriminate the states by a separable measurement. Here is a precise description of the corresponding conic program.

**Optimization Problem 1.7**

*Primal problem*

$$\begin{aligned}
\text{maximize:}\quad & \frac{1}{4}\langle\psi_1|X_1|\psi_1\rangle + \cdots + \frac{1}{4}\langle\psi_4|X_4|\psi_4\rangle\\
\text{subject to:}\quad & X_1 + \cdots + X_4 = \mathbb{1}_4 \otimes \mathbb{1}_4\\
& X_1, \ldots, X_4 \in \operatorname{Sep}(\mathbb{C}^4 : \mathbb{C}^4)
\end{aligned}$$

*Dual problem*

$$\begin{aligned}
\text{minimize:}\quad & \operatorname{Tr}(Y)\\
\text{subject to:}\quad & Y - \frac{1}{4}|\psi_k\rangle\langle\psi_k| \in \operatorname{Sep}(\mathbb{C}^4 : \mathbb{C}^4)^* \quad (1 \leq k \leq 4)\\
& Y \in \operatorname{Herm}(\mathbb{C}^4 \otimes \mathbb{C}^4)
\end{aligned}$$

Our goal will be to describe a dual-feasible solution having objective value 3/4, for this will then be an upper-bound on the probability of a correct discrimination by weak duality.

Note that the dual-feasibility of a given $Y \in \mathrm{Herm}(\mathbb{C}^4 \otimes \mathbb{C}^4)$ is equivalent to

$$Y - \frac{1}{16}\operatorname{vec}(U_k)\operatorname{vec}(U_k)^* \in \mathrm{Sep}(\mathbb{C}^4 : \mathbb{C}^4)^* \tag{1.25}$$

for

$$U_1 = \mathbb{1} \otimes \mathbb{1}, \quad U_2 = \sigma_x \otimes \sigma_x, \quad U_3 = i\sigma_y \otimes \sigma_x, \quad U_4 = \sigma_z \otimes \sigma_x. \tag{1.26}$$

In order to prove the dual-feasibility of a specific choice for $Y$ that will be specified shortly, we will make use of the following lemma.

**Lemma 1.8** (Breuer–Hall)**.** *Let $U, V \in \mathrm{U}(\mathbb{C}^n)$ be unitary operators such that $V^{\mathsf{T}}U$ is anti-symmetric: $(V^{\mathsf{T}}U)^{\mathsf{T}} = -V^{\mathsf{T}}U$. The operator*

$$Z = \mathbb{1}_n \otimes \mathbb{1}_n - \operatorname{vec}(U)\operatorname{vec}(U)^* - (\mathrm{T} \otimes \mathbb{1})\big(\operatorname{vec}(V)\operatorname{vec}(V)^*\big) \tag{1.27}$$

*is contained in* $\mathrm{Sep}(\mathbb{C}^n : \mathbb{C}^n)^*$.

*Proof.* For any unit vector $z \in \mathbb{C}^n$, we find that

$$(\mathbb{1} \otimes z)^* Z (\mathbb{1} \otimes z) = \mathbb{1} - U\overline{z}z^{\mathsf{T}}U^* - \overline{V}zz^*V^{\mathsf{T}} \geq 0. \tag{1.28}$$

This follows from the observation that the vectors $U\overline{z}$ and $\overline{V}z$ must be orthogonal unit vectors:

$$\begin{aligned} \big\langle \overline{V}z, U\overline{z} \big\rangle &= z^* V^{\mathsf{T}} U \overline{z} = \big\langle zz^{\mathsf{T}}, V^{\mathsf{T}}U \big\rangle \\ &= \big\langle (zz^{\mathsf{T}})^{\mathsf{T}}, (V^{\mathsf{T}}U)^{\mathsf{T}} \big\rangle = -\big\langle zz^{\mathsf{T}}, V^{\mathsf{T}}U \big\rangle = 0. \end{aligned} \tag{1.29}$$

It follows that

$$\langle yy^* \otimes zz^*, Z \rangle = (y \otimes z)^* Z (y \otimes z) \geq 0 \tag{1.30}$$

for every $y \in \mathbb{C}^n$. The required containment follows by convexity. □

Now define $V = \sigma_y \otimes \sigma_z$, and observe that $V^{\mathsf{T}}U_1$, $V^{\mathsf{T}}U_2$, $V^{\mathsf{T}}U_3$, and $V^{\mathsf{T}}U_4$ are all anti-symmetric. By the Breuer–Hall lemma, the operator

$$Y = \frac{1}{16}\big(\mathbb{1}_4 \otimes \mathbb{1}_4 - (\mathrm{T} \otimes \mathbb{1})(\operatorname{vec}(V)\operatorname{vec}(V)^*)\big) \tag{1.31}$$

is dual-feasible. Given that $\mathrm{Tr}(Y) = (16 - 4)/16 = 3/4$, we have obtained the claimed upper-bound on correctly discriminating these states by a separable measurement.

Lecture 2

# Max-relative entropy and conditional min-entropy

In this lecture we will first define the *max-relative entropy* and observe some of its properties. We will then define the *conditional min-entropy* in terms of the quantum max-relative entropy, derive an alternative characterization of this quantity, and consider the conditional min-entropy of a few example classes of states.

Before proceeding to the definition of the max-relative entropy, it will be helpful to consider the ordinary quantum relative entropy and its relationship to the conditional quantum entropy as a source of inspiration. Recall that the quantum relative entropy is defined as follows for all density operators $\rho$ and all positive semidefinite operators $Q$ acting on the same complex Euclidean space:

$$\mathrm{D}(\rho\|Q) = \begin{cases} \operatorname{Tr}(\rho\log\rho) - \operatorname{Tr}(\rho\log Q) & \operatorname{im}(\rho) \subseteq \operatorname{im}(Q) \\ \infty & \operatorname{im}(\rho) \not\subseteq \operatorname{im}(Q). \end{cases} \tag{2.1}$$

We can define this function more generally for any positive semidefinite operator $P$ in place of the density operator $\rho$, but our focus will be on the case where the first argument is a density operator.

One way to think about the quantum relative entropy is that it represents the *loss of efficiency*, measured in bits, that is incurred when one plans ahead for $Q$ but receives $\rho$ instead. This is highly informal, and should not be taken too seriously, but we will allow this intuitive description to suggest some useful terminology: we will refer to the second argument $Q$ in the quantum relative entropy as the *model*, and to the first argument $\rho$ as the *actual state*, for the sake of convenience.

Irrespective of how we choose to interpret the quantum relative entropy function, there is no denying its enormous utility as a "helper function," through which fundamental entropic quantities may be defined and analyzed. In particular, the conditional quantum entropy and the quantum mutual information are defined in terms of the quantum relative entropy as follows:

$$\begin{aligned} \mathrm{H}(\mathsf{X}\,|\,\mathsf{Y})_\rho &= -\,\mathrm{D}(\rho\,\|\,\mathbb{1}_{\mathcal{X}}\otimes\rho[\mathsf{Y}]), \\ \mathrm{I}(\mathsf{X}:\mathsf{Y})_\rho &= \mathrm{D}(\rho\,\|\,\rho[\mathsf{X}]\otimes\rho[\mathsf{Y}]), \end{aligned} \tag{2.2}$$

for all $\rho \in \mathrm{D}(\mathcal{X}\otimes\mathcal{Y})$. Then, though properties of the quantum relative entropy, one may establish important properties of the conditional quantum entropy and quantum mutual information. For example, through the *joint convexity* of quantum relative entropy,

$$\mathrm{D}(\lambda\rho_0 + (1-\lambda)\rho_1\,\|\,\lambda Q_0 + (1-\lambda)Q_1) \leq \lambda\,\mathrm{D}(\rho_0\,\|\,Q_0) + (1-\lambda)\,\mathrm{D}(\rho_1\,\|\,Q_1), \tag{2.3}$$

one may prove the critically important *strong subadditivity* property of von Neumann entropy, which may be expressed as

$$\mathrm{H}(\mathsf{X}|\mathsf{Y},\mathsf{Z})_\rho \leq \mathrm{H}(\mathsf{X}|\mathsf{Y})_\rho \tag{2.4}$$

for every $\rho \in \mathrm{D}(\mathcal{X}\otimes\mathcal{Y}\otimes\mathcal{Z})$.

The quantum relative entropy, and the entropic quantities it defines, tell us a great deal about the so-called *i.i.d. limit*, where an increasing number of independent copies of a given state are made available. In contrast, when our interest is in the so-called *one-shot* setting, where our concern is primarily with a single copy of a given state, the quantum relative entropy and the quantities it defines have limited value.

## 2.1 Quantum max-relative entropy

The *quantum max-relative entropy* (or just max-relative entropy for short) offers an alternative to the ordinary quantum relative entropy that is relevant in the one-shot setting. While it is a different function from the quantum relative entropy, it does possess some of the same general characteristics that make the quantum relative entropy function useful. As we will see in a couple of lectures, the ordinary quantum relative entropy can in fact be recovered from the max-relative entropy (or, to be more precise, a *smoothed* version of max-relative entropy) by applying it in the i.i.d. limit.

**Definition 2.1** (Quantum max-relative entropy)**.** For a density operator $\rho$ and a positive semidefinite operator $Q$ acting on the same complex Euclidean space, the *quantum max-relative entropy* of $\rho$ with respect to $Q$ is defined as follows:

$$\mathrm{D}_{\max}(\rho\,\|\,Q) = \inf\{\lambda \in \mathbb{R} \,:\, \rho \leq 2^{\lambda} Q\}. \tag{2.5}$$

**Remark 2.2.** The same definition may be used when $\rho$ is any positive semidefinite operator, and not necessarily a density operator. It is common, in particular, that $\rho$ is taken to be a sub-normalized state, meaning $\rho \geq 0$ and $\operatorname{Tr}(\rho) \leq 1$. In this course, however, we will focus on the case that $\rho$ is normalized.

Let us first observe the equivalence of the following statements:

1. $\mathrm{D}_{\max}(\rho \| Q) < \infty$
2. $\operatorname{im}(\rho) \subseteq \operatorname{im}(Q)$ (or, equivalently, $\ker(Q) \subseteq \ker(\rho)$)
3. $\mathrm{D}(\rho \| Q) < \infty$

In particular, the max-relative entropy is finite if and only if the ordinary quantum relative entropy is finite.

We may also observe that the max-relative entropy can be expressed through a semidefinite program. More specifically, the max-relative entropy is the *logarithm* of the optimal value of the following semidefinite program.

**Optimization Problem 2.3** (SDP for max-relative entropy)

$$\begin{array}{rl@{\qquad\qquad}rl}
 & \textit{Primal problem} & & \textit{Dual problem}\\
\text{minimize:} & \eta & \text{maximize:} & \langle \rho, X\rangle\\
\text{subject to:} & \rho \leq \eta Q & \text{subject to:} & \langle Q, X\rangle \leq 1\\
 & \eta \geq 0 & & X \in \operatorname{Pos}(\mathcal{X})
\end{array}$$

Alternatively, the max-relative entropy is the *negative logarithm* of the optimal value of the following semidefinite program.

**Optimization Problem 2.4** (Reciprocal SDP for max-relative entropy)

$$\begin{array}{rl@{\qquad\qquad}rl}
 & \textit{Primal problem} & & \textit{Dual problem}\\
\text{maximize:} & \mu & \text{minimize:} & \langle Q, Y\rangle\\
\text{subject to:} & \mu\rho \leq Q & \text{subject to:} & \langle \rho, Y\rangle \geq 1\\
 & \mu \geq 0 & & Y \in \operatorname{Pos}(\mathcal{X})
\end{array}$$

Notice that all four of the problems just suggested are strictly feasible when $\operatorname{im}(\rho) \subseteq \operatorname{im}(Q)$. Slater's theorem therefore implies that strong duality holds under this assumption for both semidefinite programs, with optimal values always being achieved in all four problems. Strong duality also holds when $\operatorname{im}(\rho) \not\subseteq \operatorname{im}(Q)$; in this case the optimal value of both the primal and dual forms in Optimization Problem 2.3 is positive infinity, while the optimal value of both the primal and dual forms in Optimization Problem 2.4 is zero.

## Two alternative characterizations of max-relative entropy

We will now take moment to observe two alternative characterizations of the max-relative entropy. For the first, observe that if $\operatorname{im}(\rho) \subseteq \operatorname{im}(Q)$, then the condition $\rho \leq 2^{\lambda} Q$ is equivalent to

$$\left\| \sqrt{Q^{+}}\rho\sqrt{Q^{+}} \right\| \leq 2^{\lambda}. \tag{2.6}$$

Therefore, we have

$$\mathrm{D}_{\max}(\rho\|Q) = \begin{cases} \log\left(\left\| \sqrt{Q^{+}}\rho\sqrt{Q^{+}} \right\|\right) & \operatorname{im}(\rho) \subseteq \operatorname{im}(Q) \\ \infty & \operatorname{im}(\rho) \not\subseteq \operatorname{im}(Q). \end{cases} \tag{2.7}$$

We may alternatively write

$$\mathrm{D}_{\max}(\rho\|Q) = \log\left(\left\| Q^{-1/2}\rho Q^{-1/2} \right\|\right), \tag{2.8}$$

with the somewhat informal understanding that the expression evaluates to $\infty$ in case $\operatorname{im}(\rho) \not\subseteq \operatorname{im}(Q)$.

The second alternative characterization of the max-relative entropy begins with the observation that the condition $\rho \leq 2^{\lambda} Q$ is equivalent to $\langle \rho, Z \rangle \leq 2^{\lambda} \langle Q, Z \rangle$ for all positive definite operators $Z$. Therefore, assuming $Q \neq 0$, we find that

$$\mathrm{D}_{\max}(\rho\|Q) = \sup_{Z>0} \log\left(\frac{\langle \rho, Z \rangle}{\langle Q, Z \rangle}\right). \tag{2.9}$$

## Interpretation of max-relative entropy

One simple and intuitive way to think about the max-relative entropy $\mathrm{D}_{\max}(\rho\|Q)$ is as follows. Suppose that one attempts to express $Q$ as a nonnegative linear combination of $\rho$ along with any other collection of positive semidefinite operators. We can amalgamate the other positive semidefinite operators and associated nonnegative scalars into a single positive semidefinite operator $R$, for the sake of focusing on the relationship between $\rho$ and $Q$, and we obtain an expression like this:

$$Q = \eta\rho + R \qquad (\text{where } R \geq 0). \tag{2.10}$$

The largest that the value $\eta$ can be, assuming we are free to choose $R$ however we wish, is precisely $2^{-\mathrm{D}_{\max}(\rho\|Q)}$.

If $Q = \sigma$ is itself a density operator, then necessarily $\eta \in [0,1]$, and we may think of this value as being a probability. The simple fact that $\eta \leq 1$ immediately yields a variant of Klein's inequality for the max-relative entropy:

$$\mathrm{D}_{\max}(\rho\|\sigma) \geq 0, \tag{2.11}$$

with equality if and only if $\rho = \sigma$. If, on the other hand, $\sigma$ is "highly dissimilar" to $\rho$, then any convex combination involving $\rho$ and yielding $\sigma$ must take the probability $\eta$ associated with $\rho$ to be small, so $\mathrm{D}_{\max}(\rho\|\sigma)$ must be large. In the extreme case that $\operatorname{im}(\rho) \not\subseteq \operatorname{im}(Q)$, then any expression of $Q$ taking the form (2.10) must have $\eta = 0$, which is consistent with $\mathrm{D}_{\max}(\rho\|\sigma) = \infty$.

## Monotonicity of max-relative entropy

Next let us observe that the max-relative entropy is monotonic with respect to the action of channels, meaning that

$$\mathrm{D}_{\max}(\Phi(\rho)\|\Phi(Q)) \leq \mathrm{D}_{\max}(\rho\|Q) \tag{2.12}$$

for all $\rho \in \mathrm{D}(\mathcal{X})$, $Q \in \operatorname{Pos}(\mathcal{X})$, and $\Phi \in \mathrm{C}(\mathcal{X}, \mathcal{Y})$. In fact, complete positivity is not required; the inequality (2.12) holds for all $\Phi$ positive and trace preserving.

Before we prove that the max-relative entropy is monotonic in the sense just described, let us noted that we cannot follow a similar route to this fact that we followed when proving the analogous fact for the ordinary quantum relative entropy in CS 766/QIC 820—which was through the joint convexity of quantum relative entropy. This is because *the max-relative entropy is not jointly convex*—and this is a sense in which it differs from the ordinary quantum relative entropy. The max-relative entropy is, however, *jointly quasi-convex*:

$$\mathrm{D}_{\max}\left(\sum_{k=1}^{n} p_k \rho_k \,\middle\|\, \sum_{k=1}^{n} p_k Q_k\right) \leq \max_{k\in\{1,\ldots,n\}} \mathrm{D}_{\max}(\rho_k\|Q_k). \tag{2.13}$$

The fact that the max-relative entropy is monotonic with respect to the action of channels, however, is not only true but is almost immediate from the definition of the max-relative entropy. Specifically, if we have $\rho \leq 2^{\lambda} Q$ for some choice of $\lambda$, then $\Phi(\rho) \leq 2^{\lambda}\Phi(Q)$ by the positivity of $\Phi$, from which (2.12) follows. The assumption that $\Phi$ preserves trace implies that $\operatorname{Tr}(\Phi(\rho)) = 1$, so that it is a suitable first argument to the max-relative entropy—but this assumption can be dropped altogether, provided that we're willing to allow $\Phi(\rho)$ as a first argument to the max-relative entropy function.

## Max-relative entropy upper-bounds relative entropy

One can prove that the max-relative entropy is at least as large as the ordinary quantum relative entropy, meaning

$$\mathrm{D}(\rho\|Q) \leq \mathrm{D}_{\max}(\rho\|Q) \tag{2.14}$$

for all density operators $\rho$ and all positive semidefinite operators $Q$.

One way to prove this is to use the fact that the logarithm is an *operator monotone* function: for all positive definite operators $P$ and $Q$ with $P \leq Q$, it is the case that $\log(P) \leq \log(Q)$. This is not a trivial fact to prove, but it is well-known, and you should have no trouble finding a proof if you search for one.

Now, suppose that $\lambda$ satisfies $\rho \leq 2^{\lambda} Q$, or equivalently $2^{-\lambda}\rho \leq Q$. We then have

$$\mathrm{D}(\rho \| Q) = \operatorname{Tr}(\rho \log \rho) - \operatorname{Tr}(\rho \log Q) \leq \operatorname{Tr}(\rho \log \rho) - \operatorname{Tr}\bigl(\rho \log\bigl(2^{-\lambda}\rho\bigr)\bigr) = \lambda, \tag{2.15}$$

and the relation (2.14) follows by minimizing over $\lambda$.

## Max-relative entropy of tensor products and block operators

The max-relative entropy is additive with respect to tensor products:

$$\mathrm{D}_{\max}\bigl(\rho_0 \otimes \rho_1 \,\big\|\, Q_0 \otimes Q_1\bigr) = \mathrm{D}_{\max}(\rho_0 \| Q_0) + \mathrm{D}_{\max}(\rho_1 \| Q_1). \tag{2.16}$$

(The ordinary relative entropy is also additive with respect to tensor products in the same way.) The characterization

$$\mathrm{D}_{\max}(\rho \| Q) = \begin{cases} \log\Bigl(\Bigl\| \sqrt{Q^{+}}\rho\sqrt{Q^{+}} \Bigr\|\Bigr) & \operatorname{im}(\rho) \subseteq \operatorname{im}(Q) \\ \infty & \operatorname{im}(\rho) \not\subseteq \operatorname{im}(Q) \end{cases} \tag{2.17}$$

offers an easy route to a proof of this fact. Observe in particular that this implies that, for every choice of $\rho$, $Q$, and a positive integer $n$, we have

$$\mathrm{D}_{\max}\bigl(\rho^{\otimes n} \,\big\|\, Q^{\otimes n}\bigr) = n\, \mathrm{D}_{\max}(\rho \| Q). \tag{2.18}$$

The max-relative entropy also obeys the following identity, for any choice of density operator $\rho_1, \ldots, \rho_n$, positive semidefinite operators $Q_1, \ldots, Q_n$, and a probability vector $(p_1, \ldots, p_n)$:

$$\mathrm{D}_{\max}\left( \sum_{k=1}^{n} p_k |k\rangle\langle k| \otimes \rho_k \,\middle\|\, \sum_{k=1}^{n} p_k |k\rangle\langle k| \otimes Q_k \right) = \max_{k \in \{1,\ldots,n\}} \mathrm{D}_{\max}(\rho_k \| Q_k). \tag{2.19}$$

Using the formula $\mathrm{D}_{\max}(\rho \| Q) = \mathrm{D}_{\max}(\rho \| \lambda Q) - \log(\eta)$, we obtain this formula for the situation in which the probabilities $p_1, \ldots, p_n$ are not included in the blocks of the second operator:

$$\begin{aligned} &\mathrm{D}_{\max}\left( \sum_{k=1}^{n} p_k |k\rangle\langle k| \otimes \rho_k \,\middle\|\, \sum_{k=1}^{n} |k\rangle\langle k| \otimes Q_k \right) \\ &\qquad = \max_{k \in \{1,\ldots,n\}} \bigl(\mathrm{D}_{\max}(\rho_k \| Q_k) + \log(p_k)\bigr). \end{aligned} \tag{2.20}$$

In contrast, the ordinary quantum relative entropy obeys this equation:

$$\mathrm{D}\left(\sum_{k=1}^{n} p_k |k\rangle\langle k| \otimes \rho_k \,\middle\|\, \sum_{k=1}^{n} p_k |k\rangle\langle k| \otimes Q_k\right) = \sum_{k=1}^{n} p_k \,\mathrm{D}(\rho_k \| Q_k). \tag{2.21}$$

Using the equation $\mathrm{D}(\rho \| Q) = \mathrm{D}(\rho \| \lambda Q) - \log(\eta)$, we then conclude that

$$\mathrm{D}\left(\sum_{k=1}^{n} p_k |k\rangle\langle k| \otimes \rho_k \,\middle\|\, \sum_{k=1}^{n} |k\rangle\langle k| \otimes Q_k\right) = \sum_{k=1}^{n} p_k \,\mathrm{D}(\rho_k \| Q_k) - \mathrm{H}(p). \tag{2.22}$$

## 2.2 Conditional min-entropy

As was already mentioned at the beginning of the lecture, the ordinary conditional quantum entropy is given by the formula

$$\mathrm{H}(\mathsf{X}|\mathsf{Y})_\rho = -\,\mathrm{D}(\rho \| \mathbb{1}_{\mathcal{X}} \otimes \rho[\mathsf{Y}]). \tag{2.23}$$

We may also observe that

$$\mathrm{D}(\rho \| \mathbb{1}_{\mathcal{X}} \otimes \rho[\mathsf{Y}]) = \inf_{\sigma \in \mathrm{D}(\mathcal{Y})} \mathrm{D}(\rho \| \mathbb{1}_{\mathcal{X}} \otimes \sigma); \tag{2.24}$$

the infimum value is always obtained when $\sigma = \rho[\mathsf{Y}]$. With this fact in mind, we define the conditional min-entropy as follows.

**Definition 2.5.** Let $\mathsf{X}$ and $\mathsf{Y}$ be registers and let $\rho \in \mathrm{D}(\mathcal{X} \otimes \mathcal{Y})$ be a state of these registers. The *conditional min-entropy* of $\mathsf{X}$ given $\mathsf{Y}$ for the state $\rho$ is defined as

$$\mathrm{H}_{\mathrm{min}}(\mathsf{X}|\mathsf{Y})_\rho = -\inf_{\sigma \in \mathrm{D}(\mathcal{Y})} \mathrm{D}_{\mathrm{max}}(\rho \| \mathbb{1}_{\mathcal{X}} \otimes \sigma). \tag{2.25}$$

**Remark 2.6.** It is not, in general, the case that the infimum in (2.25) is achieved when $\sigma = \rho[\mathsf{Y}]$.

By expanding the definition of the max-relative entropy, one may alternatively express the conditional min-entropy in the following way:

$$\begin{aligned} 2^{-\mathrm{H}_{\mathrm{min}}(\mathsf{X}|\mathsf{Y})_\rho} &= \inf\{\eta \geq 0 \,:\, \rho \leq \eta \mathbb{1}_{\mathcal{X}} \otimes \sigma,\, \sigma \in \mathrm{D}(\mathcal{Y})\} \\ &= \inf\{\mathrm{Tr}(Y) \,:\, \rho \leq \mathbb{1}_{\mathcal{X}} \otimes Y,\, Y \in \mathrm{Pos}(\mathcal{Y})\}. \end{aligned} \tag{2.26}$$

The conditional min-entropy is always at most the ordinary conditional quantum entropy: $\mathrm{H}_{\mathrm{min}}(\mathsf{X}|\mathsf{Y})_\rho \leq \mathrm{H}(\mathsf{X}|\mathsf{Y})_\rho$. This fact follows from the fact that the max-relative entropy is at least the ordinary quantum relative entropy, for then we have

$$\mathrm{D}_{\mathrm{max}}(\rho \| \mathbb{1}_{\mathcal{X}} \otimes \sigma) \geq \mathrm{D}(\rho \| \mathbb{1}_{\mathcal{X}} \otimes \sigma) \tag{2.27}$$

for all density operators $\sigma$, implying the claimed inequality.

## Semidefinite program for conditional min-entropy

It is evident from (2.26) that the conditional min-entropy can be expressed as a semidefinite program. In particular, the quantity $\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho$ is the negative logarithm of the optimal value of the following semidefinite program.

**Optimization Problem 2.7** (SDP for conditional min-entropy)

| *Primal problem* | | *Dual problem* | |
|---|---|---|---|
| maximize: | $\langle \rho, X\rangle$ | minimize: | $\operatorname{Tr}(Y)$ |
| subject to: | $\operatorname{Tr}_{\mathcal{X}}(X) = \mathbb{1}_{\mathcal{Y}}$ | subject to: | $\mathbb{1}_{\mathcal{X}} \otimes Y \geq \rho$ |
| | $X \in \operatorname{Pos}(\mathcal{X}\otimes\mathcal{Y})$ | | $Y \in \operatorname{Herm}(\mathcal{Y})$ |

The dual problem is clearly consistent with the expression (2.26), whereas the primal problem corresponds (essentially) to an optimization of a linear function (represented by the state $\rho$) over all channels $\Phi \in \mathrm{C}(\mathcal{Y},\mathcal{X})$. There is a useful and intuitive way to think about this optimization, but first we will take a moment to introduce a useful concept, the *transpose* of a channel.

**Definition 2.8.** Let $\mathcal{X}$ and $\mathcal{Y}$ be complex Euclidean spaces and let $\Phi \in \mathrm{T}(\mathcal{Y},\mathcal{X})$. The *transpose* of $\Phi$ is the unique map $\Phi^{\mathsf{T}} \in \mathrm{T}(\mathcal{X},\mathcal{Y})$ satisfying the equation

$$\big(\Phi^{\mathsf{T}} \otimes \mathbb{1}_{\mathrm{L}(\mathcal{X})}\big)\big(\operatorname{vec}(\mathbb{1}_{\mathcal{X}})\operatorname{vec}(\mathbb{1}_{\mathcal{X}})^*\big) = \big(\mathbb{1}_{\mathrm{L}(\mathcal{Y})} \otimes \Phi\big)\big(\operatorname{vec}(\mathbb{1}_{\mathcal{Y}})\operatorname{vec}(\mathbb{1}_{\mathcal{Y}})^*\big). \tag{2.28}$$

Equivalently, the map $\Phi^{\mathsf{T}} \in \mathrm{T}(\mathcal{X},\mathcal{Y})$ is the (uniquely determined) map whose Choi representation is given by

$$J(\Phi^{\mathsf{T}}) = \big(\mathbb{1}_{\mathrm{L}(\mathcal{Y})} \otimes \Phi\big)\big(\operatorname{vec}(\mathbb{1}_{\mathcal{Y}})\operatorname{vec}(\mathbb{1}_{\mathcal{Y}})^*\big). \tag{2.29}$$

Here is a short list of facts concerning this notion, all of which are straightforward to prove.

1. $(\Phi^{\mathsf{T}})^{\mathsf{T}} = \Phi$.
2. The map $\Phi \mapsto \Phi^{\mathsf{T}}$ from $\mathrm{T}(\mathcal{Y},\mathcal{X})$ to $\mathrm{T}(\mathcal{X},\mathcal{Y})$, is linear, one-to-one, and onto.
3. $\Phi^{\mathsf{T}} \in \mathrm{CP}(\mathcal{X},\mathcal{Y})$ if and only if $\Phi \in \mathrm{CP}(\mathcal{Y},\mathcal{X})$.
4. $\Phi^{\mathsf{T}}$ is unital if and only if $\Phi$ preserves trace.

Finally, one may observe that $\Phi^{\mathsf{T}}$ is (as you might have guessed) the map that is obtained by taking any Kraus representation of $\Phi$ and transposing the Kraus operators.

Returning to Optimization Problem 2.7, let us consider the set $\mathcal{A}$ of primal feasible operators, which can be expressed in multiple ways based on the facts about the transpose of a map just listed:

$$\begin{aligned}\mathcal{A} &= \{X \in \mathrm{Pos}(\mathcal{X}\otimes\mathcal{Y}) \,:\, \mathrm{Tr}_{\mathcal{X}}(X) = \mathbb{1}_{\mathcal{Y}}\} \\ &= \bigl\{ \bigl(\Phi \otimes \mathbb{1}_{\mathrm{L}(\mathcal{Y})}\bigr)\bigl(\operatorname{vec}(\mathbb{1}_{\mathcal{Y}})\operatorname{vec}(\mathbb{1}_{\mathcal{Y}})^*\bigr) \,:\, \Phi \in \mathrm{C}(\mathcal{Y},\mathcal{X})\bigr\} \\ &= \bigl\{ \bigl(\mathbb{1}_{\mathrm{L}(\mathcal{X})} \otimes \Phi^{\mathsf{T}}\bigr)\bigl(\operatorname{vec}(\mathbb{1}_{\mathcal{X}})\operatorname{vec}(\mathbb{1}_{\mathcal{X}})^*\bigr) \,:\, \Phi \in \mathrm{C}(\mathcal{Y},\mathcal{X})\bigr\} \\ &= \bigl\{ \bigl(\mathbb{1}_{\mathrm{L}(\mathcal{X})} \otimes \Psi\bigr)\bigl(\operatorname{vec}(\mathbb{1}_{\mathcal{X}})\operatorname{vec}(\mathbb{1}_{\mathcal{X}})^*\bigr) \,:\, \Psi \in \mathrm{CP}(\mathcal{X},\mathcal{Y}),\ \Psi(\mathbb{1}_{\mathcal{X}}) = \mathbb{1}_{\mathcal{Y}}\bigr\}.\end{aligned} \tag{2.30}$$

The optimal value of the semidefinite program is $2^{-\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho}$, so

$$\begin{aligned}&2^{-\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho} \\ &= \sup\bigl\{\bigl\langle \rho, (\mathbb{1}_{\mathrm{L}(\mathcal{X})}\otimes\Psi)\bigl(\operatorname{vec}(\mathbb{1}_{\mathcal{X}})\operatorname{vec}(\mathbb{1}_{\mathcal{X}})^*\bigr)\bigr\rangle : \Psi \in \mathrm{CP}(\mathcal{X},\mathcal{Y}),\ \Psi(\mathbb{1}_{\mathcal{X}}) = \mathbb{1}_{\mathcal{Y}}\bigr\} \\ &= \sup\bigl\{\bigl\langle (\mathbb{1}_{\mathrm{L}(\mathcal{X})}\otimes\Psi^*)(\rho), \operatorname{vec}(\mathbb{1}_{\mathcal{X}})\operatorname{vec}(\mathbb{1}_{\mathcal{X}})^*\bigr\rangle : \Psi \in \mathrm{CP}(\mathcal{X},\mathcal{Y}),\ \Psi(\mathbb{1}_{\mathcal{X}}) = \mathbb{1}_{\mathcal{Y}}\bigr\} \\ &= \sup\bigl\{\bigl\langle (\mathbb{1}_{\mathrm{L}(\mathcal{X})}\otimes\Xi)(\rho), \operatorname{vec}(\mathbb{1}_{\mathcal{X}})\operatorname{vec}(\mathbb{1}_{\mathcal{X}})^*\bigr\rangle : \Xi \in \mathrm{C}(\mathcal{Y},\mathcal{X})\bigr\}.\end{aligned} \tag{2.31}$$

That is,

$$2^{-\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho} = n \cdot \sup_{\Xi\in\mathrm{C}(\mathcal{Y},\mathcal{X})} \bigl\langle \bigl(\mathbb{1}_{\mathrm{L}(\mathcal{X})}\otimes\Xi\bigr)(\rho), \tau\bigr\rangle \tag{2.32}$$

where

$$\tau = \frac{1}{n}\sum_{a,b=1}^{n} |a\rangle\langle b| \otimes |a\rangle\langle b| \quad \text{and} \quad n = \dim(\mathcal{X}). \tag{2.33}$$

In words, $2^{-\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho}$ is equal to $\dim(\mathcal{X})$ times the maximum squared-fidelity, over all channels $\Xi \in \mathrm{C}(\mathcal{Y},\mathcal{X})$, between the state $(\mathbb{1}_{\mathrm{L}(\mathcal{X})}\otimes\Xi)(\rho)$ and the canonical maximally entangled state $\tau \in \mathrm{D}(\mathcal{X}\otimes\mathcal{X})$.

## 2.3 Examples

We will conclude the lecture by considering the conditional min-entropy of a few classes of states.

**Example 2.9.** Suppose $\mathsf{Y}$ is trivial (i.e., one-dimensional), so that $\rho \in \mathrm{D}(\mathcal{X})$. We then find that

$$\begin{aligned}\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho &= -\inf_{\sigma\in\mathrm{D}(\mathcal{Y})} \mathrm{D}_{\max}(\rho\|\mathbb{1}_{\mathcal{X}}\otimes\sigma) \\ &= -\mathrm{D}_{\max}(\rho\|\mathbb{1}_{\mathcal{X}}) \\ &= -\log\lambda_1(\rho).\end{aligned} \tag{2.34}$$

Naturally, we omit the register $\mathsf{Y}$ from this notation when it is trivial:

$$\mathrm{H}_{\min}(\mathsf{X})_\rho = \mathrm{H}_{\min}(\rho) = -\log\lambda_1(\rho). \tag{2.35}$$

**Example 2.10.** Suppose $\rho = \sigma \otimes \xi$ for $\sigma \in \mathrm{D}(\mathcal{X})$ and $\xi \in \mathrm{D}(\mathcal{Y})$. Through a similar calculation to the previous example, we find that

$$\begin{aligned} \mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho &= -\inf_{\xi' \in \mathrm{D}(\mathcal{Y})} \mathrm{D}_{\max}(\sigma \otimes \xi \| \mathbb{1}_{\mathcal{X}} \otimes \xi') \\ &= -\mathrm{D}_{\max}(\sigma \| \mathbb{1}_{\mathcal{X}}) \\ &= \mathrm{H}_{\min}(\mathsf{X})_\sigma. \end{aligned} \tag{2.36}$$

This is natural: if the registers X and Y are completely uncorrelated, the conditional min-entropy of X given Y is simply the min-entropy of X.

**Example 2.11.** Next, suppose that we have a separable state: $\rho \in \mathrm{SepD}(\mathcal{X} : \mathcal{Y})$. Then, for any channel $\Xi \in \mathrm{C}(\mathcal{Y}, \mathcal{X})$ we have

$$(\mathbb{1}_{\mathrm{L}(\mathcal{X})} \otimes \Xi)(\rho) \in \mathrm{SepD}(\mathcal{X} : \mathcal{X}); \tag{2.37}$$

applying a channel locally to one part of a separable state always results in a separable state. The inner-product between any separable state and the canonical maximally entangled state $\tau$ is at most $1/n$ (as we proved in CS 766/QIC 820), and therefore

$$2^{-\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho} = n \cdot \sup_{\Xi \in \mathrm{C}(\mathcal{Y}, \mathcal{X})} \big\langle \big(\mathbb{1}_{\mathrm{L}(\mathcal{X})} \otimes \Xi\big)(\rho), \tau \big\rangle \leq n \cdot \frac{1}{n} = 1. \tag{2.38}$$

The conditional min-entropy of every separable state is therefore nonnegative.

By similar reasoning, for every PPT state $\rho \in \mathrm{PPT}(\mathcal{X} : \mathcal{Y}) \cap \mathrm{D}(\mathcal{X} \otimes \mathcal{Y})$ it is the case that $\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho \geq 0$.

**Example 2.12.** Suppose that the $\tau$ can be recovered perfectly by applying a channel locally to Y for the state $\rho \in \mathrm{D}(\mathcal{X} \otimes \mathcal{Y})$. This is equivalent to $\rho$ taking the form

$$\rho = (\mathbb{1}_{\mathcal{X}} \otimes V)(\tau \otimes \xi)(\mathbb{1}_{\mathcal{X}} \otimes V)^* \tag{2.39}$$

for some choice of a density operator $\xi \in \mathrm{D}(\mathcal{Z})$ and an isometry $V \in \mathrm{U}(\mathcal{X} \otimes \mathcal{Z}, \mathcal{Y})$. Then we have

$$\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho = -\log(n), \tag{2.40}$$

which is the minimum possible value for the conditional min-entropy.

**Example 2.13.** Finally, suppose that $\rho$ is a classical-quantum state:

$$\rho = \sum_{a=1}^{n} p(a) \, |a\rangle\langle a| \otimes \xi_a. \tag{2.41}$$

We find that

$$\begin{aligned} 2^{-\,\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho} &= n \cdot \sup_{\Xi \in \mathrm{C}(\mathcal{Y},\mathcal{X})} \big\langle \big(\mathbb{1}_{\mathrm{L}(\mathcal{X})} \otimes \Xi\big)(\rho), \tau \big\rangle \\ &= \sup_{\Xi \in \mathrm{C}(\mathcal{Y},\mathcal{X})} \sum_{a=1}^{n} p(a) \langle a|\, \Xi(\xi_a)\, |a\rangle, \end{aligned} \tag{2.42}$$

with the simplification to the second line being possible because $\rho$ is a classical-quantum state. This has the following intuitive meaning: $\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho$ is the negative logarithm of the optimal correctness probability to identify a state chosen randomly according to the ensemble corresponding to $\rho$.

Lecture 3

# Smoothing and optimizing max-relative entropy

Recall the definition of the max-relative entropy, which was introduced in the previous lecture:

$$\mathrm{D}_{\max}(\rho\|Q) = \inf\{\lambda \in \mathbb{R} : \rho \leq 2^{\lambda} Q\}. \tag{3.1}$$

In this lecture we will consider what happens when we minimize this function over various choices of $\rho$ and $Q$. Two common situations in which this is done are as follows:

1. *Smoothing.* For a given $\rho$, $Q$, and $\varepsilon > 0$, the *smoothed max-relative entropy* of $\rho$ with respect to $Q$ is defined as

$$\mathrm{D}^{\varepsilon}_{\max}(\rho\|Q) = \inf_{\xi \in \mathcal{B}_{\varepsilon}(\rho)} \mathrm{D}_{\max}(\xi\|Q), \tag{3.2}$$

   where $\mathcal{B}_{\varepsilon}(\rho)$ denotes the set of states that are $\varepsilon$-close to $\rho$ with respect to some notion of distance.

2. *Optimizing over models.* For a given $\rho$ and a convex set $\mathcal{C}$ of possible choices of models $Q$, we may consider the quantity

$$\mathrm{D}_{\max}(\rho\|\mathcal{C}) = \inf_{Q \in \mathcal{C}} \mathrm{D}_{\max}(\rho\|Q), \tag{3.3}$$

   which measures in a certain sense which $Q \in \mathcal{C}$ incurs the least loss of efficiency when serving as a model for the state $\rho$.

In both cases, the optimizations can be represented as conic programs. We will consider the two types of optimizations separately in the sections that follow.

## 3.1 Smoothed max-relative entropy

Let us begin with the smoothed max-relative entropy, where one takes the minimum value of the max-relative entropy over all choices of actual states that are close to a given state, as suggested above. The idea is that the smoothed max-relative entropy reflects a tolerance for small errors, which we often have or would like to express when analyzing operationally defined notions. Without smoothing, the max-relative entropy can sometimes, in certain settings at least, have unwanted hyper-sensitivities that smoothing eliminates.

### Definition

As it turns out, there is not a single agreed upon definition for the smoothed max-relative entropy; different authors sometimes choose different notions of distance with respect to which the smoothing is done, which translates to different choices for the set $\mathcal{B}_\varepsilon(\rho)$ in (3.2). In addition, the operator $\xi$ is sometimes allowed to range not only over density operators, but also over sub-normalized density operators, and in this case the definition of the max-relative entropy is extended in the most straightforward way to accommodate such operators.

It is typically the case, however, that the notions of distance with respect to which the smoothed max-relative entropy is defined are based on either the trace distance or the fidelity function. Through the Fuchs–van de Graaf inequalities, one finds that the resulting definitions of smoothed max-relative entropy are roughly equivalent, and are certainly quite similar in a qualitative sense.

For the sake of concreteness, we will define the smoothed max-relative entropy in terms of the trace distance, as the following definition makes precise.

**Definition 3.1** (Smoothed max-relative entropy)**.** For a density operator $\rho \in \mathrm{D}(\mathcal{X})$, a positive semidefinite operator $Q \in \mathrm{Pos}(\mathcal{X})$, and a real number $\varepsilon \in (0,1)$, the *$\varepsilon$-smoothed relative max-entropy* of $\rho$ with respect to $Q$ is defined as

$$\mathrm{D}^{\varepsilon}_{\mathrm{max}}(\rho \| Q) = \min_{\xi \in \mathcal{B}_\varepsilon(\rho)} \mathrm{D}_{\mathrm{max}}(\xi \| Q), \tag{3.4}$$

where

$$\mathcal{B}_\varepsilon(\rho) = \bigl\{ \xi \in \mathrm{D}(\mathcal{X}) \,:\, \tfrac{1}{2}\|\rho - \xi\|_1 \leq \varepsilon \bigr\}. \tag{3.5}$$

A couple of other common choices for $\mathcal{B}_\varepsilon(\rho)$ are these:

$$\begin{aligned} \mathcal{B}_\varepsilon(\rho) &= \bigl\{ \xi \in \mathrm{D}(\mathcal{X}) \,:\, \mathrm{F}(\xi,\rho)^2 \geq 1 - \varepsilon \bigr\}, \\ \mathcal{B}_\varepsilon(\rho) &= \bigl\{ \xi \in \mathrm{Pos}(\mathcal{X}) \,:\, \mathrm{F}(\xi,\rho)^2 \geq 1 - \varepsilon^2,\, \mathrm{Tr}(\xi) \leq 1 \bigr\}. \end{aligned} \tag{3.6}$$

## Optimizing over arbitrary closed and convex sets of states

To better understand the smoothed max-relative entropy, it is helpful to consider a more general set-up. Suppose that $\mathcal{C} \subseteq \mathrm{D}(\mathcal{X})$ is any closed and compact set of density operators, let $Q \in \mathrm{Pos}(\mathcal{X})$ be given, and consider the problem of minimizing $\eta$ over all choices of $\xi \in \mathcal{C}$ and $\eta \in \mathbb{R}$ that satisfy $\xi \leq \eta Q$. This problem can be expressed as a conic problem as will now be described.

First, define a set $\mathcal{K} \subset \mathbb{R} \oplus \mathrm{Herm}(\mathcal{X}) \oplus \mathbb{R} \oplus \mathrm{Herm}(\mathcal{X})$ as follows:

$$\mathcal{K} = \bigl\{(\lambda, \lambda\xi, \eta, P) \,:\, \eta, \lambda \geq 0,\, \xi \in \mathcal{C},\, P \in \mathrm{Pos}(\mathcal{X})\bigr\}. \tag{3.7}$$

This set is a closed and convex cone. One may think of $\mathcal{K}$ as being the Cartesian product of three sets: the first is

$$\mathcal{L} = \bigl\{(\lambda, \lambda\xi) \,:\, \lambda \geq 0,\, \xi \in \mathcal{C}\bigr\}, \tag{3.8}$$

the second is the set of all nonnegative real numbers $\eta \geq 0$, and the third is $\mathrm{Pos}(\mathcal{X})$. All three of these sets are closed and convex cones; in the case of $\mathcal{L}$ this follows from the assumption that $\mathcal{C}$ is a compact and convex set. It should be noted that the construction of a convex cone $\mathcal{L}$ from a convex set $\mathcal{C}$ like this is both common and useful.

Now, the optimization problem suggested above is evidently equivalent to the problem of minimizing the inner-product

$$\bigl\langle (0,0,1,0), (\lambda, \lambda\xi, \eta, P) \bigr\rangle \tag{3.9}$$

over all $(\lambda, \lambda\xi, \eta, P) \in \mathcal{K}$, subject to the affine linear constraints that

$$\lambda = 1 \quad \text{and} \quad \eta Q = \lambda\xi + P. \tag{3.10}$$

By defining a linear map $\phi : \mathbb{R} \oplus \mathrm{Herm}(\mathcal{X}) \oplus \mathbb{R} \oplus \mathrm{Herm}(\mathcal{X}) \rightarrow \mathbb{R} \oplus \mathrm{Herm}(\mathcal{X})$ as

$$\phi(\lambda, X, \eta, Y) = (\lambda, \eta Q - X - Y), \tag{3.11}$$

these affine linear constraints may be expressed as

$$\phi(\lambda, \lambda\xi, \eta, P) = (1, 0). \tag{3.12}$$

The optimization problem being considered is therefore the primal form of a conic program, which is stated below (together with a simplified expression of its dual form) as Optimization Problem 3.2.

The dual form of this optimization problem is given by the maximization of the objective function

$$\bigl\langle (1,0), (\mu, Z) \bigr\rangle \tag{3.13}$$

subject to the constraint

$$(0,0,1,0) - \phi^*(\mu, Z) \in \mathcal{K}^*, \tag{3.14}$$

where $(\mu, Z)$ ranges over the space $\mathbb{R} \oplus \text{Herm}(\mathcal{X})$ (as is dictated by Optimization Problem 1.4 in Lecture 1). To simplify this problem, we must compute the adjoint map $\phi^*$ and try to understand what $\mathcal{K}^*$ looks like. The adjoint of $\phi$ may simply be computed, and one obtains

$$\phi^*(\mu, Z) = (\mu, -Z, \langle Q, Z \rangle, -Z). \tag{3.15}$$

As for the dual cone to $\mathcal{K}$, an element $(\delta, H, \gamma, R)$ is contained in $\mathcal{K}^*$ if and only if

$$\big\langle (\delta, H, \gamma, R), (\lambda, \lambda\xi, \eta, P) \big\rangle = \delta\lambda + \lambda\langle H, \xi \rangle + \gamma\eta + \langle R, P \rangle \geq 0 \tag{3.16}$$

for all $\lambda, \eta \geq 0$, $\xi \in \mathcal{C}$, and $P \in \text{Pos}(\mathcal{X})$. This is equivalent to the requirement that $\gamma \geq 0$, $R \in \text{Pos}(\mathcal{X})$, and

$$\delta \geq -\langle H, \xi \rangle \tag{3.17}$$

for all $\xi \in \mathcal{C}$. By defining a function

$$\psi_{\mathcal{C}}(H) = \inf_{\xi \in \mathcal{C}} \langle \xi, H \rangle, \tag{3.18}$$

we may alternatively express $\mathcal{K}^*$ as follows:

$$\mathcal{K}^* = \big\{ (\delta, H, \gamma, R) \,:\, \delta \geq -\psi_{\mathcal{C}}(H),\ \gamma \geq 0,\ R \in \text{Pos}(\mathcal{X}) \big\}. \tag{3.19}$$

The dual problem is therefore a maximization of $\delta$ subject to the constraint that $\delta \leq \psi_{\mathcal{C}}(Z)$, $\langle Q, Z \rangle \leq 1$, and $Z \in \text{Pos}(\mathcal{X})$, or, equivalently, a maximization of $\psi_{\mathcal{C}}(Z)$ over all $Z \in \text{Pos}(\mathcal{Z})$ satisfying $\langle Q, Z \rangle \leq 1$. We obtain the following simplified expression of this conic program.

**Optimization Problem 3.2**

$$\begin{array}{ll@{\qquad\qquad}rl}
\multicolumn{2}{c}{\textit{Primal problem}} & \multicolumn{2}{c}{\textit{Dual problem}} \\
\text{minimize:} & \eta & \text{maximize:} & \inf\{\langle \xi, Z \rangle \,:\, \xi \in \mathcal{C}\} \\
\text{subject to:} & \xi \leq \eta Q, & \text{subject to:} & \langle Q, Z \rangle \leq 1, \\
 & \xi \in \mathcal{C}, & & Z \in \text{Pos}(\mathcal{X}). \\
 & \eta \geq 0. & &
\end{array}$$

It can be shown, through the use of Slater's theorem, that strong duality always holds for Optimization Problem 3.2.

### Conic program for smoothed max-relative entropy

At this point, one may substitute the set $\mathcal{B}_\varepsilon(\rho)$ for $\mathcal{C}$ in Optimization Problem 3.2 to obtain a conic program for the smoothed max-relative entropy. To be precise, the smoothed max-relative entropy $\mathrm{D}^{\varepsilon}_{\max}(\rho\|Q)$ is the logarithm of the optimal value of this conic program. One may also note that if $\mathcal{C} = \{\rho\}$, then the semidefinite program for the ordinary (non-smoothed) max-relative entropy is recovered.

Naturally, other notions of smoothing can be considered by making alternative choices for the set $\mathcal{C}$.

## 3.2 Minimizing over a convex set of models

Next we may consider what happens when we minimize the max-relative entropy over a compact and convex set in the second coordinate. That is, for $\mathcal{C} \subseteq \mathrm{Pos}(\mathcal{X})$ being a compact and convex set, we may consider the quantity

$$\mathrm{D}_{\max}(\rho\|\mathcal{C}) = \inf_{Q\in\mathcal{C}} \mathrm{D}_{\max}(\rho\|Q) \tag{3.20}$$

for a given choice of $\rho \in \mathrm{D}(\mathcal{X})$.

### Conic program formulation

Going through a similar process to the one above, we obtain the following conic program.

**Optimization Problem 3.3**

| *Primal problem* | | *Dual problem* | |
|---|---|---|---|
| minimize: | $\eta$ | maximize: | $\langle \rho, Z\rangle$ |
| subject to: | $\rho \leq \eta Q$ | subject to: | $\sup\{\langle Q, Z\rangle \,:\, Q\in\mathcal{C}\} \leq 1$ |
| | $Q \in \mathcal{C}$, | | $Z \in \mathrm{Pos}(\mathcal{X})$ |
| | $\eta \geq 0$ | | |

Similar to before, the value $\mathrm{D}_{\max}(\rho\|\mathcal{C})$ defined above is the logarithm of the optimal value of this conic program. The constraint in the dual problem can alternatively be written

$$\theta_{\mathcal{C}}(Z) \leq 1, \tag{3.21}$$

where

$$\theta_{\mathcal{C}}(H) = \sup_{Q\in\mathcal{C}} \langle Q, H\rangle \tag{3.22}$$

is the so-called *support function* of the convex set $\mathcal{C}$.

**Example 3.4.** We have, in fact, already seen an example in which we minimize over models drawn from a convex set: the *conditional min-entropy*. The conditional min-entropy of X given Y, for the state $\rho \in \mathrm{D}(\mathcal{X} \otimes \mathcal{Y})$, is given by

$$\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho = -\inf_{\sigma \in \mathrm{D}(\mathcal{Y})} \mathrm{D}_{\max}(\rho \| \mathbb{1}_{\mathcal{X}} \otimes \sigma) = -\inf_{Q \in \mathcal{C}} \mathrm{D}_{\max}(\rho \| Q) \tag{3.23}$$

for $\mathcal{C} = \{\mathbb{1}_{\mathcal{X}} \otimes \sigma : \sigma \in \mathrm{D}(\mathcal{Y})\}$. Observe that

$$\begin{aligned} \theta_{\mathcal{C}}(Z) &= \sup_{Q \in \mathcal{C}} \langle Q, Z \rangle = \sup_{\sigma \in \mathrm{D}(\mathcal{Y})} \langle \mathbb{1}_{\mathcal{X}} \otimes \sigma, Z \rangle \\ &= \sup_{\sigma \in \mathrm{D}(\mathcal{Y})} \langle \sigma, \mathrm{Tr}_{\mathcal{X}}(Z) \rangle = \lambda_1\bigl(\mathrm{Tr}_{\mathcal{X}}(Z)\bigr). \end{aligned} \tag{3.24}$$

By the dual formulation of the conic program above, we find that

$$\begin{aligned} \mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho &= \sup\bigl\{\log\langle \rho, Z \rangle : \lambda_1\bigl(\mathrm{Tr}_{\mathcal{X}}(Z)\bigr) \leq 1,\, Z \in \mathrm{Pos}(\mathcal{X} \otimes \mathcal{Y})\bigr\} \\ &= \sup\bigl\{\log\langle \rho, Z \rangle : \mathrm{Tr}_{\mathcal{X}}(Z) \leq \mathbb{1}_{\mathcal{Y}},\, Z \in \mathrm{Pos}(\mathcal{X} \otimes \mathcal{Y})\bigr\} \\ &= \sup\bigl\{\log\langle \rho, Z \rangle : \mathrm{Tr}_{\mathcal{X}}(Z) = \mathbb{1}_{\mathcal{Y}},\, Z \in \mathrm{Pos}(\mathcal{X} \otimes \mathcal{Y})\bigr\}, \end{aligned} \tag{3.25}$$

which is consistent with the primal problem in our semidefinite program for the conditional min-entropy from Lecture 2.

## Divergence from a convex set of states

For a convex set of states $\mathcal{C} \subseteq \mathrm{D}(\mathcal{X})$, it is typical that one views the quantity

$$\mathrm{D}(\rho \| \mathcal{C}) = \inf_{\sigma \in \mathcal{C}} \mathrm{D}(\rho \| \sigma) \tag{3.26}$$

(where it should be stressed that it is the ordinary quantum relative entropy, not the max-relative entropy, that appears in this equation) as a measure of distance (or divergence) of $\rho$ from $\mathcal{C}$. For example, the *relative entropy of entanglement* of a state $\rho \in \mathrm{D}(\mathcal{Y} \otimes \mathcal{Z})$ is given by

$$\mathrm{REE}(\mathsf{Y} : \mathsf{Z})_\rho = \mathrm{D}(\rho \| \mathrm{SepD}(\mathcal{Y} : \mathcal{Z})) = \inf_{\sigma \in \mathrm{SepD}(\mathcal{Y}:\mathcal{Z})} \mathrm{D}(\rho \| \sigma). \tag{3.27}$$

We may consider a similar notion for the *max-relative entropy* in place of the ordinary relative entropy:

$$\mathrm{D}_{\max}(\rho \| \mathcal{C}) = \inf_{\sigma \in \mathcal{C}} \mathrm{D}_{\max}(\rho \| \sigma). \tag{3.28}$$

To better understand this quantity, let us expand the definition of the max-relative entropy, so that we obtain

$$\mathrm{D}_{\max}(\rho \| \mathcal{C}) = \inf\bigl\{\lambda \in \mathbb{R} : \rho \leq 2^{\lambda}\sigma,\, \sigma \in \mathcal{C}\bigr\}. \tag{3.29}$$

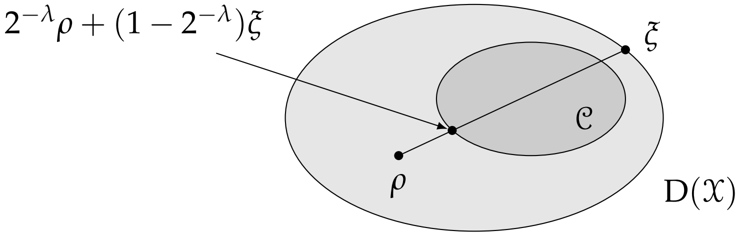
$2^{-\lambda}\rho + (1-2^{-\lambda})\xi$
$\xi$
$\mathcal{C}$
$\rho$
$\mathrm{D}(\mathcal{X})$

Figure 3.1: $\mathrm{D}_{\max}(\rho\|\mathcal{C})$ is the minimum value of $\lambda$ for which $2^{-\lambda}\rho + (1-2^{-\lambda})\xi$ is in $\mathcal{C}$, for some choice of a density operator $\xi$.

The inequality on the right-hand side of this equation can be expressed as an equality through the introduction of a positive semidefinite slack variable $P$, which yields

$$\mathrm{D}_{\max}(\rho\|\mathcal{C}) = \inf\bigl\{\lambda \in \mathbb{R} \,:\, \rho + P = 2^{\lambda}\sigma,\ \sigma \in \mathcal{C},\ P \in \mathrm{Pos}(\mathcal{X})\bigr\}. \tag{3.30}$$

As $\rho$ and $\sigma$ are density operators, the slack variable $P$ must have trace equal to $2^{\lambda}-1$ in order for the equality $\rho + P = 2^{\lambda}\sigma$ to hold. We may therefore replace $P$ by $(2^{\lambda}-1)\xi$ for a density operator $\xi$, and we obtain

$$\mathrm{D}_{\max}(\rho\|\mathcal{C}) = \inf\bigl\{\lambda \in \mathbb{R} \,:\, \rho + \bigl(2^{\lambda}-1\bigr)\xi = 2^{\lambda}\sigma,\ \sigma \in \mathcal{C},\ \xi \in \mathrm{D}(\mathcal{X})\bigr\}, \tag{3.31}$$

which is equivalent to

$$\mathrm{D}_{\max}(\rho\|\mathcal{C}) = \inf\bigl\{\lambda \in \mathbb{R} : 2^{-\lambda}\rho + \bigl(1-2^{-\lambda}\bigr)\xi \in \mathcal{C},\ \xi \in \mathrm{D}(\mathcal{X})\bigr\}. \tag{3.32}$$

This expression reveals that the quantity $\mathrm{D}_{\max}(\rho\|\mathcal{C})$ has the simple and intuitive interpretation suggested by Figure 3.1. For $\lambda = \mathrm{D}_{\max}(\rho\|\mathcal{C})$, the quantity $2^{\lambda}-1$ is sometimes called the (global or generalized) *robustness* of $\rho$ with respect to $\mathcal{C}$.

## 3.3 Minimizing over both states and models

Finally, and very briefly, it should be noted that one can simultaneously minimize over both arguments in the max-relative entropy. That is, if $\mathcal{C} \subseteq \mathrm{D}(\mathcal{X})$ and $\mathcal{D} \subseteq \mathrm{Pos}(\mathcal{X})$ are convex and compact sets, one may consider the quantity

$$\mathrm{D}_{\max}(\mathcal{C}\|\mathcal{D}) = \inf_{\substack{\sigma\in\mathcal{C}\\ Q\in\mathcal{D}}} \mathrm{D}_{\max}(\rho\|Q). \tag{3.33}$$

This value is the logarithm of the optimal value of the following conic program.

**Optimization Problem 3.5**

$$
\begin{array}{rl@{\qquad\qquad}rl}
\multicolumn{2}{c}{\textit{Primal problem}} & \multicolumn{2}{c}{\textit{Dual problem}} \\
\text{minimize:} & \eta & \text{maximize:} & \psi_{\mathcal{C}}(Z) \\
\text{subject to:} & \rho \leq \eta Q & \text{subject to:} & \theta_{\mathcal{D}}(Z) \leq 1 \\
& \rho \in \mathcal{C} & & Z \in \operatorname{Pos}(\mathcal{X}) \\
& Q \in \mathcal{D} & & \\
& \eta \geq 0 & &
\end{array}
$$

where

$$\psi_{\mathcal{C}}(Z) = \inf_{\rho \in \mathcal{C}} \langle \rho, Z \rangle \qquad \text{and} \qquad \theta_{\mathcal{D}}(Z) = \sup_{Q \in \mathcal{D}} \langle Q, Z \rangle. \tag{3.34}$$

Lecture 4

# Regularization of the smoothed max-relative entropy

In this lecture we will prove an important theorem concerning the smoothed max-relative entropy, which is that by regularizing the smoothed max-relative entropy we obtain the ordinary quantum relative entropy:

$$\lim_{n\to\infty} \frac{\mathrm{D}^{\varepsilon}_{\max}\bigl(\rho^{\otimes n}\big\|\sigma^{\otimes n}\bigr)}{n} = \mathrm{D}(\rho\|\sigma) \tag{4.1}$$

for all density operators $\rho, \sigma \in \mathrm{D}(\mathcal{X})$ and all $\varepsilon \in (0,1)$. For the sake of clarity, recall that we define the smoothed max-relative entropy with respect to *trace-distance smoothing*:

$$\mathrm{D}^{\varepsilon}_{\max}(\rho\|\sigma) = \inf_{\xi\in\mathcal{B}_{\varepsilon}(\rho)} \mathrm{D}_{\max}(\xi\|\sigma) \tag{4.2}$$

where

$$\mathcal{B}_{\varepsilon}(\rho) = \bigl\{\xi \in \mathrm{D}(\mathcal{X}) \,:\, \tfrac{1}{2}\|\rho - \xi\|_1 \leq \varepsilon\bigr\}. \tag{4.3}$$

## Bibliographic remarks

Lemma 4.4, which in some sense is the engine that drives the proof we will discuss, is due to Bjelaković and Siegmund-Schultze (`arXiv:quant-ph/0307170`), who used it to prove the so-called *quantum Stein lemma*, and through it obtained an alternative proof of the monotonicity of quantum relative entropy.

The more direct route from Bjelaković and Siegmund-Schultze's lemma to the regularization (4.1) to be followed in this lecture appears in the following as-of-yet unpublished manuscript:

> Shitikanth Kashyap, Ashwin Nayak, and Michael Saks. Asymptotic equipartition for quantum relative entropy revisited. Manuscript, 2014.

## 4.1 Strong typicality

The general notion of *typicality* is fundamentally important in information theory, and there is a sense in which it goes hand-in-hand with the concept of entropy. We will begin the lecture with a brief and directed summary of *strong typicality*, which is a particular formulation of typicality that is convenient for the proof.

First let us introduce some notation. Supposing that $\Sigma$ is an alphabet, for every string $a_1 \cdots a_n \in \Sigma^n$ and symbol $a \in \Sigma$, we write

$$N(a \,|\, a_1 \cdots a_n) = \bigl|\{k \in \{1,\ldots,n\} \,:\, a_k = a\}\bigr|, \tag{4.4}$$

which is simply the number of times the symbol $a$ occurs in the string $a_1 \cdots a_n$. With respect to that notation, strong typicality is defined as follows.

**Definition 4.1.** Let $\Sigma$ be an alphabet, let $p \in \mathcal{P}(\Sigma)$ be a probability vector, let $n$ be a positive integer, and let $\delta > 0$ be a positive real number. A string $a_1 \cdots a_n \in \Sigma^n$ is *$\delta$-strongly typical with respect to $p$* if

$$\left| \frac{N(a \,|\, a_1 \cdots a_n)}{n} - p(a) \right| \leq p(a)\delta \tag{4.5}$$

for every $a \in \Sigma$. The set of all $\delta$-strongly typical strings of length $n$ with respect to $p$ is denoted $S_{n,\delta}(p)$.

What the definition expresses is that the proportion of each symbol in a strongly typical string is approximately what one would expect if the individual symbols were chosen independently at random according to the probability vector $p$. Notice that because it is the quantity $p(a)\delta$, as opposed to $\delta$, that appears on the right-hand side of the inequality in the definition, we have that the error tolerance for the frequency with which each symbol appears shrinks proportionately as the probability for that symbol to appear shrinks—and if $p(a) = 0$ for some $a \in \Sigma$, then a strongly typical string cannot include the symbol $a$ at all.

Next we will prove two basic facts concerning the notion of strong typicality. The two facts are stated as the lemmas that follow.

**Lemma 4.2.** *Let $\Sigma$ be an alphabet, let $p \in \mathcal{P}(\Sigma)$ be a probability vector, let $n$ be a positive integer, and let $\delta > 0$ be a positive real number. It is the case that*

$$\sum_{a_1 \cdots a_n \in S_{n,\delta}(p)} p(a_1) \cdots p(a_n) \geq 1 - 2 \sum_{\substack{a \in \Sigma \\ p(a) > 0}} \exp\bigl(-2n\delta^2 p(a)^2\bigr). \tag{4.6}$$

*Proof.* Suppose first that $a \in \Sigma$ is fixed, and consider the probability that a string $a_1 \cdots a_n \in \Sigma^n$, where each symbol is selected independently at random according to the probability vector $p$, satisfies

$$\left| \frac{N(a \,|\, a_1 \cdots a_n)}{n} - p(a) \right| > p(a)\delta. \tag{4.7}$$

To upper-bound this probability, one may define $X_1, \ldots, X_n$ to be independent and identically distributed random variables, taking value 1 with probability $p(a)$ and value 0 otherwise, so that the probability of the event (4.7) is equal to

$$\Pr\left( \left| \frac{X_1 + \cdots + X_n}{n} - p(a) \right| > p(a)\delta \right). \tag{4.8}$$

If it is the case that $p(a) > 0$, then Hoeffding's inequality implies that

$$\Pr\left( \left| \frac{X_1 + \cdots + X_n}{n} - p(a) \right| > p(a)\delta \right) \leq 2\exp\bigl(-2n\delta^2 p(a)^2\bigr), \tag{4.9}$$

while it is the case that

$$\Pr\left( \left| \frac{X_1 + \cdots + X_n}{n} - p(a) \right| > p(a)\delta \right) = 0 \tag{4.10}$$

in case $p(a) = 0$. The lemma follows from the union bound. □

**Lemma 4.3.** *Let $\Sigma$ be an alphabet, let $p \in \mathcal{P}(\Sigma)$ be a probability vector, let $n$ be a positive integer, let $\delta > 0$ be a positive real number, let $a_1 \cdots a_n \in S_{n,\delta}(p)$ be a $\delta$-strongly typical string with respect to $p$, and let $\phi : \Sigma \rightarrow [0, \infty)$ be a nonnegative real-valued function. The following inequality is satisfied:*

$$\left| \frac{\phi(a_1) + \cdots + \phi(a_n)}{n} - \sum_{a \in \Sigma} p(a)\phi(a) \right| \leq \delta \sum_{a \in \Sigma} p(a)\phi(a). \tag{4.11}$$

*Proof.* The inequality (4.11) follows from the definition of strong typicality together with the triangle inequality:

$$\begin{aligned}
&\left| \frac{\phi(a_1) + \cdots + \phi(a_n)}{n} - \sum_{a \in \Sigma} p(a)\phi(a) \right| \\
&\qquad = \left| \sum_{a \in \Sigma} \left( \frac{N(a \,|\, a_1 \cdots a_n)}{n} - p(a) \right) \phi(a) \right| \qquad (4.12)\\
&\qquad \leq \sum_{a \in \Sigma} \left| \frac{N(a \,|\, a_1 \cdots a_n)}{n} - p(a) \right| \phi(a) \leq \delta \sum_{a \in \Sigma} p(a)\phi(a),
\end{aligned}$$

as required. □

## 4.2 Lemmas

Next we will prove two lemmas that are needed for the proof of the main theorem to which this lecture is devoted. The first of these lemmas is the one due to Bjelaković and Siegmund-Schultze mentioned at the start of the lecture.

**Lemma 4.4.** *Let $\rho, \sigma \in \mathrm{D}(\mathcal{X})$ be density operators for which $\mathrm{im}(\rho) \subseteq \mathrm{im}(\sigma)$ and let $\delta > 0$ be a positive real number. There exist positive real numbers $K$ and $\mu$ such that, for every positive integer $n$, there exists a projection operator $\Pi_n$ acting on $\mathcal{X}^{\otimes n}$ satisfying $\left[\Pi_n, \sigma^{\otimes n}\right] = 0$,*

$$\left\langle \Pi_n, \rho^{\otimes n} \right\rangle \geq 1 - K \exp(-\mu n), \tag{4.13}$$

*and*

$$2^{(1+\delta) n \operatorname{Tr}(\rho \log(\sigma))} \Pi_n \leq \Pi_n \sigma^{\otimes n} \Pi_n \leq 2^{(1-\delta) n \operatorname{Tr}(\rho \log(\sigma))} \Pi_n. \tag{4.14}$$

*Proof.* By considering a spectral decomposition of $\sigma$, one may select an alphabet $\Sigma$, an orthonormal set $\{x_a : a \in \Sigma\} \subset \mathcal{X}$, and a probability vector $q \in \mathcal{P}(\Sigma)$ such that

$$\sigma = \sum_{a \in \Sigma} q(a) x_a x_a^* \tag{4.15}$$

and $q(a) > 0$ for all $a \in \Sigma$. Define a new probability vector $p \in \mathcal{P}(\Sigma)$ as

$$p(a) = x_a^* \rho x_a \tag{4.16}$$

for every $a \in \Sigma$. The fact that $p$ is indeed a probability vector follows from the assumption that $\rho$ is a density operator with $\mathrm{im}(\rho) \subseteq \mathrm{im}(\sigma)$.

Real numbers $K$ and $\mu$ satisfying the requirements of the lemma may now be selected as follows:

$$\begin{aligned} K &= 2|\mathrm{supp}(p)|, \\ \mu &= 2\delta^2 \min\{p(a)^2 : a \in \Sigma,\ p(a) > 0\}. \end{aligned} \tag{4.17}$$

Toward a verification that $K$ and $\mu$ satisfying the requirements of the lemma, let

$$\Pi_n = \sum_{a_1 \cdots a_n \in S_{n,\delta}(p)} x_{a_1} x_{a_1}^* \otimes \cdots \otimes x_{a_n} x_{a_n}^*, \tag{4.18}$$

where $S_{n,\delta}(p)$ denotes the set of $\delta$-strongly typical sequences with respect to the probability vector $p$, for every positive integer $n$. The condition $\left[\Pi_n, \sigma^{\otimes n}\right] = 0$ is immediate, while the bound

$$\begin{aligned} \left\langle \Pi_n, \rho^{\otimes n} \right\rangle &= \sum_{a_1 \cdots a_n \in S_{n,\delta}(p)} p(a_1) \cdots p(a_n) \\ &\geq 1 - 2 \sum_{\substack{a \in \Sigma \\ p(a) > 0}} \exp\left(-2n\delta^2 p(a)^2\right) \geq 1 - K \exp(-\mu n) \end{aligned} \tag{4.19}$$

follows directly from Lemma 4.2.

It remains to prove the inequalities in (4.14). As

$$\Pi_n \sigma^{\otimes n} \Pi_n = \sum_{a_1 \cdots a_n \in S_{n,\delta}(p)} q(a_1) \cdots q(a_n) x_{a_1} x_{a_1}^* \otimes \cdots \otimes x_{a_n} x_{a_n}^*, \tag{4.20}$$

these inequalities are equivalent to

$$-(1-\delta) n \operatorname{Tr}(\rho \log(\sigma)) \leq -\sum_{k=1}^{n} \log(q(a_k)) \leq -(1+\delta) n \operatorname{Tr}(\rho \log(\sigma)) \tag{4.21}$$

for every $a_1 \cdots a_n \in S_{n,\delta}(p)$. By taking $\phi(a) = -\log(q(a))$ for every $a \in \Sigma$ in Lemma 4.3, so that

$$\sum_{a \in \Sigma} p(a) \phi(a) = -\operatorname{Tr}(\rho \log(\sigma)), \tag{4.22}$$

the inequalities (4.21) are obtained, which completes the proof. □

The next lemma is just a simple technical fact concerning the inner-product of a product of projection operators with a density operator.

**Lemma 4.5.** *Let $\rho \in \mathrm{D}(\mathcal{X})$ be a density operator, let $\varepsilon > 0$ be a positive real number, and let $\Pi$ and $\Delta$ be projection operators on $\mathcal{X}$ that satisfy $\langle \Pi, \rho \rangle \geq 1 - \varepsilon$ and $\langle \Delta, \rho \rangle \geq 1 - \varepsilon$. It is the case that*

$$\langle \Delta \Pi \Delta, \rho \rangle \geq 1 - 6\varepsilon. \tag{4.23}$$

*Proof.* Observe first that

$$\langle \Delta \Pi \Delta, \rho \rangle = \operatorname{Tr}\bigl(\Pi \Delta \rho \Delta \Pi\bigr) = \bigl\| \Pi \Delta \sqrt{\rho} \bigr\|_2^2 \geq \bigl| \bigl\langle \sqrt{\rho}, \Pi \Delta \sqrt{\rho} \bigr\rangle \bigr|^2 = |\langle \Delta \Pi, \rho \rangle|^2, \tag{4.24}$$

where the inequality is by the Cauchy–Schwarz inequality. Next, by the identity

$$\mathbb{1} = (\mathbb{1} - \Delta)(\mathbb{1} - \Pi) + \Delta + \Pi - \Delta \Pi, \tag{4.25}$$

one sees that

$$\begin{aligned} \langle \Delta \Pi, \rho \rangle &= \langle \Delta, \rho \rangle + \langle \Pi, \rho \rangle - 1 + \langle (\mathbb{1} - \Delta)(\mathbb{1} - \Pi), \rho \rangle \\ &\geq 1 - 2\varepsilon + \langle (\mathbb{1} - \Delta)(\mathbb{1} - \Pi), \rho \rangle. \end{aligned} \tag{4.26}$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned} \bigl| \langle (\mathbb{1} - \Delta)(\mathbb{1} - \Pi), \rho \rangle \bigr| &= \bigl| \bigl\langle (\mathbb{1} - \Pi) \sqrt{\rho}, (\mathbb{1} - \Delta) \sqrt{\rho} \bigr\rangle \bigr| \\ &\leq \bigl\| (\mathbb{1} - \Pi) \sqrt{\rho} \bigr\|_2 \bigl\| (\mathbb{1} - \Delta) \sqrt{\rho} \bigr\|_2 = \sqrt{\langle \mathbb{1} - \Pi, \rho \rangle} \sqrt{\langle \mathbb{1} - \Delta, \rho \rangle} \leq \varepsilon, \end{aligned} \tag{4.27}$$

and therefore

$$\langle (\mathbb{1} - \Delta)(\mathbb{1} - \Pi), \rho \rangle \geq -\varepsilon. \tag{4.28}$$

Consequently,

$$\big|\langle \Delta\Pi, \rho\rangle\big| \geq 1 - 3\varepsilon, \tag{4.29}$$

and therefore

$$\big\langle \Delta\Pi\Delta, \rho\big\rangle \geq (1 - 3\varepsilon)^2 \geq 1 - 6\varepsilon, \tag{4.30}$$

as required. □

**Remark 4.6.** If $\Delta$ commutes with $\rho$, then the bound established by the previous lemma can be improved to

$$\langle \Delta\Pi\Delta, \rho\rangle \geq 1 - 2\varepsilon. \tag{4.31}$$

To see that this is so, note that

$$\rho - \Delta\rho\Delta = (\mathbb{1} - \Delta)\rho(\mathbb{1} - \Delta), \tag{4.32}$$

and therefore

$$\langle \Delta\Pi\Delta, \rho\rangle = \langle \Pi, \rho\rangle + \langle \Pi, \Delta\rho\Delta - \rho\rangle = \langle \Pi, \rho\rangle - \langle \Pi, (\mathbb{1} - \Delta)\rho(\mathbb{1} - \Delta)\rangle. \tag{4.33}$$

Because $(\mathbb{1} - \Delta)\rho(\mathbb{1} - \Delta)$ is positive semidefinite and $\Pi \leq \mathbb{1}$, it follows that

$$\langle \Delta\Pi\Delta, \rho\rangle \geq \langle \Pi, \rho\rangle - \operatorname{Tr}\big((\mathbb{1} - \Delta)\rho(\mathbb{1} - \Delta)\big) = \langle \Pi, \rho\rangle - \langle \mathbb{1} - \Delta, \rho\rangle \geq 1 - 2\varepsilon. \tag{4.34}$$

**Remark 4.7.** The proof of the lemma above can be extended to the assumptions $\langle \Pi, \rho\rangle \geq 1 - \varepsilon$ and $\langle \Delta, \rho\rangle \geq 1 - \delta$ to obtain

$$\langle \Delta\Pi\Delta, \rho\rangle \geq \Big(1 - \varepsilon - \delta - \sqrt{\varepsilon\delta}\Big)^2. \tag{4.35}$$

We're also going to make use of Winter's gentle measurement lemma, which is a very useful and well-known fact.

**Lemma 4.8** (Winter's gentle measurement lemma)**.** *Let $\mathcal{X}$ be a complex Euclidean space, let $\rho \in \mathrm{D}(\mathcal{X})$ be a density operator, and let $P \in \mathrm{Pos}(\mathcal{X})$ be a positive semidefinite operator satisfying $P \leq \mathbb{1}$ and $\langle P, \rho\rangle > 0$. This inequality is satisfied:*

$$\mathrm{F}\left(\rho, \frac{\sqrt{P}\rho\sqrt{P}}{\langle P, \rho\rangle}\right) \geq \sqrt{\langle P, \rho\rangle}. \tag{4.36}$$

*Proof.* Observe that for any two positive semidefinite operators $Q$ and $R$, it is necessarily the case that $\mathrm{F}(R, QRQ) = \langle R, Q\rangle$. Indeed,

$$\sqrt{\sqrt{R}QRQ\sqrt{R}} = \sqrt{\big(\sqrt{R}Q\sqrt{R}\big)^2} = \sqrt{R}Q\sqrt{R}, \tag{4.37}$$

and therefore

$$\mathrm{F}(R, QRQ) = \operatorname{Tr}\left(\sqrt{\sqrt{R}QRQ\sqrt{R}}\right) = \operatorname{Tr}\left(\sqrt{R}Q\sqrt{R}\right) = \langle R, Q \rangle. \tag{4.38}$$

By this formula, along with the square root scaling of the fidelity function, one finds that

$$\mathrm{F}\left(\rho, \frac{\sqrt{P}\rho\sqrt{P}}{\langle P, \rho\rangle}\right) = \frac{1}{\sqrt{\langle P, \rho\rangle}}\,\mathrm{F}\left(\rho, \sqrt{P}\rho\sqrt{P}\right) = \frac{\langle\sqrt{P}, \rho\rangle}{\sqrt{\langle P, \rho\rangle}}. \tag{4.39}$$

Finally, under the assumption $0 \leq P \leq \mathbb{1}$, it is the case that $\sqrt{P} \geq P$, and therefore $\langle\sqrt{P}, \rho\rangle \geq \langle P, \rho\rangle$, from which the lemma follows. □

**Remark 4.9.** We will actually only need the lemma for $P$ being a projection operator, in which case $\sqrt{P} = P$, and so the lemma holds with equality in (4.36).

## 4.3 Main theorem

Now we're ready for the main theorem and its proof. Here it is.

**Theorem 4.10.** *Let $\rho, \sigma \in \mathrm{D}(\mathcal{X})$ be density operators. For every $\varepsilon \in (0, 1)$, the following equality holds.*

$$\lim_{n\to\infty} \frac{\mathrm{D}^{\varepsilon}_{\max}(\rho^{\otimes n} \| \sigma^{\otimes n})}{n} = \mathrm{D}(\rho\|\sigma). \tag{4.40}$$

*Proof.* Let us first consider the case that $\operatorname{im}(\rho) \not\subseteq \operatorname{im}(\sigma)$. In this case the right-hand side of (4.40) is infinite, and so we must prove the same for the left-hand side. This follows from the fact that $\mathrm{D}^{\varepsilon}_{\max}(\rho^{\otimes n} \| \sigma^{\otimes n})$ is infinite for all but finitely many positive integers $n$. Indeed, let $\Lambda$ be the projection onto $\operatorname{im}(\sigma)$, so that $\sigma = \Lambda\sigma\Lambda$ and $\langle\Lambda, \rho\rangle < 1$. Now, for a given choice of $n$, one has that if $\xi_n \in \mathrm{D}(\mathcal{X}^{\otimes n})$ satisfies $\operatorname{im}(\xi_n) \subseteq \operatorname{im}(\sigma^{\otimes n})$, then

$$\mathrm{F}(\xi_n, \rho^{\otimes n}) = \mathrm{F}(\xi_n, \Lambda^{\otimes n}\rho^{\otimes n}\Lambda^{\otimes n}) \leq \sqrt{\operatorname{Tr}(\Lambda^{\otimes n}\rho^{\otimes n}\Lambda^{\otimes n})} < \langle\Lambda, \rho\rangle^{\frac{n}{2}}, \tag{4.41}$$

and therefore

$$\frac{1}{2}\left\|\xi_n - \rho^{\otimes n}\right\|_1 > 1 - \langle\Lambda, \rho\rangle^{\frac{n}{2}} \tag{4.42}$$

by one of the Fuchs–van de Graaf inequalities. The right-hand side of this inequality must exceed $\varepsilon$ for all but finitely many positive integers $n$, by the fact that $\langle\Lambda, \rho\rangle < 1$. It follows that for all but finitely many positive integers $n$, there are no elements of $\mathcal{B}_{\varepsilon}(\rho^{\otimes n})$ whose images are contained in $\operatorname{im}(\sigma^{\otimes n})$, which implies for any such $n$ that $\mathrm{D}^{\varepsilon}_{\max}(\rho^{\otimes n} \| \sigma^{\otimes n}) = \infty$.

For the remainder of the proof it will be assumed that $\operatorname{im}(\rho) \subseteq \operatorname{im}(\sigma)$. Let $\delta > 0$ be chosen arbitrarily, and for each positive integer $n$, let $\Pi_n$ be the projection whose existence is guaranteed by Lemma 4.4 for $\rho$, $\sigma$, $\delta$, and $n$, and also let $\Delta_n$ be the projection whose existence is guaranteed by Lemma 4.4, again for $\rho$, $\delta$, and $n$, but where $\sigma$ is replaced by $\rho$. Thus, we have $[\Pi_n, \sigma^{\otimes n}] = 0$ and $[\Delta_n, \rho^{\otimes n}] = 0$, and the following inequalities are satisfied:

$$2^{(1+\delta)n \operatorname{Tr}(\rho \log(\sigma))} \Pi_n \leq \Pi_n \sigma^{\otimes n} \Pi_n \leq 2^{(1-\delta)n \operatorname{Tr}(\rho \log(\sigma))} \Pi_n, \tag{4.43}$$

$$2^{-(1+\delta)n \operatorname{H}(\rho)} \Delta_n \leq \Delta_n \rho^{\otimes n} \Delta_n \leq 2^{-(1-\delta)n \operatorname{H}(\rho)} \Delta_n, \tag{4.44}$$

$$\langle \Delta_n, \rho^{\otimes n} \rangle \geq 1 - K \exp(-\mu n), \tag{4.45}$$

$$\langle \Pi_n, \rho^{\otimes n} \rangle \geq 1 - K \exp(-\mu n), \tag{4.46}$$

where $K$ and $\mu$ are positive constants independent of $n$.

It will first be proved that

$$\lim_{n \to \infty} \frac{\operatorname{D}^{\varepsilon}_{\max}\big(\rho^{\otimes n} \big\| \sigma^{\otimes n}\big)}{n} \leq \operatorname{D}(\rho \| \sigma). \tag{4.47}$$

The projection operator $\Delta_n$ commutes with $\rho^{\otimes n}$, and therefore

$$\langle \Delta_n \Pi_n \Delta_n, \rho^{\otimes n} \rangle \geq 1 - 2K \exp(-\mu n) \tag{4.48}$$

by Remark 4.6. It is therefore the case that $\langle \Delta_n, \rho^{\otimes n} \rangle$, $\langle \Pi_n, \rho^{\otimes n} \rangle$, and $\langle \Delta_n \Pi_n \Delta_n, \rho^{\otimes n} \rangle$ are all positive for all but finitely many $n$, and we will restrict our attention to those $n$ for which these values are all positive.

Define a density operator

$$\xi_n = \frac{\Pi_n \Delta_n \rho^{\otimes n} \Delta_n \Pi_n}{\langle \Delta_n \Pi_n \Delta_n, \rho^{\otimes n} \rangle} \tag{4.49}$$

for all values of $n$ under consideration. We will begin by proving a bound on the trace distance between $\xi_n$ and $\rho^{\otimes n}$. To this end, observe that

$$\frac{1}{2} \big\| \xi_n - \rho^{\otimes n} \big\|_1 \leq \frac{1}{2} \big\| \xi_n - \tau_n \big\|_1 + \frac{1}{2} \big\| \tau_n - \rho^{\otimes n} \big\|_1 \tag{4.50}$$

for

$$\tau_n = \frac{\Delta_n \rho^{\otimes n} \Delta_n}{\big\langle \Delta_n, \rho^{\otimes n} \big\rangle}, \tag{4.51}$$

and notice that

$$\xi_n = \frac{\Pi_n \tau_n \Pi_n}{\langle \Pi_n, \tau_n \rangle}. \tag{4.52}$$

By Winter's gentle measurement lemma and one of the Fuchs–van de Graaf inequalities, we find that

$$\frac{1}{2}\big\|\xi_n - \tau_n\big\|_1 = \frac{1}{2}\left\|\frac{\Pi_n\tau_n\Pi_n}{\langle\Pi_n,\tau_n\rangle} - \tau_n\right\|_1 \leq \sqrt{1-\langle\Pi_n,\tau_n\rangle}, \tag{4.53}$$

and because

$$\langle\Pi_n,\tau_n\rangle = \frac{\langle\Delta_n\Pi_n\Delta_n,\rho^{\otimes n}\rangle}{\langle\Delta_n,\rho^{\otimes n}\rangle} \geq \langle\Delta_n\Pi_n\Delta_n,\rho^{\otimes n}\rangle, \tag{4.54}$$

we obtain

$$\frac{1}{2}\big\|\xi_n - \tau_n\big\|_1 \leq \sqrt{1-\langle\Delta_n\Pi_n\Delta_n,\rho^{\otimes n}\rangle} \leq \sqrt{2K\exp(-\mu n)}. \tag{4.55}$$

Along similar (although simpler) lines, we find that

$$\frac{1}{2}\big\|\tau_n - \rho^{\otimes n}\big\|_1 \leq \sqrt{1-\langle\Delta_n,\rho^{\otimes n}\rangle} \leq \sqrt{K\exp(-\mu n)}. \tag{4.56}$$

These upper bounds are decreasing to 0 (exponentially quickly, as it happens), and therefore

$$\frac{1}{2}\big\|\xi_n - \rho^{\otimes n}\big\|_1 \leq \varepsilon, \tag{4.57}$$

or equivalently $\xi_n \in \mathcal{B}_\varepsilon(\rho^{\otimes n})$, implying

$$\mathrm{D}^{\varepsilon}_{\max}\big(\rho^{\otimes n}\big\|\sigma^{\otimes n}\big) \leq \mathrm{D}_{\max}\big(\xi_n\big\|\sigma^{\otimes n}\big), \tag{4.58}$$

for all but finitely many $n$. Let us further restrict our attention to these values of $n$.

Next, we will use the inequalities (4.43) and (4.44) to obtain an upper-bound on $\mathrm{D}_{\max}\big(\xi_n\big\|\sigma^{\otimes n}\big)$. First, by (4.44), together with $\Delta_n \leq \mathbb{1}$ and $\Pi_n^2 = \Pi_n$, we find that

$$\Pi_n\Delta_n\rho^{\otimes n}\Delta_n\Pi_n \leq 2^{-(1-\delta)n\,\mathrm{H}(\rho)}\Pi_n\Delta_n\Pi_n \leq 2^{-(1-\delta)n\,\mathrm{H}(\rho)}\Pi_n. \tag{4.59}$$

Second, by (4.44), together with the fact that $[\Pi_n,\sigma^{\otimes n}] = 0$, we have

$$\Pi_n \leq 2^{-(1+\delta)n\,\mathrm{Tr}(\rho\log(\sigma))}\Pi_n\sigma^{\otimes n}\Pi_n \leq 2^{-(1+\delta)n\,\mathrm{Tr}(\rho\log(\sigma))}\sigma^{\otimes n}. \tag{4.60}$$

Combining (4.59) and (4.60), we obtain

$$\Pi_n\Delta_n\rho^{\otimes n}\Delta_n\Pi_n \leq 2^{n\,\mathrm{D}(\rho\|\sigma)+\delta n(\mathrm{H}(\rho)-\mathrm{Tr}(\rho\log(\sigma)))}\sigma^{\otimes n}. \tag{4.61}$$

Accounting for the normalization of $\xi_n$ and making use of (4.48), we find that

$$\begin{aligned}\mathrm{D}_{\max}\big(\xi_n\big\|\sigma^{\otimes n}\big) \leq n\,\mathrm{D}(\rho\|\sigma) + \delta n(\mathrm{H}(\rho) - \mathrm{Tr}(\rho\log(\sigma)))\\ - \log\big(1-2K\exp(-\mu n)\big).\end{aligned} \tag{4.62}$$

At this point we may conclude that

$$\lim_{n\to\infty} \frac{\mathrm{D}^{\varepsilon}_{\max}\big(\rho^{\otimes n}\big\|\sigma^{\otimes n}\big)}{n} \leq \mathrm{D}(\rho\|\sigma) + \delta(\mathrm{H}(\rho) - \mathrm{Tr}(\rho\log(\sigma))), \tag{4.63}$$

and as $\delta$ was an arbitrarily chosen positive real number, we obtain the required inequality (4.47).

Now we will prove the reverse inequality

$$\lim_{n\to\infty} \frac{\mathrm{D}^{\varepsilon}_{\max}\big(\rho^{\otimes n}\big\|\sigma^{\otimes n}\big)}{n} \geq \mathrm{D}(\rho\|\sigma). \tag{4.64}$$

Let $\delta > 0$ again be chosen arbitrarily. For every positive integer $n$ it is the case that

$$\big\langle \sigma^{\otimes n}, \Pi_n\Delta_n\Pi_n\big\rangle = \big\langle \Delta_n, \Pi_n\sigma^{\otimes n}\Pi_n\big\rangle \leq 2^{(1-\delta)n\,\mathrm{Tr}(\rho\log(\sigma))}\big\langle \Delta_n, \Pi_n\big\rangle \tag{4.65}$$

by (4.43), as well as

$$\big\langle \Delta_n, \Pi_n\big\rangle \leq 2^{(1+\delta)n\,\mathrm{H}(\rho)}\big\langle \Delta_n\rho^{\otimes n}\Delta_n, \Pi_n\big\rangle \leq 2^{(1+\delta)n\,\mathrm{H}(\rho)} \tag{4.66}$$

by (4.44). It therefore follows that the operator

$$Z_n = 2^{n\,\mathrm{D}(\rho\|\sigma)-\delta n(\mathrm{H}(\rho)-\mathrm{Tr}(\rho\log(\sigma)))}\Pi_n\Delta_n\Pi_n \tag{4.67}$$

is positive semidefinite and satisfies $\langle\sigma^{\otimes n}, Z\rangle \leq 1$. By an inspection of the dual form of Optimization Problem 3.2, the conic program for the exponential of the smoothed max-relative entropy, we conclude that

$$\begin{aligned}\mathrm{D}^{\varepsilon}_{\max}\big(\rho^{\otimes n}\big\|\sigma^{\otimes n}\big) \geq{} & n\,\mathrm{D}(\rho\|\sigma) - \delta n(\mathrm{H}(\rho) - \mathrm{Tr}(\rho\log(\sigma)))\\ & + \inf_{\xi_n\in\mathcal{B}_\varepsilon(\rho^{\otimes n})} \log\langle\Pi_n\Delta_n\Pi_n, \xi_n\rangle.\end{aligned} \tag{4.68}$$

For every $\xi_n \in \mathcal{B}_\varepsilon(\rho^{\otimes n})$ we have, by virtue of the fact that $\rho^{\otimes n} - \xi_n$ is traceless and $0 \leq \Pi_n\Delta_n\Pi_n \leq \mathbb{1}$, that

$$\big\langle \Pi_n\Delta_n\Pi_n, \rho^{\otimes n} - \xi_n\big\rangle \leq \frac{1}{2}\big\|\rho^{\otimes n} - \xi_n\big\|_1 \leq \varepsilon \tag{4.69}$$

and therefore

$$\begin{aligned}\langle\Pi_n\Delta_n\Pi_n, \xi_n\rangle &= \langle\Pi_n\Delta_n\Pi_n, \rho^{\otimes n}\rangle + \langle\Pi_n\Delta_n\Pi_n, \xi_n - \rho^{\otimes n}\rangle\\ &\geq 1 - 6K\exp(-\mu n) - \varepsilon\end{aligned} \tag{4.70}$$

by Lemma 4.5. Consequently,

$$\begin{aligned}\mathrm{D}^{\varepsilon}_{\max}\big(\rho^{\otimes n}\big\|\sigma^{\otimes n}\big) \geq{} & n\,\mathrm{D}(\rho\|\sigma) - \delta n(\mathrm{H}(\rho) - \mathrm{Tr}(\rho\log(\sigma)))\\ & + \log\big(1 - 6K\exp(-\mu n) - \varepsilon\big).\end{aligned} \tag{4.71}$$

Given the assumption $\varepsilon \in (0,1)$, one concludes that $\log(1 - 6K\exp(-\mu n) - \varepsilon)$ converges to a constant value as $n$ goes to infinity. It follows that

$$\lim_{n\to\infty} \frac{\mathrm{D}^{\varepsilon}_{\max}(\rho^{\otimes n} \| \sigma^{\otimes n})}{n} \geq \mathrm{D}(\rho\|\sigma) - \delta(\mathrm{H}(\rho) - \mathrm{Tr}(\rho\log(\sigma))). \tag{4.72}$$

Once again, as $\delta$ was an arbitrarily chosen positive real number, the required inequality (4.64) follows. □

**Remark 4.11.** An alternative way to argue the closeness of $\xi_n$ to $\rho^{\otimes n}$ is to use a different known equality concerning the fidelity function, which is that

$$\mathrm{F}\left(AA^*, BB^*\right) = \left\| A^*B \right\|_1 \tag{4.73}$$

for any choice of operators $A, B \in \mathrm{L}(\mathcal{X}, \mathcal{Y})$. (This fact is closely connected with Uhlmann's theorem, and can be found as Lemma 3.21 in my book *Theory of Quantum Information*.) In the present case, we obtain

$$\begin{aligned} \mathrm{F}\big(\Pi_n\Delta_n\rho^{\otimes n}\Delta_n\Pi_n, \rho^{\otimes n}\big) &= \left\| \sqrt{\rho^{\otimes n}}\Delta_n\Pi_n\sqrt{\rho^{\otimes n}} \right\|_1 \\ &\geq \left| \mathrm{Tr}\left( \sqrt{\rho^{\otimes n}}\Delta_n\Pi_n\sqrt{\rho^{\otimes n}} \right) \right| = \langle \Delta_n\Pi_n\Delta_n, \rho^{\otimes n} \rangle, \end{aligned} \tag{4.74}$$

where the last equality makes use of $[\Delta_n, \rho^{\otimes n}] = 0$. A suitable bound on the trace distance between $\xi_n$ and $\rho^{\otimes n}$ is obtained through the Fuchs–van de Graaf inequalities.

And we will conclude with two corollaries.

It is not important that $\sigma$ is a density operator in Theorem 4.10—it is true for arbitrary positive semidefinite models. The only part of the proof that depends on the scaling of $\sigma$ occurs in the proof of Lemma 4.4, where $q \in \mathcal{P}(\Sigma)$ implies that $\phi(a) = -\log(q(a))$ is nonnegative. Although it would not be difficult to modify this portion of the proof slightly to handle arbitrary positive semidefinite models, it is perhaps simpler to observe it as a fairly straightforward corollary of Theorem 4.10.

**Corollary 4.12.** *If $\rho$ is a density operator and $Q$ is any positive semidefinite operator, we have*

$$\lim_{n\to\infty} \frac{\mathrm{D}^{\varepsilon}_{\max}(\rho^{\otimes n} \| Q^{\otimes n})}{n} = \mathrm{D}(\rho\|Q). \tag{4.75}$$

*Proof.* Let $\sigma = Q/\mathrm{Tr}(Q)$. Then

$$\begin{aligned} \mathrm{D}^{\varepsilon}_{\max}(\rho^{\otimes n} \| Q^{\otimes n}) &= \mathrm{D}^{\varepsilon}_{\max}(\rho^{\otimes n} \| \mathrm{Tr}(Q)^n\sigma^{\otimes n}) \\ &= \mathrm{D}^{\varepsilon}_{\max}(\rho^{\otimes n} \| \sigma^{\otimes n}) - n\log(\mathrm{Tr}(Q)) \end{aligned}$$

so

$$\lim_{n\to\infty}\frac{\mathrm{D}^{\varepsilon}_{\max}\big(\rho^{\otimes n}\big\|Q^{\otimes n}\big)}{n} = \lim_{n\to\infty}\frac{\mathrm{D}^{\varepsilon}_{\max}\big(\rho^{\otimes n}\big\|\sigma^{\otimes n}\big)}{n} - \log(\mathrm{Tr}(Q))$$
$$= \mathrm{D}(\rho\|\sigma) - \log(\mathrm{Tr}(Q))$$
$$= \mathrm{D}(\rho\|Q),$$

as required. □

The second and final corollary is quite spectacular. Of course it is well-known, and we proved it in CS 766/QIC 820, but now we have a completely different alternative proof.

**Corollary 4.13** (Monotonicity of quantum relative entropy)**.** *Let* $\rho,\sigma \in \mathrm{D}(\mathcal{X})$ *be density operators and let* $\Phi \in \mathrm{C}(\mathcal{X},\mathcal{Y})$ *be a channel, for complex Euclidean spaces* $\mathcal{X}$ *and* $\mathcal{Y}$*. It is the case that* $\mathrm{D}(\Phi(\rho)\|\Phi(\sigma)) \leq \mathrm{D}(\rho\|\sigma)$*.*

*Proof.* Observe first that the smoothed max-relative entropy is monotonic:

$$\begin{aligned}\mathrm{D}^{\varepsilon}_{\max}(\Phi(\rho)\|\Phi(\sigma)) &\leq \inf_{\xi\in\mathcal{B}_{\varepsilon}(\rho)} \mathrm{D}_{\max}(\Phi(\xi)\|\Phi(\sigma))\\ &\leq \inf_{\xi\in\mathcal{B}_{\varepsilon}(\rho)} \mathrm{D}_{\max}(\xi\|\sigma) = \mathrm{D}^{\varepsilon}_{\max}(\rho\|\sigma)\end{aligned} \tag{4.76}$$

by the monotonicity of the max-relative entropy. Therefore,

$$\begin{aligned}\mathrm{D}(\Phi(\rho)\|\Phi(\sigma)) &= \lim_{n\to\infty}\frac{\mathrm{D}^{\varepsilon}_{\max}\big(\Phi(\rho)^{\otimes n}\big\|\Phi(\sigma)^{\otimes n}\big)}{n}\\ &= \lim_{n\to\infty}\frac{\mathrm{D}^{\varepsilon}_{\max}\big(\Phi^{\otimes n}(\rho^{\otimes n})\big\|\Phi^{\otimes n}(\sigma^{\otimes n})\big)}{n}\\ &\leq \lim_{n\to\infty}\frac{\mathrm{D}^{\varepsilon}_{\max}\big(\rho^{\otimes n}\big\|\sigma^{\otimes n}\big)}{n}\\ &= \mathrm{D}(\rho\|\sigma),\end{aligned} \tag{4.77}$$

as required. □

Lecture 5

# Min-relative entropy, conditional max-entropy, and hypothesis-testing relative entropy

In this lecture we will discuss a few additional generalized entropy measures, namely the min-relative entropy, the conditional max-entropy, and the hypothesis-testing relative entropy. We will discuss various properties of these quantities, and relate them to the other entropic quantities we have previously discussed.

## 5.1 Min-relative entropy

We will begin with the min-relative entropy.

**Definition 5.1** (Quantum min-relative entropy)**.** Let $\rho \in \mathrm{D}(\mathcal{X})$ be a density operator and let $Q \in \mathrm{Pos}(\mathcal{X})$ be a positive semidefinite operator, for $\mathcal{X}$ a complex Euclidean space. The *quantum min-relative entropy* (or min-relative entropy, for short) of $\rho$ with respect to $Q$ is defined as

$$\mathrm{D}_{\min}(\rho \| Q) = -\log\bigl(\mathrm{F}(\rho, Q)^2\bigr), \tag{5.1}$$

where

$$\mathrm{F}(\rho, Q) = \Bigl\| \sqrt{\rho}\sqrt{Q} \Bigr\|_1 \tag{5.2}$$

is the fidelity between $\rho$ and $Q$.

**Remark 5.2.** In the case that $\mathrm{F}(\rho, Q) = 0$, which is equivalent to $\mathrm{im}(\rho) \perp \mathrm{im}(Q)$, one is to interpret that $\mathrm{D}_{\min}(\rho \| Q) = \infty$.

## Elementary observations

Here are a couple of relevant properties of the min-relative entropy that follow directly from known properties of the fidelity function.

1. A variant of Klein's inequality holds for the min-relative entropy. That is, if $\rho, \sigma \in \mathrm{D}(\mathcal{X})$ are density operators, then $\mathrm{D}_{\mathrm{min}}(\rho\|\sigma) \geq 0$, with equality if and only if $\rho = \sigma$.
2. The min-relative entropy is monotonic with respect to the action of channels. That is, for every choice of $\rho \in \mathrm{D}(\mathcal{X})$, $Q \in \mathrm{Pos}(\mathcal{X})$, and $\Phi \in \mathrm{C}(\mathcal{X}, \mathcal{Y})$, it is the case that
$$\mathrm{D}_{\mathrm{min}}(\Phi(\rho)\|\Phi(Q)) \leq \mathrm{D}_{\mathrm{min}}(\rho\|Q).$$
This is true, in fact, for all positive and trace-preserving maps $\Phi \in \mathrm{T}(\mathcal{X}, \mathcal{Y})$.

One can identify additional properties of the min-relative entropy through its very direct connection to the fidelity function, which we know to have many interesting and remarkable properties.

## Relationship to the quantum relative entropy

The following theorem reveals that the min-relative entropy is upper-bounded by the ordinary quantum relative entropy.

**Theorem 5.3.** *Let $\rho \in \mathrm{D}(\mathcal{X})$ be a density operator and let $Q \in \mathrm{Pos}(\mathcal{X})$ be a positive semidefinite operator, for $\mathcal{X}$ a complex Euclidean space. It is the case that*

$$\mathrm{D}_{\mathrm{min}}(\rho\|Q) \leq \mathrm{D}(\rho\|Q). \tag{5.3}$$

*Proof.* The theorem is trivial in the case $\mathrm{im}(\rho) \not\subseteq \mathrm{im}(Q)$, as the right-hand side of (5.3) is infinite in this case, so the remainder of the proof is focused on the case $\mathrm{im}(\rho) \subseteq \mathrm{im}(Q)$. There is no loss of generality in assuming that $Q$ is positive definite in this case, as the values of $\mathrm{D}_{\mathrm{min}}(\rho\|Q)$ and $\mathrm{D}(\rho\|Q)$ then do not change if $\mathcal{X}$ is replaced by $\mathrm{im}(Q)$.

Define a function $\phi : (-1, 1) \rightarrow \mathbb{R}$ as

$$\phi(\alpha) = -\ln \mathrm{Tr}(\rho^{1-\alpha} Q^{\alpha}). \tag{5.4}$$

We are using the natural logarithm because it will simplify the calculus that will soon be considered. It is not really important to the proof that this function is defined on the entire interval $(-1, 1)$, we only require that the function is defined on the interval $[0,1/2]$ and is differentiable at $\alpha = 0$. But, in any case, $\phi$ is differentiable at every point $\alpha \in (-1,1)$, with the derivative being given by

$$\phi'(\alpha) = \frac{\operatorname{Tr}\bigl(\rho^{1-\alpha} Q^{\alpha} (\ln(\rho) - \ln(Q))\bigr)}{\operatorname{Tr}\bigl(\rho^{1-\alpha} Q^{\alpha}\bigr)}. \tag{5.5}$$

Notice in particular that

$$\phi'(0) = \operatorname{Tr}(\rho \ln(\rho)) - \operatorname{Tr}(\rho \ln(Q)) = \frac{1}{\log(e)} \operatorname{D}(\rho \| Q). \tag{5.6}$$

We also observe that

$$\phi(1/2) = -\ln \operatorname{Tr}\Bigl(\sqrt{\rho}\sqrt{Q}\Bigr) \geq \frac{1}{2\log(e)} \operatorname{D}_{\min}(\rho \| Q) \tag{5.7}$$

with the inequality following from

$$\operatorname{F}(\rho, Q) = \Bigl\| \sqrt{\rho}\sqrt{Q} \Bigr\|_1 \geq \operatorname{Tr}\Bigl(\sqrt{\rho}\sqrt{Q}\Bigr). \tag{5.8}$$

Finally, noting that $\phi(0) = 0$, we see that the theorem will follow from a demonstration that

$$\phi'(0) \geq \frac{\phi(1/2) - \phi(0)}{1/2}. \tag{5.9}$$

This in turn will follow from a demonstration that $\phi$ is a concave function.

To prove that $\phi$ is concave, it suffices to compute its second derivative and observe that its value is non-positive. To make this as simple as possible, and to avoid a messy calculation, let us use the spectral theorem to write

$$\rho = \sum_{a \in \Sigma} p(a) x_a x_a^* \quad \text{and} \quad Q = \sum_{b \in \Gamma} q(b) y_b y_b^* \tag{5.10}$$

for alphabets $\Sigma$ and $\Gamma$, orthonormal sets $\{x_a : a \in \Sigma\}$ and $\{y_b : b \in \Gamma\}$, a probability vector $p \in \mathcal{P}(\Sigma)$, and $q \in (0,\infty)^{\Gamma}$ being a vector of positive real numbers. Let us also define a function

$$r_{a,b}(\alpha) = \frac{|\langle x_a, y_b \rangle|^2 p(a)^{1-\alpha} q(b)^{\alpha}}{\operatorname{Tr}(\rho^{1-\alpha} Q^{\alpha})} \tag{5.11}$$

for every $(a,b) \in \Sigma \times \Gamma$ and $\alpha \in (-1,1)$, so that

$$\phi'(\alpha) = \sum_{(a,b) \in \Sigma \times \Gamma} r_{a,b}(\alpha) \bigl(\ln(p(a)) - \ln(q(b))\bigr). \tag{5.12}$$

We may then express the second derivative of $\phi$ as

$$\begin{aligned}\phi''(\alpha) = \Biggl(\sum_{(a,b)\in\Sigma\times\Gamma} r_{a,b}(\alpha)\bigl(\ln(p(a)) - \ln(q(b))\bigr)\Biggr)^2 \\ - \sum_{(a,b)\in\Sigma\times\Gamma} r_{a,b}(\alpha)\bigl(\ln(p(a) - \ln(q(b))\bigr)^2.\end{aligned} \tag{5.13}$$

Observing that $r_{a,b}(\alpha) \geq 0$ and

$$\sum_{(a,b)\in\Sigma\times\Gamma} r_{a,b}(\alpha) = 1 \tag{5.14}$$

for every $\alpha \in (-1, 1)$, we find that $\phi''(\alpha)$ is non-positive by Jensen's inequality, which completes the proof. □

## 5.2 Conditional max-entropy

Next we will discuss the conditional max-entropy, which is defined through the min-relative entropy in precisely the same way that the conditional min-entropy is defined through the max-relative entropy.

**Definition 5.4** (Conditional max-entropy)**.** Let $\rho \in \mathrm{D}(\mathcal{X}\otimes\mathcal{Y})$ be a state of a pair of registers $(\mathsf{X}, \mathsf{Y})$. The *conditional max-entropy* of $\mathsf{X}$ given $\mathsf{Y}$ for the state $\rho$ is defined as

$$\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Y})_\rho = -\inf_{\sigma\in\mathrm{D}(\mathcal{Y})} \mathrm{D}_{\min}(\rho\|\mathbb{1}_{\mathcal{X}}\otimes\sigma). \tag{5.15}$$

Equivalently,

$$\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Y})_\rho = \sup_{\sigma\in\mathrm{D}(\mathcal{Y})} \log\bigl(\mathrm{F}(\rho, \mathbb{1}_{\mathcal{X}}\otimes\sigma)^2\bigr). \tag{5.16}$$

We obtain from Theorem 5.3 that $\mathrm{H}(\mathsf{X}|\mathsf{Y})_\rho \leq \mathrm{H}_{\max}(\mathsf{X}|\mathsf{Y})_\rho$, and so

$$\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})_\rho \leq \mathrm{H}(\mathsf{X}|\mathsf{Y})_\rho \leq \mathrm{H}_{\max}(\mathsf{X}|\mathsf{Y})_\rho. \tag{5.17}$$

Writing $n = \dim(\mathcal{X})$ and $\omega = \mathbb{1}_{\mathcal{X}}/n$, we see from the definition of the conditional max-entropy that

$$\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Y})_\rho = \log(n) + \sup_{\sigma\in\mathrm{D}(\mathcal{Y})} \log\bigl(\mathrm{F}(\rho, \omega\otimes\sigma)^2\bigr). \tag{5.18}$$

## Semidefinite program for conditional max-entropy

One way to compute the value $\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Y})_\rho$ is to use the semidefinite program for the fidelity function that was discussed in CS 766/QIC 820, obtaining the following semidefinite program.

**Optimization Problem 5.5** (SDP for conditional max-entropy)

*Primal problem*

$$\begin{aligned}
\text{maximize:}\quad & \frac{1}{2}\operatorname{Tr}(X) + \frac{1}{2}\operatorname{Tr}(X^*)\\
\text{subject to:}\quad & \begin{pmatrix} \rho & X \\ X^* & \mathbb{1}_{\mathcal{X}} \otimes \sigma \end{pmatrix} \geq 0,\\
& X \in \mathrm{L}(\mathcal{X}\otimes\mathcal{Y}),\\
& \sigma \in \mathrm{D}(\mathcal{Y})
\end{aligned}$$

*Dual problem*

$$\begin{aligned}
\text{minimize:}\quad & \frac{1}{2}\langle \rho, Y\rangle + \frac{1}{2}\lambda_1(\operatorname{Tr}_{\mathcal{X}}(Z))\\
\text{subject to:}\quad & \begin{pmatrix} Y & -\mathbb{1}_{\mathcal{X}\otimes\mathcal{Y}} \\ -\mathbb{1}_{\mathcal{X}\otimes\mathcal{Y}} & Z \end{pmatrix} \geq 0,\\
& Y, Z \in \mathrm{Pos}(\mathcal{X}\otimes\mathcal{Y}).
\end{aligned}$$

It is the case that

$$\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Y})_\rho = 2\log(\alpha), \tag{5.19}$$

for

$$\alpha = \sup_{\sigma\in\mathrm{D}(\mathcal{Y})} \mathrm{F}(\rho, \mathbb{1}_{\mathcal{X}}\otimes\sigma) \tag{5.20}$$

being the optimal value of this semidefinite program.

**Remark 5.6.** The dual problem may be simplified to obtain the expression

$$\alpha = \inf_{Z>0}\left(\frac{\langle\rho, Z\rangle}{2} + \frac{\|\operatorname{Tr}_{\mathcal{X}}(Z^{-1})\|}{2}\right) \tag{5.21}$$

for the optimal value of Optimization Problem 5.5. Using the arithmetic-geometric mean inequality, one may conclude that

$$\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Y})_\rho = \inf_{Z>0}\bigl(\log\langle\rho, Z\rangle + \log\|\operatorname{Tr}_{\mathcal{X}}(Z^{-1})\|\bigr) \tag{5.22}$$

## Examples

We may again consider a few examples of classes of states, to gain some intuition on the conditional max-entropy.

**Example 5.7.** For any choice of $\sigma \in \mathrm{D}(\mathcal{X})$ and $\xi \in \mathrm{D}(\mathcal{Y})$, it is the case that

$$\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Y})_{\sigma\otimes\xi} = 2\log\operatorname{Tr}\sqrt{\sigma}. \tag{5.23}$$

**Example 5.8.** Calculating the conditional max-entropy $\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Y})_\rho$ for a classical-quantum state $\rho$ of the form

$$\rho = \sum_{k=1}^{n} p_k |k\rangle\langle k| \otimes \xi_k \tag{5.24}$$

yields

$$\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Y})_\rho = \log \sup_{\sigma \in \mathrm{D}(\mathcal{Y})} \left( \sum_{k=1}^{n} \sqrt{p_k}\, \mathrm{F}(\xi_k, \sigma) \right)^2. \tag{5.25}$$

**Example 5.9.** Let $n = \dim(\mathcal{X})$, let

$$\tau = \frac{1}{n} \sum_{a,b=1}^{n} |a\rangle\langle b| \otimes |a\rangle\langle b| \tag{5.26}$$

and suppose that $\tau$ can be recovered perfectly by applying a channel locally to $\mathsf{Y}$ for the state $\rho \in \mathrm{D}(\mathcal{X} \otimes \mathcal{Y})$. This is equivalent to $\rho$ taking the form

$$\rho = (\mathbb{1}_{\mathcal{X}} \otimes V)(\tau \otimes \xi)(\mathbb{1}_{\mathcal{X}} \otimes V)^* \tag{5.27}$$

for some choice of a density operator $\xi \in \mathrm{D}(\mathcal{Z})$ and an isometry $V \in \mathrm{U}(\mathcal{X} \otimes \mathcal{Z}, \mathcal{Y})$. Then we have

$$\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Y})_\rho = -\log(n), \tag{5.28}$$

just like the conditional min-entropy and conditional quantum entropy.

## A relationship between conditional min- and max-entropy

We will now prove a fundamental relationship between conditional min- and max-entropy, which is stated in the following theorem.

**Theorem 5.10.** *Let* $\mathsf{X}$, $\mathsf{Y}$, *and* $\mathsf{Z}$ *be registers and assume the triple* $(\mathsf{X}, \mathsf{Y}, \mathsf{Z})$ *is in a pure state* $uu^*$, *for* $u \in \mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{Z}$ *a unit vector. It is the case that*

$$\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y}) + \mathrm{H}_{\max}(\mathsf{X}|\mathsf{Z}) = 0. \tag{5.29}$$

*Proof.* First let us prove

$$2^{-\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})} \leq 2^{\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Z})}. \tag{5.30}$$

Let

$$\rho = \operatorname{Tr}_{\mathcal{Z}}(uu^*) \tag{5.31}$$

and choose $\Phi \in \mathrm{C}(\mathcal{Y}, \mathcal{X})$ to be a channel for which

$$2^{-\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})} = \mathrm{F}\big((\mathbb{1}_{\mathrm{L}(\mathcal{X})} \otimes \Phi)(\rho), \operatorname{vec}(\mathbb{1}_{\mathcal{X}}) \operatorname{vec}(\mathbb{1}_{\mathcal{X}})^*\big)^2. \tag{5.32}$$

Among the many nice properties that the fidelity function possesses is the fact that if $Q_0, Q_1 \in \operatorname{Pos}(\mathcal{U})$ and $P_0 \in \operatorname{Pos}(\mathcal{U} \otimes \mathcal{V})$ satisfies $\operatorname{Tr}_{\mathcal{V}}(P_0) = Q_0$, then

$$\mathrm{F}(Q_0, Q_1) = \max\bigl\{\mathrm{F}(P_0, P_1) \,:\, P_1 \in \operatorname{Pos}(\mathcal{U} \otimes \mathcal{V}),\ \operatorname{Tr}_{\mathcal{V}}(P_1) = Q_1\bigr\}. \tag{5.33}$$

From this fact we find that

$$2^{-\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})} = \mathrm{F}\bigl(\bigl(\mathbb{1}_{\mathrm{L}(\mathcal{X})} \otimes \Phi \otimes \mathbb{1}_{\mathrm{L}(\mathcal{Z})}\bigr)(uu^*), \operatorname{vec}(\mathbb{1}_{\mathcal{X}}) \operatorname{vec}(\mathbb{1}_{\mathcal{X}})^* \otimes \sigma\bigr)^2 \tag{5.34}$$

for some state $\sigma \in \mathrm{D}(\mathcal{Z})$, as operators of the form $\operatorname{vec}(\mathbb{1}_{\mathcal{X}}) \operatorname{vec}(\mathbb{1}_{\mathcal{X}})^* \otimes \sigma$ are the only operators that leave $\operatorname{vec}(\mathbb{1}_{\mathcal{X}}) \operatorname{vec}(\mathbb{1}_{\mathcal{X}})^*$ when $\mathcal{Z}$ is traced out. The fidelity is nondecreasing under the partial trace on the second tensor factor of $\mathcal{X}$, and therefore

$$2^{-\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})} \leq \mathrm{F}\bigl(\operatorname{Tr}_{\mathcal{Y}}(uu^*), \mathbb{1}_{\mathcal{X}} \otimes \sigma\bigr)^2 \leq 2^{\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Z})}. \tag{5.35}$$

Now let us prove the reverse inequality

$$2^{\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Z})} \leq 2^{-\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})}, \tag{5.36}$$

which is based on a similar idea. Choose a density operator $\sigma \in \mathrm{D}(\mathcal{Z})$ so that

$$2^{\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Z})} = \mathrm{F}\bigl(\operatorname{Tr}_{\mathcal{Y}}(uu^*), \mathbb{1}_{\mathcal{X}} \otimes \sigma\bigr)^2. \tag{5.37}$$

The operator $\mathbb{1}_{\mathcal{X}} \otimes \sigma$ can be purified as

$$\operatorname{vec}\bigl(\mathbb{1}_{\mathcal{X}} \otimes \sqrt{\sigma}\bigr) \operatorname{vec}\bigl(\mathbb{1}_{\mathcal{X}} \otimes \sqrt{\sigma}\bigr)^* \in \operatorname{Pos}(\mathcal{X} \otimes \mathcal{Z} \otimes \mathcal{X} \otimes \mathcal{Z}). \tag{5.38}$$

Every extension of $\operatorname{Tr}_{\mathcal{Y}}(uu^*)$ to an element of $\operatorname{Pos}(\mathcal{X} \otimes \mathcal{Z} \otimes \mathcal{X} \otimes \mathcal{Z})$ can be expressed as

$$\bigl(\mathbb{1}_{\mathrm{L}(\mathcal{X})} \otimes \Xi \otimes \mathbb{1}_{\mathrm{L}(\mathcal{Z})}\bigr)(uu^*) \tag{5.39}$$

for a channel $\Xi \in \mathrm{C}(\mathcal{Y}, \mathcal{Z} \otimes \mathcal{X})$, meaning that these are exactly the operators that leave $\operatorname{Tr}_{\mathcal{Y}}(uu^*)$ when the first tensor factor of $\mathcal{Z}$ and the second tensor factor of $\mathcal{X}$ are traced out. By the same fact regarding the fidelity function from before, we find that

$$2^{\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Z})} = \mathrm{F}\bigl(\bigl(\mathbb{1}_{\mathrm{L}(\mathcal{X})} \otimes \Xi \otimes \mathbb{1}_{\mathrm{L}(\mathcal{Z})}\bigr)(uu^*), \operatorname{vec}\bigl(\mathbb{1}_{\mathcal{X}} \otimes \sqrt{\sigma}\bigr) \operatorname{vec}\bigl(\mathbb{1}_{\mathcal{X}} \otimes \sqrt{\sigma}\bigr)^*\bigr)^2 \tag{5.40}$$

for some choice of a channel $\Xi \in \mathrm{C}(\mathcal{Y}, \mathcal{Z} \otimes \mathcal{X})$. Because the fidelity is nondecreasing under the partial trace on both copies of $\mathcal{Z}$, we obtain

$$2^{\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Z})} \leq \mathrm{F}\bigl(\bigl(\mathbb{1}_{\mathrm{L}(\mathcal{X})} \otimes \Phi\bigr)(\operatorname{Tr}_{\mathcal{Z}}(uu^*)), \operatorname{vec}\bigl(\mathbb{1}_{\mathcal{X}}\bigr) \operatorname{vec}\bigl(\mathbb{1}_{\mathcal{X}}\bigr)^*\bigr)^2 \tag{5.41}$$

for $\Phi = \operatorname{Tr}_{\mathcal{Z}} \circ\, \Xi \in \mathrm{C}(\mathcal{Y}, \mathcal{X})$. This implies

$$2^{\mathrm{H}_{\max}(\mathsf{X}|\mathsf{Z})} \leq 2^{-\mathrm{H}_{\min}(\mathsf{X}|\mathsf{Y})}, \tag{5.42}$$

as required. □

## 5.3 Hypothesis-testing relative entropy

We will define the *hypothesis-testing relative entropy* as follows.

**Definition 5.11.** Let $\rho \in \mathrm{D}(\mathcal{X})$, $Q \in \mathrm{Pos}(\mathcal{X})$, and $\varepsilon \in [0,1]$. The *$\varepsilon$-hypothesis-testing relative entropy* of $\rho$ with respect to $Q$ is defined as

$$\mathrm{D}_{\mathrm{H}}^{\varepsilon}(\rho \| Q) = -\inf\bigl\{\log\langle Q, X\rangle \,:\, X \in \mathrm{Pos}(\mathcal{X}),\, \langle \rho, X\rangle \geq 1,\, \varepsilon X \leq \mathbb{1}\bigr\}. \tag{5.43}$$

### Elementary observations

Before we try to understand the intuitive meaning of this quantity, let us note a few simple things about it.

First, we see that $\mathrm{D}_{\mathrm{H}}^{\varepsilon}(\rho\|Q) = \infty$ is possible:

1. $\mathrm{D}_{\mathrm{H}}^{0}(\rho\|Q) = \infty$ if and only if $\mathrm{im}(\rho) \not\subseteq \mathrm{im}(Q)$.
2. For $\varepsilon \in (0,1)$ we have $\mathrm{D}_{\mathrm{H}}^{\varepsilon}(\rho\|Q) = \infty$ if and only if
$$\bigl\langle \Pi_{\ker(Q)}, \rho \bigr\rangle \geq \varepsilon. \tag{5.44}$$
3. $\mathrm{D}_{\mathrm{H}}^{1}(\rho\|Q) = \infty$ if and only if $\mathrm{im}(\rho) \perp \mathrm{im}(Q)$.

Next, notice that as $\varepsilon$ decreases, the infimum decreases, because decreasing $\varepsilon$ means relaxing the constraint $\varepsilon X \leq \mathbb{1}$, and therefore the $\varepsilon$-hypothesis-testing relative entropy increases: $\delta \leq \varepsilon$ implies $\mathrm{D}_{\mathrm{H}}^{\delta}(\rho\|Q) \geq \mathrm{D}_{\mathrm{H}}^{\varepsilon}(\rho\|Q)$. Stated another way, taking $\varepsilon$ to be smaller means taking the $\varepsilon$-hypothesis-testing relative entropy to be a stronger notion of divergence. (The same may be said about the $\varepsilon$-smoothed max-relative entropy.)

Continuing on, although our primary focus will be on the range of values $\varepsilon \in (0,1)$, we will take a moment to consider the extreme cases $\varepsilon = 0$ and $\varepsilon = 1$. If it is the case that $\varepsilon = 0$, the constrain $\varepsilon X \leq \mathbb{1}$ is trivially satisfied. We obtain precisely the max-relative entropy, as an examination of the dual form of Optimization Problem 2.4 reveals:

$$\mathrm{D}_{\mathrm{H}}^{0}(\rho\|Q) = \mathrm{D}_{\max}(\rho\|Q). \tag{5.45}$$

At the other extreme, we may consider $\varepsilon = 1$, so that the constraint $\varepsilon X \leq \mathbb{1}$ becomes $X \leq \mathbb{1}$. The infimum is evidently achieved for $X = \Pi_{\mathrm{im}(\rho)}$, and so we obtain

$$\mathrm{D}_{\mathrm{H}}^{1}(\rho\|Q) = -\log\bigl\langle Q, \Pi_{\mathrm{im}(\rho)}\bigr\rangle. \tag{5.46}$$

This quantity has also been called the min-relative entropy by some, but obviously we will not use this name given that we have already used it for something else—the name *1-hypothesis-testing relative entropy* will do just fine.

## Semidefinite programming characterization

The definition of the hypothesis-testing relative entropy immediately suggests a semidefinite programming characterization. Specifically, the value $\mathrm{D}_{\mathrm{H}}^{\varepsilon}(\rho\|Q)$ is the negative logarithm of the optimal value of the following semidefinite program.

**Optimization Problem 5.12** (SDP for hypothesis-testing relative entropy)

*Primal problem*

$$\begin{aligned}\text{minimize:}\quad & \langle Q, X\rangle\\ \text{subject to:}\quad & \langle \rho, X\rangle \geq 1,\\ & \varepsilon X \leq \mathbb{1},\\ & X \in \mathrm{Pos}(\mathcal{X}).\end{aligned}$$

*Dual problem*

$$\begin{aligned}\text{maximize:}\quad & \lambda - \mathrm{Tr}(Y)\\ \text{subject to:}\quad & \lambda\rho \leq Q + \varepsilon Y,\\ & Y \in \mathrm{Pos}(\mathcal{X}),\\ & \lambda \geq 0.\end{aligned}$$

The primal problem is strictly feasible provided that $\varepsilon \in [0,1)$. In particular,

$$X = \begin{cases} \frac{(1+\varepsilon)\mathbb{1}}{2\varepsilon} & \text{if } \varepsilon \in (0,1)\\ 2\mathbb{1} & \text{if } \varepsilon = 0\end{cases} \tag{5.47}$$

is strictly primal feasible. Strict primal feasibility is, on the other hand, impossible when $\varepsilon = 1$. The dual problem is strictly feasible when $\varepsilon \in (0,1]$, by $\lambda = 1$ and $Y = 2\mathbb{1}/\varepsilon$ for instance. In the case $\varepsilon = 0$, the dual problem is strictly feasible if and only if $\mathrm{im}(\rho) \subseteq \mathrm{im}(Q)$.

Therefore, strong duality holds and the optimal values are achieved for all choices of $\rho$ and $Q$ when $\varepsilon \in (0,1)$, by Slater's theorem. We also have strong duality and an optimal value achieved in the primal problem when $\varepsilon = 1$, again by Slater's theorem, as $X = \mathbb{1}$ is primal feasible (although not strictly so); and strong duality also holds in the case $\varepsilon = 0$, as our examination of Optimization Problem 2.4 has already revealed.

## Interpretation

One way to interpret the $\varepsilon$-hypothesis-testing relative entropy, at least in the case $\varepsilon > 0$, begins with observation that

$$\mathrm{D}_{\mathrm{H}}^{\varepsilon}(\rho\|Q) = -\inf\biggl\{\log\frac{\langle Q, P\rangle}{\varepsilon} \;:\; 0 \leq P \leq \mathbb{1},\; \langle \rho, P\rangle \geq \varepsilon\biggr\}. \tag{5.48}$$

For the sake of the discussion that follows, let us consider the case that $Q = \sigma$ is a density operator.

We can now consider a test being performed that aims to distinguish between the states $\rho$ and $\sigma$. Think of $\sigma$ as representing an idealized model for a state, corresponding to the null-hypothesis of the test being performed, whereas $\rho$ is the actual state of the system being tested. The operator $P$ may be associated with a measurement operator, the outcome corresponding to which is to be seen as a signal supporting an alternative hypothesis. If we were to measure $\rho$ with respect to such a measurement, the alternative hypothesis may not be signaled with high probability, but the probability is at least $\varepsilon$. Think of the probability $\varepsilon$ as representing how small of a signal one is willing to tolerate in support of an alternative hypothesis. The value $2^{-\mathrm{D}_{\mathrm{H}}^{\varepsilon}(\rho\|Q)}$ is then equal to the smallest possible value for $\langle Q, P\rangle/\varepsilon$ that we could achieve by selecting $P$ optimally. Informally speaking, $\mathrm{D}_{\mathrm{H}}^{\varepsilon}(\rho\|\sigma)$ is a measure of how surprising it would be to obtain the outcome corresponding to $P$ in the idealized case represented by $\sigma$.

The precise scaling in the definition above has been selected so that a variant of Klein's inequality holds, provided $\varepsilon \in [0,1)$. That is, for all $\varepsilon \in [0,1)$ it is the case that $\mathrm{D}_{\mathrm{H}}^{\varepsilon}(\rho\|\sigma) \geq 0$, with equality if and only if $\rho = \sigma$. Indeed, if $\sigma \neq \rho$, we may choose a unit vector $u$ for which $\langle uu^*, \rho\rangle > \langle uu^*, \sigma\rangle$, and then consider

$$X = (1-\delta)\mathbb{1} + \frac{\delta}{\langle uu^*, \rho\rangle}\, uu^* \tag{5.49}$$

in the primal problem, for $\delta \in (0,1)$. The operator $X$ is clearly positive semidefinite, and it is the case that $\langle \sigma, X\rangle < \langle \rho, X\rangle = 1$. The constraint $\varepsilon X \leq \mathbb{1}$ is satisfied so long as

$$1 - \delta + \frac{\delta}{\langle uu^*, \rho\rangle} \leq \frac{1}{\varepsilon}, \tag{5.50}$$

which is so for all sufficiently small $\delta$, as $1/\varepsilon$ is strictly larger than 1 by the assumption $\varepsilon < 1$. By selecting any such $\delta$, one obtains a primal feasible $X$ having objective value strictly smaller than 1, which implies that $\mathrm{D}_{\mathrm{H}}^{\varepsilon}(\rho\|\sigma)$ is positive. In the case $\varepsilon = 1$, one has $\mathrm{D}_{\mathrm{H}}^{1}(\rho\|\sigma) = 0$ if and only if $\mathrm{im}(\sigma) \subseteq \mathrm{im}(\rho)$.

## Monotonicity under the action of channels

Suppose that $\Phi$ is a trace-preserving and positive map. Consider the dual form of Optimization Problem 5.12. If it is the case that $\lambda \geq 0$ and $Y \in \mathrm{Pos}(\mathcal{X})$ satisfy the constraint

$$\lambda\rho \leq Q + \varepsilon Y, \tag{5.51}$$

then by the positivity of $\Phi$ it must also hold that

$$\lambda\Phi(\rho) \leq \Phi(Q) + \varepsilon\Phi(Y), \tag{5.52}$$

and therefore $(\lambda, \Phi(Y))$ is dual-feasible for the instance of this problem corresponding to $\mathrm{D}^{\varepsilon}_{\mathrm{H}}(\Phi(\rho)\|\Phi(Q))$, with the same objective value being achieved by the assumption that $\Phi$ preserves trace. It follows that

$$2^{-\mathrm{D}^{\varepsilon}_{\mathrm{H}}(\Phi(\rho)\|\Phi(Q))} \geq 2^{-\mathrm{D}^{\varepsilon}_{\mathrm{H}}(\rho\|Q)}, \tag{5.53}$$

or, equivalently,

$$\mathrm{D}^{\varepsilon}_{\mathrm{H}}(\Phi(\rho)\|\Phi(Q)) \leq \mathrm{D}^{\varepsilon}_{\mathrm{H}}(\rho\|Q). \tag{5.54}$$

## Relationship to smoothed max-relative entropy

We will conclude our discussion of the hypothesis-testing relative entropy by observing its close relationship to the smoothed max-relative entropy. The two theorems that follow reveal this close relationship.

Recall that $\mathrm{D}^{\varepsilon}_{\max}(\rho\|Q)$ is the logarithm of the optimal value of this conic program:

**Optimization Problem 5.13** (SDP for smoothed max-relative entropy)

*Primal problem*

$$\begin{aligned} \text{minimize:}\quad & \eta \\ \text{subject to:}\quad & \xi \leq \eta Q \\ & \xi \in \mathcal{B}_{\varepsilon}(\rho), \\ & \eta \geq 0 \end{aligned}$$

*Dual problem*

$$\begin{aligned} \text{maximize:}\quad & \psi^{\varepsilon}_{\rho}(Z) \\ \text{subject to:}\quad & \langle Q, Z\rangle \leq 1 \\ & Z \in \mathrm{Pos}(\mathcal{X}) \end{aligned}$$

where

$$\psi^{\varepsilon}_{\rho}(Z) = \inf_{\xi \in \mathcal{B}_{\varepsilon}(\rho)} \langle \xi, Z\rangle. \tag{5.55}$$

In the theorems that follow, we will again use trace-distance smoothing, as we did in Lecture 4:

$$\mathcal{B}_{\varepsilon}(\rho) = \bigl\{\xi \in \mathrm{D}(\mathcal{X}) \,:\, \tfrac{1}{2}\|\rho - \xi\|_1 \leq \varepsilon\bigr\}. \tag{5.56}$$

**Theorem 5.14.** *Let $\rho \in \mathrm{D}(\mathcal{X})$ and $Q \in \mathrm{Pos}(\mathcal{X})$ be operators satisfying* $\mathrm{im}(\rho) \subseteq \mathrm{im}(Q)$ *and let $\varepsilon \in (0,1)$. It is the case that*

$$\mathrm{D}^{\sqrt{\varepsilon}}_{\max}(\rho\|Q) \leq \mathrm{D}^{\varepsilon}_{H}(\rho\|Q) + \log\Bigl(\frac{1}{\varepsilon(1-\varepsilon)}\Bigr). \tag{5.57}$$

*Proof.* Consider any operator $Z \in \mathrm{Pos}(\mathcal{X})$ that satisfies $\langle Q, Z\rangle \leq 1$, and let

$$Z = \sum_{k=1}^{n} \lambda_k(Z)\, z_k z_k^{*} \tag{5.58}$$

be a spectral decomposition of $Z$. We will use the basis $\{z_1,\ldots,z_n\}$ to construct a family of feasible solutions $X$ to the primal form of Optimization Problem 5.12.

In particular, for each real number $\lambda \in \mathbb{R}$, define a subset $S_\lambda \subseteq \{1,\ldots,n\}$ as

$$S_\lambda = \bigl\{k \in \{1,\ldots,n\} \,:\, \bigl\langle z_k z_k^*, \rho - 2^\lambda Q \bigr\rangle > 0 \bigr\}, \tag{5.59}$$

and define

$$\Pi_\lambda = \sum_{k \in S_\lambda} z_k z_k^*. \tag{5.60}$$

Observe that

$$\bigl\langle \Pi_\lambda, \rho \bigr\rangle \geq 2^\lambda \bigl\langle \Pi_\lambda, Q \bigr\rangle, \tag{5.61}$$

with the equality being strict so long as $\Pi_\lambda$ is nonzero. The operator

$$X = \frac{\Pi_\lambda}{\varepsilon} \tag{5.62}$$

is therefore feasible for primal form of Optimization Problem 5.12 provided that $\langle \Pi_\lambda, \rho \rangle \geq \varepsilon$, and with this observation in mind it is informative to consider the supremum over all such values of $\lambda$:

$$\gamma = \sup\bigl\{\lambda \in \mathbb{R} \,:\, \langle \Pi_\lambda, \rho \rangle \geq \varepsilon \bigr\}. \tag{5.63}$$

In particular, for every $\delta > 0$ we see that the operator

$$X = \frac{\Pi_{\gamma-\delta}}{\varepsilon} \tag{5.64}$$

is primal feasible, and so we obtain the inequalities

$$2^{-\operatorname{D}_{\mathrm{H}}^{\varepsilon}(\rho\|Q)} \leq \frac{\langle Q, \Pi_{\gamma-\delta} \rangle}{\varepsilon} \leq \frac{2^{-\gamma+\delta}}{\varepsilon} \langle \rho, \Pi_{\gamma-\delta} \rangle \leq \frac{2^{-\gamma+\delta}}{\varepsilon}. \tag{5.65}$$

As these inequalities hold for every $\delta > 0$, it follows that

$$\operatorname{D}_{\mathrm{H}}^{\varepsilon}(\rho\|Q) \geq \gamma + \log(\varepsilon). \tag{5.66}$$

On the other hand, for any choice of $\delta > 0$, it must be that $\langle \Pi_{\gamma+\delta}, \rho \rangle < \varepsilon$. The density operator

$$\xi = \frac{(\mathbb{1} - \Pi_{\gamma+\delta})\rho(\mathbb{1} - \Pi_{\gamma+\delta})}{\langle \mathbb{1} - \Pi_{\gamma+\delta}, \rho \rangle} \tag{5.67}$$

therefore satisfies

$$\operatorname{F}(\rho, \xi) \geq \sqrt{1 - \varepsilon} \tag{5.68}$$

by Winter's gentle measurement lemma, which implies that

$$\frac{1}{2}\|\rho-\xi\|_1 \leq \sqrt{\varepsilon} \tag{5.69}$$

by one of the Fuchs–van de Graaf inequalities. Consequently,

$$\begin{aligned}\psi_\rho^{\sqrt{\varepsilon}}(Z) \leq \langle \xi, Z\rangle &= \frac{1}{\langle \mathbb{1}-\Pi_{\gamma+\delta},\rho\rangle}\sum_{k\notin S_{\gamma+\delta}} \lambda_k(Z)\langle z_k z_k^*,\rho\rangle \\ &\leq \frac{2^{\gamma+\delta}}{1-\varepsilon}\sum_{k\notin S_{\gamma+\delta}} \lambda_k(Z)\langle z_k z_k^*, Q\rangle \leq \frac{2^{\gamma+\delta}}{1-\varepsilon}\langle Z, Q\rangle \leq \frac{2^{\gamma+\delta}}{1-\varepsilon}.\end{aligned} \tag{5.70}$$

As this is so for all $\delta > 0$, it follows that

$$\log \psi_\rho^{\sqrt{\varepsilon}}(Z) \leq \gamma - \log(1-\varepsilon), \tag{5.71}$$

and therefore

$$\log \psi_\rho^{\sqrt{\varepsilon}}(Z) \leq \mathrm{D}_{\mathrm{H}}^{\varepsilon}(\rho\|Q) + \log\Bigl(\frac{1}{\varepsilon(1-\varepsilon)}\Bigr). \tag{5.72}$$

By optimizing over all operators $Z \in \mathrm{Pos}(\mathcal{X})$ that satisfy $\langle Q, Z\rangle \leq 1$ we obtain the required inequality. □

**Theorem 5.15.** *Let $\rho \in \mathrm{D}(\mathcal{X})$ and $Q \in \mathrm{Pos}(\mathcal{X})$ be operators satisfying $\mathrm{im}(\rho) \subseteq \mathrm{im}(Q)$, let $\varepsilon \in (0,1)$, and let $\delta \in (0, 1-\varepsilon)$. It is the case that*

$$\mathrm{D}_{H}^{\varepsilon+\delta}(\rho\|Q) \leq \mathrm{D}_{\max}^{\varepsilon}(\rho\|Q) - \log\Bigl(\frac{\delta}{\varepsilon+\delta}\Bigr). \tag{5.73}$$

*Proof.* Suppose that $\xi \in \mathcal{B}_\varepsilon(\rho)$ satisfies

$$\xi \leq 2^{\mathrm{D}_{\max}^{\varepsilon}(\rho\|Q)} Q. \tag{5.74}$$

Given that $\xi \in \mathcal{B}_\varepsilon(\rho)$ we have that

$$\rho \leq \xi + R \tag{5.75}$$

for $R \in \mathrm{Pos}(\mathcal{X})$ satisfying $\mathrm{Tr}(R) \leq \varepsilon$.

Now consider the quantity $\mathrm{D}_{\mathrm{H}}^{\varepsilon+\delta}(\rho\|Q)$, and in particular consider the choices

$$\lambda = 2^{-\mathrm{D}_{\max}^{\varepsilon}(\rho\|Q)} \quad \text{and} \quad Y = \frac{2^{-\mathrm{D}_{\max}^{\varepsilon}(\rho\|Q)}R}{\varepsilon+\delta} \tag{5.76}$$

in the dual form of Optimization Problem 5.12 (with $\varepsilon$ being replaced by $\varepsilon+\delta$). These choices represent a feasible solution to this conic program, and the objective value is at least

$$2^{-\mathrm{D}_{\max}^{\varepsilon}(\rho\|Q)}\left(1 - \frac{\varepsilon}{\varepsilon+\delta}\right) \tag{5.77}$$

It follows that

$$\mathrm{D}_{\mathrm{H}}^{\varepsilon+\delta}(\rho\,\|\,Q) \leq \mathrm{D}_{\max}^{\varepsilon}(\rho\,\|\,Q) - \log\Big(\frac{\delta}{\varepsilon+\delta}\Big), \tag{5.78}$$

which completes the proof. □

**Corollary 5.16.** *Let $\rho \in \mathrm{D}(\mathcal{X})$ and $Q \in \mathrm{Pos}(\mathcal{X})$ be operators. For every $\varepsilon \in (0,1)$, we have*

$$\lim_{n\to\infty} \frac{\mathrm{D}_{\mathrm{H}}^{\varepsilon}\big(\rho^{\otimes n}\,\|\,Q^{\otimes n}\big)}{n} = \mathrm{D}(\rho\,\|\,Q). \tag{5.79}$$

Lecture 6

# Nonlocal games and Tsirelson's theorem

In this lecture we will discuss *nonlocal games*, which offer a model through which the phenomenon of nonlocality is commonly studied. We will then narrow our focus to *XOR games*, which are a highly restricted form of nonlocal games that can, perhaps surprisingly, be analyzed through semidefinite programming. This is made possible by *Tsirelson's theorem*, which we will prove in this lecture.

## 6.1 Nonlocal games

We will begin by introducing the *nonlocal games* model. A nonlocal game is a hypothetical game in which two cooperating players, *Alice* and *Bob*, each receive a question from a *referee*, and then respond with an answer. The referee randomly selects the questions according to a known distribution, and, upon receiving answers from Alice and Bob, decides whether they win or lose. The following definition makes this notion precise in mathematical terms.

**Definition 6.1.** A *nonlocal game* is a 6-tuple $G = (X, Y, A, B, \pi, V)$, where

1. $X$, $Y$, $A$, and $B$ are finite and nonempty sets,
2. $\pi \in \mathcal{P}(\mathcal{X} \times Y)$ is a probability vector, and
3. $V : A \times B \times X \times Y \rightarrow \{0, 1\}$ is a predicate.

In this definition, the sets $X$ and $Y$ are the sets of questions, and $A$ and $B$ are the sets of answers, for Alice and Bob, respectively. The probability vector $\pi$ determines the probability with which each pair of questions $(x, y) \in X \times Y$ is selected by the referee, and $V$ determines whether or not a pair of answers $(a, b)$ wins or loses for a given pair of questions $(x, y)$. For a given pair of questions

$(x, y) \in X \times Y$ and a pair of answers $(a, b) \in A \times B$, we write the value of the predicate as $V(a, b|x, y)$, because that's the way Ben Toner prefers it to be written—as it helps to stress the idea that $(a, b)$ either wins or loses given that the question pair $(x, y)$ was selected.

**Example 6.2** (The CHSH game)**.** The CHSH game (named after Clauser, Horn, Shimony, and Holt) is a nonlocal game in which the questions and answers correspond to binary values, $X = Y = A = B = \{0, 1\}$, the probability vector $\pi$ is uniform,

$$\pi(0,0) = \pi(0,1) = \pi(1,0) = \pi(1,1) = \frac{1}{4}, \tag{6.1}$$

and the predicate $V$ is defined as

$$V(a, b|x, y) = \begin{cases} 1 & \text{if } a \oplus b = x \wedge y \\ 0 & \text{if } a \oplus b \neq x \wedge y, \end{cases} \tag{6.2}$$

where $a \oplus b$ denotes the XOR of $a$ and $b$, and $x \wedge y$ denotes the AND of $x$ and $y$.

Intuitively speaking, if the referee selects any of the question pairs $(0, 0)$, $(0, 1)$, or $(1, 0)$, then Alice and Bob must provide a pair of answers $(a, b)$ for which $a = b$ in order to win, while if the referee selects the question pair $(1, 1)$, the answer $(a, b)$ wins when $a \neq b$.

**Example 6.3** (The FFL game)**.** The FFL game (named after Fortnow, Feige, and Lovász) is a nonlocal game in which the questions and answers correspond to binary values, $X = Y = A = B = \{0, 1\}$, the probability vector $\pi$ is given by

$$\pi(0,0) = \pi(0,1) = \pi(1,0) = \frac{1}{3}, \quad \pi(1,1) = 0, \tag{6.3}$$

and the predicate $V$ is defined as

$$V(a, b|x, y) = \begin{cases} 1 & \text{if } a \vee x \neq b \vee y \\ 0 & \text{if } a \vee x = b \vee y, \end{cases} \tag{6.4}$$

where $a \vee x$ denotes the OR of $a$ and $x$, and similar for $b \vee y$.

Intuitively speaking, if the referee asks the question pair $(0, 0)$, then exactly one of Alice and Bob, but not both, must respond with the answer 1 in order to win. However, if the question pair is either $(0, 1)$ or $(1, 0)$, then the player who received 0 must answer 0 to win (and it does not matter what the player who received the question 1 answers).

**Example 6.4** (Graph coloring games)**.** Suppose that $H = (V, E)$ is an undirected graph and $k$ is a positive integer. Let us also define $n = |V|$ and $m = |E|$, and assume $m \geq 1$. We may form a nonlocal game in the following way. The question sets are both equal to the set of vertices, $X = Y = \{1, \ldots, n\}$, and the answer sets are given by $A = B = \{1, \ldots, k\}$, which we may intuitively think about as colors. The probability vector $\pi$ is defined as follows:

$$\pi(x, y) = \begin{cases} \frac{1}{2n} & \text{if } x = y \\ \frac{1}{4m} & \text{if } \{x, y\} \in E \\ 0 & \text{otherwise.} \end{cases} \tag{6.5}$$

In words, the referee flips a fair coin, and if the outcome is heads, it randomly selects a vertex and sends it to both players, and if the outcome is tails, it randomly selects an edge and then sends the two incident vertices to the two players (again at random). The predicate is defined as

$$V(a, b|x, y) = \begin{cases} 1 & \text{if } x = y \text{ and } a = b \\ 1 & \text{if } x \neq y \text{ and } a \neq b \\ 0 & \text{otherwise.} \end{cases} \tag{6.6}$$

The idea is that if Alice and Bob receive the same vertex, they should answer with the same color, while if they receive different (adjacent) vertices, they should answer with different colors.

## Strategies

The definition of a nonlocal game does not, in itself, specify or restrict the sorts of strategies that Alice and Bob might employ when playing. There are, in fact, different types of strategies that are of interest. Let us start with a short summary of the strategy types that are of interest for this lecture.

1. *Deterministic strategies.* In a deterministic strategy, Alice must deterministically choose her answer $a$ based on her question $x$ alone, and likewise Bob must choose $b$ based on $y$ alone. A deterministic strategy may therefore be described as a pair of functions $(f, g)$, where $f : X \to A$ and $g : Y \to B$.

   Notice that when we consider such a strategy, there is an implicit assumption that Alice cannot see Bob's question (or answer), and likewise Bob cannot see Alice's question (or answer). This sort of implicit assumption is also in place for the other strategy types listed below, and is what makes nonlocal games interesting and motivates their name.

2. *Randomized strategies.* Rather than choosing their answers deterministically, Alice and Bob could choose to make use of randomness when selecting their answers. The randomness could be in the form of local randomness, where Alice and Bob individually generate random numbers to assist in the selection of their answers, or it could be in the form of shared randomness, which one might view as having been generated by Alice and Bob at some point in the past.

   As it turns out, randomized strategies are not helpful to Alice and Bob, assuming their goal is to maximize the probability that they win. This is because randomized strategies can simply be viewed as the random selection of a deterministic strategy, and Alice and Bob might as well just select the optimal deterministic strategy—the average winning probability obviously cannot be larger than the maximum winning probability over all deterministic strategies.

3. *Entangled strategies.* An entangled strategy is one in which Alice and Bob make use of a shared quantum state when playing a nonlocal game. That is, Alice holds a register $\mathsf{A}$ and Bob holds a register $\mathsf{B}$, where $(\mathsf{A}, \mathsf{B})$ is in a joint state $\rho \in \mathrm{D}(\mathcal{A} \otimes \mathcal{B})$, prior to the referee sending the questions. Upon receiving a question $x \in X$, Alice measures the register $\mathsf{A}$ with respect to a measurement described by a collection of measurement operators

   $$\left\{P_a^x \,:\, a \in A\right\} \subset \mathrm{Pos}(\mathcal{A}), \tag{6.7}$$

   and likewise Bob measures $\mathsf{B}$ with respect to a measurement described by measurement operators

   $$\left\{Q_b^y \,:\, b \in B\right\} \subset \mathrm{Pos}(\mathcal{B}). \tag{6.8}$$

   To be clear, Alice's measurement depends on her question $x \in X$ and Bob's measurement depends on his question $y \in Y$; they each have a measurement for each possible question they might receive.

   Given such a strategy, we see that the probability that Alice and Bob respond to a question pair $(x, y)$ with an answer pair $(a, b)$ is equal to

   $$\left\langle P_a^x \otimes Q_b^y, \rho \right\rangle. \tag{6.9}$$

   Note that $\rho$ is not actually required to be entangled by the definition of an entangled strategy, but also note that if $\rho$ is separable, then the strategy will be equivalent to a classical randomized strategy. So, entanglement is what makes this sort of strategy different from a classical strategy, which perhaps explains the name *entangled strategy*.

There are other types of strategies that are often considered in the study of nonlocal games, including *commuting operator strategies* and *no-signaling strategies*—we will discuss commuting operator strategies in the lecture following the next one. One could also consider *global strategies*, in which there is no implicit assumption that Alice and Bob are separated, so that $(a, b)$ can depend arbitrarily on $(x, y)$, but this class of strategies is not very interesting in a setting in which the nonlocality of Alice and Bob is relevant.

## Values of games

When we speak of the *value* of a nonlocal game, we're referring to the supremum probability with which Alice and Bob can win the game, with respect to whatever class of strategies we might wish to consider. For this lecture we will focus on two values: the *classical value* and the *entangled value*.

**Definition 6.5** (Classical value of a nonlocal game)**.** The *classical value* of a nonlocal game $G = (X, Y, A, B, \pi, V)$, which is denoted $\omega(G)$, is given by a maximization of the winning probability over all deterministic strategies:

$$\omega(G) = \max_{f,g} \sum_{(x,y)\in X\times Y} \pi(x,y) V\bigl(f(x), g(y) | x, y\bigr), \tag{6.10}$$

where the maximum is over all $f : X \rightarrow A$ and $g : Y \rightarrow B$.

**Remark 6.6.** As was already discussed, there is need to differentiate between deterministic and randomized values of nonlocal games, because they are the same—and so the name *classical value* is justified.

**Definition 6.7** (Entangled value of a nonlocal game)**.** The *entangled value* of a nonlocal game $G = (X, Y, A, B, \pi, V)$, which is denoted $\omega^*(G)$, is the supremum of the winning probabilities

$$\sum_{(x,y)\in X\times Y} \pi(x,y) \sum_{(a,b)\in A\times B} V\bigl(a, b | x, y\bigr) \bigl\langle P_a^x \otimes Q_b^y, \rho \bigr\rangle, \tag{6.11}$$

over all choices of complex Euclidean spaces $\mathcal{A}$ and $\mathcal{B}$, states $\rho \in \mathrm{D}(\mathcal{A} \otimes \mathcal{B}\}$, and sets of measurements

$$\bigl\{P_a^x \,:\, a \in A\bigr\}_{x\in X} \subset \mathrm{Pos}(\mathcal{A}) \quad \text{and} \quad \bigl\{Q_b^y \,:\, b \in B\bigr\}_{y\in Y} \subset \mathrm{Pos}(\mathcal{B}). \tag{6.12}$$

That is, the entangled value is the supremum winning probability over all entangled strategies.

**Remark 6.8.** There are nonlocal games for which the winning probability is never achieved, so it is necessary to use the supremum in this definition. The principal issue is that the dimensions of the spaces $\mathcal{A}$ and $\mathcal{B}$ are not bounded as one ranges over all entangled strategies.

**Example 6.9** (CHSH game values)**.** Letting $G$ denote the CHSH game, we have that the classical value of this game is $\omega(G) = 3/4$. This may be verified by checking that the winning probability of each of the 16 possible deterministic strategies is at most $3/4$, and of course that some of those strategies win with probability $3/4$.

The entangled value of the CHSH game is $\omega^*(G) = \cos^2(\pi/8) \approx 0.85$. The fact that this is so will emerge as a simple corollary to Tsirelson's theorem—which is fitting given that the inequality $\omega^*(G) \leq \cos^2(\pi/8)$ is a rephrasing of an inequality known as Tsirelson's bound.

**Example 6.10** (FFL game values)**.** If we let $G$ denote the FFL game, then we have that its classical value and quantum value agree: $\omega(G) = \omega^*(G) = 2/3$. The fact that $\omega(G) = 2/3$ is easily established by testing all deterministic classical strategies. I will ask you to prove that $\omega^*(G) = 2/3$ as a homework problem. One way to do this is to prove that even the so-called *no-signaling value*, which upper-bounds the quantum value, of the FFL game is 2/3. The no-signaling value can be computed through linear programming.

**Example 6.11** (Graph coloring game values)**.** If $G$ is the graph coloring game determined by a graph $H$ and an integer $k$, there we see that $\omega(G) = 1$ if and only if the chromatic number of $H$ is at most $k$. That is, given any *perfect deterministic strategy*, meaning one that wins with certainty, it is possible to recover a $k$-coloring of $H$, meaning an assignment of colors $\{1, \ldots, k\}$ to the vertices of $H$ such that no two adjacent vertices share the same color.

There are known examples of graphs $H$ and choices of $k$ for which the associated nonlocal game $G$ satisfies $\omega(G) < 1$ but $\omega^*(G) = 1$.

## 6.2 XOR games

XOR games are a restricted type of nonlocal game $G = (X, Y, A, B, \pi, V)$ in which both players answer binary values, so that $A = B = \{0, 1\}$, and for which the predicate $V$ takes the form

$$V(a, b|x, y) = \begin{cases} 1 & \text{if } a \oplus b = f(x, y) \\ 0 & \text{if } a \oplus b \neq f(x, y) \end{cases} \tag{6.13}$$

for some choice of a function $f : X \times Y \rightarrow \{0,1\}$. Intuitively speaking, the function $f$ specifies whether $a$ and $b$ should agree or disagree in order to be a winning answer, for each question pair $(x, y)$. Notice that exactly one of the two possibilities, meaning the possibilities that $a$ and $b$ agree or disagree, always wins for each question pair, while the other possibility loses.

As every XOR game is uniquely determined by the sets $X$ and $Y$, the probability vector $\pi \in \mathcal{P}(X \times Y)$, and the function $f : X \times Y \rightarrow \{0,1\}$, we will identify the corresponding game $G$ with the quadruple $(X, Y, \pi, f)$ when it is convenient to do that. For example, the CHSH game is an example of an XOR game, corresponding to the the quadruple $(\{0,1\}, \{0,1\}, \pi, f)$, for $\pi$ the uniform probability vector and $f(x, y) = x \wedge y$ being the AND function.

## Bias of an XOR game

When analyzing XOR games, it is often convenient to consider the *bias* of games rather than their value. For a given XOR game $G = (X, Y, \pi, f)$, and any strategy for $G$, we define the bias of that strategy, for that game, to be the probability it wins *minus* the probability it loses—which happens to be the same thing as twice the probability it wins minus 1. The *bias of a game* is defined to be the supremum bias over all strategies under consideration for that game. We will write $\varepsilon(G)$ and $\varepsilon^*(G)$ to denote the classical and quantum biases for $G$, and so we have

$$\varepsilon(G) = 2\omega(G) - 1 \quad \text{and} \quad \varepsilon^*(G) = 2\omega^*(G) - 1, \tag{6.14}$$

or, alternatively,

$$\omega(G) = \frac{1}{2} + \frac{\varepsilon(G)}{2} \quad \text{and} \quad \omega^*(G) = \frac{1}{2} + \frac{\varepsilon^*(G)}{2}. \tag{6.15}$$

## XOR game strategies described by observables

Let $G = (X, Y, \pi, f)$ be an XOR game, and consider any entangled strategy for that game, represented by a state $\rho \in \mathrm{D}(\mathcal{A} \otimes \mathcal{B}\}$ and measurement operators

$$\{P_0^x, P_1^x\}_{x \in X} \subset \mathrm{Pos}(\mathcal{A}) \quad \text{and} \quad \{Q_0^y, Q_1^y\}_{y \in Y} \subset \mathrm{Pos}(\mathcal{B}). \tag{6.16}$$

If we consider the expression

$$\sum_{x,y \in X \times Y} \pi(x, y)(-1)^{f(x,y)} \big\langle \big(P_0^x - P_1^x\big) \otimes \big(Q_0^y - Q_1^y\big), \rho \big\rangle \tag{6.17}$$

for a few moments, we find that it agrees with the bias of the strategy just described. By defining $A_x = P_0^x - P_1^x$ for each $x \in X$ and $B_y = Q_0^y - Q_1^y$ for each

$y \in Y$, we may express this quantity as

$$\sum_{x,y \in X \times Y} \pi(x,y)(-1)^{f(x,y)} \langle A_x \otimes B_y, \rho \rangle. \tag{6.18}$$

The operators $A_x$ and $B_y$ may be viewed as representing *observables* in the parlance of quantum mechanics.

Notice that as one ranges over all binary-valued measurements $\{R_0, R_1\}$, the operator $R_0 - R_1$ ranges over all Hermitian operators $H$ with $\|H\| \leq 1$. Therefore, the bias of a game $G$ is given by the supremum value of the expression (6.18), over all choices of $\{A_x : x \in X\} \subset \mathrm{Herm}(\mathcal{A})$, $\{B_y : y \in Y\} \subset \mathrm{Herm}(\mathcal{B})$, and $\rho \in \mathrm{D}(\mathcal{A} \otimes \mathcal{B})$, subject to the constraints $\|A_x\| \leq 1$ for every $x \in X$ and $\|B_y\| \leq 1$ for every $y \in Y$.

## 6.3 Tsirelson's theorem

Now we will prove the theorem of Tsirelson mentioned previously. Let us begin with a statement of the theorem.

**Theorem 6.12** (Tsirelson's theorem). *For every choice of finite and nonempty sets $X$ and $Y$ and an operator $M \in \mathrm{L}(\mathbb{R}^Y, \mathbb{R}^X)$, the following statements are equivalent.*

1. *There exist complex Euclidean spaces $\mathcal{A}$ and $\mathcal{B}$, a density operator $\rho \in \mathrm{D}(\mathcal{A} \otimes \mathcal{B})$, and two collections $\{A_x : x \in X\} \subset \mathrm{Herm}(\mathcal{A})$ and $\{B_y : y \in Y\} \subset \mathrm{Herm}(\mathcal{B})$ of operators such that $\|A_x\| \leq 1$, $\|B_y\| \leq 1$, and*

$$M(x,y) = \langle A_x \otimes B_y, \rho \rangle \tag{6.19}$$

   *for all $x \in X$ and $y \in Y$.*

2. *There exist positive semidefinite operators $R \in \mathrm{Pos}(\mathbb{C}^X)$ and $S \in \mathrm{Pos}(\mathbb{C}^Y)$, with $R(x,x) = 1$ and $S(y,y) = 1$ for all $x \in X$ and $y \in Y$, such that*

$$\begin{pmatrix} R & M \\ M^* & S \end{pmatrix} \geq 0. \tag{6.20}$$

**Remark 6.13.** The second statement in the theorem is equivalent to one in which the requirement that $R$ and $S$ have real number entries is added. In particular, if $R_0$ and $S_0$ satisfy the conditions listed in the second statement of the theorem, then so too will

$$R = \frac{R_0 + R_0^{\mathsf{T}}}{2} \quad \text{and} \quad S = \frac{S_0 + S_0^{\mathsf{T}}}{2}, \tag{6.21}$$

by virtue of the fact that $M$ has real-number entries and

$$\begin{pmatrix} R & M \\ M^* & S \end{pmatrix} = \frac{1}{2}\begin{pmatrix} R_0 & M \\ M^* & S_0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} R_0 & M \\ M^* & S_0 \end{pmatrix}^{\mathsf{T}} \tag{6.22}$$

is a positive semidefinite operator whose diagonal entries are all equal to 1.

The first statement of the theorem says that the operator $M$, which is best viewed as a matrix indexed by pairs $(x, y) \in X \times Y$ in this case, describes exactly the values in the expression (6.18) that depend upon the strategy under consideration. The second statement of the theorem is a suprisingly simple condition on $M$—and it may come at no surprise to learn that it will be used to define semidefinite programs to calculate XOR game biases. The fact that these two statements are exactly the same thing is a remarkable thing of beauty.

## Weyl–Brauer operators

The proof of Tsirelson's theorem will make use of a collection of unitary and Hermitian operators known as *Weyl–Brauer operators*.

**Definition 6.14.** Let $N$ be a positive integer and let $\mathcal{Z} = \mathbb{C}^2$. The *Weyl–Brauer operators of order* $N$ are the operators $V_1, \ldots, V_{2N+1} \in \mathrm{L}(\mathcal{Z}^{\otimes N})$ defined as

$$\begin{aligned} V_{2k-1} &= \sigma_z^{\otimes(k-1)} \otimes \sigma_x \otimes \mathbb{1}^{\otimes(N-k)}, \\ V_{2k} &= \sigma_z^{\otimes(k-1)} \otimes \sigma_y \otimes \mathbb{1}^{\otimes(N-k)}, \end{aligned} \tag{6.23}$$

for all $k \in \{1, \ldots, N\}$, as well as

$$V_{2N+1} = \sigma_z^{\otimes N}, \tag{6.24}$$

where $\mathbb{1}$, $\sigma_x$, $\sigma_y$, and $\sigma_z$ denote the Pauli operators:

$$\mathbb{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{6.25}$$

**Example 6.15.** In the case $N = 3$, the Weyl–Brauer operators $V_1, \ldots, V_7$ are

$$\begin{aligned} V_1 &= \sigma_x \otimes \mathbb{1} \otimes \mathbb{1} \\ V_2 &= \sigma_y \otimes \mathbb{1} \otimes \mathbb{1} \\ V_3 &= \sigma_z \otimes \sigma_x \otimes \mathbb{1} \\ V_4 &= \sigma_z \otimes \sigma_y \otimes \mathbb{1} \\ V_5 &= \sigma_z \otimes \sigma_z \otimes \sigma_x \\ V_6 &= \sigma_z \otimes \sigma_z \otimes \sigma_y \\ V_7 &= \sigma_z \otimes \sigma_z \otimes \sigma_z. \end{aligned} \tag{6.26}$$

A proposition summarizing the properties of the Weyl–Brauer operators that are relevant to the proof of Tsirelson's theorem follows.

**Proposition 6.16.** *Let $N$ be a positive integer, let $V_1, \ldots, V_{2N+1}$ denote the Weyl–Brauer operators of order $N$. For every unit vector $u \in \mathbb{R}^{2N+1}$, the operator*

$$\sum_{k=1}^{2N+1} u(k) V_k \tag{6.27}$$

*is both unitary and Hermitian, and for any two vectors $u, v \in \mathbb{R}^{2N+1}$, it holds that*

$$\frac{1}{2^N} \left\langle \sum_{j=1}^{2N+1} u(j) V_j, \sum_{k=1}^{2N+1} v(k) V_k \right\rangle = \langle u, v \rangle. \tag{6.28}$$

*Proof.* Each operator $V_k$ is Hermitian, and therefore the operator (6.27) is Hermitian as well.

The Pauli operators anti-commute in pairs:

$$\sigma_x \sigma_y = -\sigma_y \sigma_x, \quad \sigma_x \sigma_z = -\sigma_z \sigma_x, \quad \text{and} \quad \sigma_y \sigma_z = -\sigma_z \sigma_y. \tag{6.29}$$

By an inspection of the definition of the Weyl–Brauer operators, it follows that $V_1, \ldots, V_{2N+1}$ also anti-commute in pairs:

$$V_j V_k = -V_k V_j \tag{6.30}$$

for distinct choices of $j, k \in \{1, \ldots, 2N+1\}$. Moreover, each $V_k$ is unitary (as well as being Hermitian), and therefore $V_k^2 = \mathbb{1}^{\otimes N}$. It follows that

$$\begin{aligned}\left( \sum_{k=1}^{2N+1} u(k) V_k \right)^2 &= \sum_{k=1}^{2N+1} u(k)^2 V_k^2 + \sum_{1 \leq j < k \leq 2N+1} u(j) u(k) \bigl( V_j V_k + V_k V_j \bigr) \\ &= \sum_{k=1}^{2N+1} u(k)^2 \mathbb{1}^{\otimes N} = \mathbb{1}^{\otimes N},\end{aligned} \tag{6.31}$$

and therefore (6.27) is unitary.

Next, observe that

$$\langle V_j, V_k \rangle = \begin{cases} 2^N & \text{if } j = k \\ 0 & \text{if } j \neq k. \end{cases} \tag{6.32}$$

Therefore, one has

$$\begin{aligned}&\frac{1}{2^N} \left\langle \sum_{j=1}^{2N+1} u(j) V_j, \sum_{k=1}^{2N+1} v(k) V_k \right\rangle \\ &\quad = \frac{1}{2^N} \sum_{j=1}^{2N+1} \sum_{k=1}^{2N+1} u(j) v(k) \langle V_j, V_k \rangle = \sum_{k=1}^{2N+1} u(k) v(k) = \langle u, v \rangle,\end{aligned} \tag{6.33}$$

as required. □

## Proof of Tsirelson's theorem

*Proof of Theorem 6.12.* For the sake of simplifying notation, we will make the assumption that $X = \{1, \ldots, n\}$ and $Y = \{1, \ldots, m\}$.

Assume that the first statement is true, and define an operator

$$K = \begin{pmatrix} \operatorname{vec}\bigl((A_1 \otimes \mathbb{1})\sqrt{\rho}\bigr)^* \\ \vdots \\ \operatorname{vec}\bigl((A_n \otimes \mathbb{1})\sqrt{\rho}\bigr)^* \\ \operatorname{vec}\bigl((\mathbb{1} \otimes B_1)\sqrt{\rho}\bigr)^* \\ \vdots \\ \operatorname{vec}\bigl((\mathbb{1} \otimes B_m)\sqrt{\rho}\bigr)^* \end{pmatrix} \in \mathrm{L}(\mathcal{A} \otimes \mathcal{B} \otimes \mathcal{A} \otimes \mathcal{B}, \mathbb{C}^n \oplus \mathbb{C}^m). \tag{6.34}$$

The operator $KK^* \in \operatorname{Pos}\bigl(\mathbb{C}^n \oplus \mathbb{C}^m\bigr)$ may be written in a block form as

$$KK^* = \begin{pmatrix} P & M \\ M^* & Q \end{pmatrix} \tag{6.35}$$

for $P \in \operatorname{Pos}(\mathbb{C}^n)$ and $Q \in \operatorname{Pos}(\mathbb{C}^m)$; the fact that the off-diagonal blocks are as claimed follows from the calculation

$$\bigl\langle (A_j \otimes \mathbb{1})\sqrt{\rho}, (\mathbb{1} \otimes B_k)\sqrt{\rho} \bigr\rangle = \bigl\langle A_j \otimes B_k, \rho \bigr\rangle = M(j,k). \tag{6.36}$$

For each $j \in \{1, \ldots, n\}$ one has

$$P(j,j) = \bigl\langle (A_j \otimes \mathbb{1})\sqrt{\rho}, (A_j \otimes \mathbb{1})\sqrt{\rho} \bigr\rangle = \bigl\langle A_j^2 \otimes \mathbb{1}, \rho \bigr\rangle, \tag{6.37}$$

which is necessarily a nonnegative real number in the interval $[0,1]$; and through a similar calculation, one finds that $Q(k,k)$ is also a nonnegative integer in the interval $[0,1]$ for each $k \in \{1, \ldots, m\}$. A nonnegative real number may be added to each diagonal entry of this operator to yield another positive semidefinite operator, so one has that statement 2 holds.

Next, let us assume statement 2 holds. As was explained in Remark 6.13, we are free to assume that all of the entries of $R$ and $S$ are real numbers.

Now, a matrix with real number entries is positive semidefinite if and only if it is the Gram matrix of a collection of real vectors, and therefore there must exist real vectors $\{u_1, \ldots, u_n, v_1, \ldots, v_m\}$ such that

$$\langle u_j, v_k \rangle = M(j,k) \tag{6.38}$$

for all $j \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, m\}$, as well as

$$\langle u_{j_0}, u_{j_1} \rangle = R(j_0, j_1) \quad \text{and} \quad \langle v_{k_0}, v_{k_1} \rangle = S(k_0, k_1) \tag{6.39}$$

for all $j_0, j_1 \in \{1, \ldots, n\}$ and $k_0, k_1 \in \{1, \ldots, m\}$. There are $n + m$ of these vectors, and therefore they span a real vector space of dimension at most $n + m$, so there is no loss of generality in assuming $u_1, \ldots, u_n, v_1, \ldots, v_m \in \mathbb{R}^{n+m}$. Observe that these vectors are all unit vectors, as the diagonal entries of $R$ and $S$ represent their norm squared.

Choose $N$ so that $2N + 1 \geq n + m$ and let $\mathcal{Z} = \mathbb{C}^2$. Define operators $A_1, \ldots, A_n$ and $B_1, \ldots, B_m$, all acting on $\mathrm{L}(\mathcal{Z}^{\otimes N})$, as

$$A_j = \sum_{i=1}^{n+m} u_j(i) V_i \quad \text{and} \quad B_k = \sum_{i=1}^{n+m} v_k(i) V_i^{\mathsf{T}} \tag{6.40}$$

for each $j \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, m\}$, where $V_1, \ldots, V_{n+m}$ are the first $n + m$ Weyl–Brauer operators of order $N$. By Proposition 6.16, each of these operators is both unitary and Hermitian, and therefore each of these operators has spectral norm equal to 1.

Finally, define

$$\rho = \frac{1}{2^N} \operatorname{vec}\bigl(\mathbb{1}^{\otimes N}\bigr) \operatorname{vec}\bigl(\mathbb{1}^{\otimes N}\bigr)^* \in \mathrm{D}\bigl(\mathcal{Z}^{\otimes N} \otimes \mathcal{Z}^{\otimes N}\bigr). \tag{6.41}$$

Applying Proposition 6.16 again gives

$$\bigl\langle A_j \otimes B_k, \rho \bigr\rangle = \frac{1}{2^N} \bigl\langle A_j, B_k^{\mathsf{T}} \bigr\rangle = \langle u_j, v_k \rangle = M(j,k), \tag{6.42}$$

for each $j \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, m\}$.

We have proved that statement 2 implies statement 1, for the spaces $\mathcal{A} = \mathcal{Z}^{\otimes N}$ and $\mathcal{B} = \mathcal{Z}^{\otimes N}$, and so the proof is complete. □

Lecture 7

# A semidefinite program for the entangled bias of XOR games

In this lecture we will discuss a couple of applications of Tsirelson's theorem, centering primarily on the semidefinite programming formulation for the entangled bias of XOR games that it yields.

## 7.1 The semidefinite program

Let us begin with the semidefinite program that is suggested by Tsirelson's theorem. Assume that $G = (X, Y, \pi, f)$ is an XOR game, and recall (as was discussed in the previous lecture) that the entangled bias of $G$ is the supremum of the values

$$\sum_{(x,y)\in X\times Y} \pi(x,y)(-1)^{f(x,y)} \langle A_x \otimes B_y, \rho\rangle, \tag{7.1}$$

taken over all choices for complex Euclidean spaces $\mathcal{A}$ and $\mathcal{B}$, a state $\rho \in \mathrm{D}(\mathcal{A}\otimes\mathcal{B})$, and Hermitian contractions

$$\{A_x \,:\, x \in X\} \subset \mathrm{Herm}(\mathcal{A}) \quad \text{and} \quad \{B_y \,:\, y \in Y\} \subset \mathrm{Herm}(\mathcal{B}). \tag{7.2}$$

By Tsirelson's theorem, this is equivalent to the supremum of the values

$$\sum_{(x,y)\in X\times Y} \pi(x,y)(-1)^{f(x,y)} M(x,y), \tag{7.3}$$

taken over all $M \in \mathrm{L}(\mathbb{R}^Y, \mathbb{R}^X)$ for which there exist $R \in \mathrm{Pos}(\mathbb{C}^X)$ and $S \in \mathrm{Pos}(\mathbb{C}^Y)$ for which $R(x,x) = 1$ for all $x \in X$, $S(y,y) = 1$ for all $y \in Y$, and

$$\begin{pmatrix} R & M \\ M^* & S \end{pmatrix} \in \mathrm{Pos}(\mathbb{C}^X \oplus \mathbb{C}^Y). \tag{7.4}$$

With this fact in mind, let us consider the following semidefinite program. First, let us write $\mathcal{X} = \mathbb{C}^X$ and $\mathcal{Y} = \mathbb{C}^Y$ for brevity, and let $\Delta \in \mathrm{C}(\mathcal{X} \oplus \mathcal{Y})$ denote the completely dephasing channel acting on $\mathcal{X} \oplus \mathcal{Y}$, which zeros out all of the off-diagonal entries of its input and leaves the diagonal entries alone. Define an operator $D \in \mathrm{L}(\mathcal{Y}, \mathcal{X})$ as

$$D(x,y) = \pi(x,y)(-1)^{f(x,y)} \tag{7.5}$$

for every $x \in X$ and $y \in Y$, and let $H \in \mathrm{Herm}(\mathcal{X} \oplus \mathcal{Y})$ be defined as

$$H = \frac{1}{2}\begin{pmatrix} 0 & D \\ D^* & 0 \end{pmatrix}. \tag{7.6}$$

The semidefinite program to be considered is described by the triple $(\Delta, H, \mathbb{1}_{\mathcal{X}\oplus\mathcal{Y}})$. The primal and dual problems associated with this semidefinite program take the following form.

**Optimization Problem 7.1** (SDP for XOR game bias, unsimplified)

| *Primal problem* | | *Dual problem* | |
|---|---|---|---|
| maximize: | $\langle H, Z\rangle$ | minimize: | $\mathrm{Tr}(W)$ |
| subject to: | $\Delta(Z) = \mathbb{1}_{\mathcal{X}\oplus\mathcal{Y}}$, | subject to: | $\Delta(W) \geq H$, |
| | $Z \in \mathrm{Pos}(\mathcal{X}\oplus\mathcal{Y})$. | | $W \in \mathrm{Herm}(\mathcal{X}\oplus\mathcal{Y})$. |

Note that in the dual problem formulation we have used the fact that the completely dephasing channel is self-dual: $\Delta = \Delta^*$. Strong duality and the achievability of the optimal values in both the primal and dual problems follow from Slater's theorem; strictly feasible solutions are given by $Z = \mathbb{1}_{\mathcal{X}\oplus\mathcal{Y}}$ in the primal and $W = \lambda\mathbb{1}_{\mathcal{X}\oplus\mathcal{Y}}$ for a sufficiently large $\lambda$ in the dual.

Let us now examine both the primal and dual problems, beginning with the primal problem. Our principal order of business with the primal problem, which is fairly straightforward, is to verify that its optimal value indeed agrees with the entangled bias of the XOR game $G$.

Suppose first that $M \in \mathrm{L}(\mathbb{R}^Y, \mathbb{R}^X)$ is such that there exist $R \in \mathrm{Pos}(\mathcal{X})$ and $S \in \mathrm{Pos}(\mathcal{Y})$ for which $R(x,x) = 1$ for all $x \in X$, $S(y,y) = 1$ for all $y \in Y$, and

$$Z = \begin{pmatrix} R & M \\ M^* & S \end{pmatrix} \in \mathrm{Pos}(\mathcal{X}\oplus\mathcal{Y}). \tag{7.7}$$

The operator $Z$ is then primal feasible, as the constraint $\Delta(Z) = \mathbb{1}_{\mathcal{X}\oplus\mathcal{Y}}$ is equivalent to $R$ and $S$ having diagonal entries equal to one. The objective value for $Z$ is equal to

$$\langle H, Z\rangle = \frac{1}{2}\langle D, M\rangle + \frac{1}{2}\langle D^*, M^*\rangle = \langle D, M\rangle, \tag{7.8}$$

owing to the fact that $D$ and $M$ have real number entries. Noting that

$$\langle D, M\rangle = \sum_{(x,y)\in X\times Y} \pi(x,y)(-1)^{f(x,y)} M(x,y), \tag{7.9}$$

we find that the optimal value of the semidefinite program is at least the entangled bias of $G$.

To see that the optimal value of the semidefinite program is no greater than the entangled bias of $G$, consider any $Z \in \operatorname{Pos}(\mathcal{X}\oplus\mathcal{Y})$, which can be expressed as

$$Z = \begin{pmatrix} R & K \\ K^* & S \end{pmatrix} \tag{7.10}$$

for some choice of $R \in \operatorname{Pos}(\mathcal{X})$, $S \in \operatorname{Pos}(\mathcal{Y})$, and $K \in \mathrm{L}(\mathcal{Y},\mathcal{X})$. Again the constraint $\Delta(Z) = \mathbb{1}_{\mathcal{X}\oplus\mathcal{Y}}$ is equivalent to $R$ and $S$ having diagonal entries equal to one, and the only issue remaining is that $K$ might not have real number entries. However, by expressing the objective function in terms of the block structure of $H$ and $Z$, we find that

$$\langle H, Z\rangle = \frac{1}{2}\langle D, K\rangle + \frac{1}{2}\langle D^*, K^*\rangle = \langle D, M\rangle \tag{7.11}$$

for

$$M = \frac{K+\overline{K}}{2}, \tag{7.12}$$

following from the fact that $D$ has real entries:

$$\langle D^*, K^*\rangle = \overline{\langle D, K\rangle} = \langle \overline{D}, \overline{K}\rangle = \langle D, \overline{K}\rangle. \tag{7.13}$$

Proceeding in much the same way as in the previous lecture, we see that

$$\frac{1}{2}\begin{pmatrix} R & K \\ K^* & S \end{pmatrix} + \frac{1}{2}\begin{pmatrix} R & K \\ K^* & S \end{pmatrix}^{\mathsf{T}} = \begin{pmatrix} \frac{1}{2}R + \frac{1}{2}R^{\mathsf{T}} & M \\ M^* & \frac{1}{2}S + \frac{1}{2}S^{\mathsf{T}} \end{pmatrix} \tag{7.14}$$

is positive semidefinite, and the diagonal entries of $\frac{1}{2}R + \frac{1}{2}R^{\mathsf{T}}$ and $\frac{1}{2}S + \frac{1}{2}S^{\mathsf{T}}$, are all equal to one, and therefore the objective value $\langle D, M\rangle$ is no greater than the entangled bias of $G$.

Now let us consider the dual problem, in which one maximizes $\operatorname{Tr}(W)$ over all $W \in \operatorname{Herm}(\mathcal{X}\oplus\mathcal{Y})$, subject to the constraint that

$$\Delta(W) \geq H. \tag{7.15}$$

Notice that the off-diagonal entries of $W$ have absolutely no influence on this problem: they are zeroed out by $\Delta$ in the constraint, and they do not influence the objective function. For this reason there is no generality lost in restricting one's attention to $W$ taking the form

$$W = \frac{1}{2}\begin{pmatrix} \operatorname{Diag}(u) & 0 \\ 0 & \operatorname{Diag}(v) \end{pmatrix} \tag{7.16}$$

for vectors $u \in \mathbb{R}^X$ and $v \in \mathbb{R}^Y$. (Here we're including the factor of $1/2$ for the sake of convenience—we're free to scale the vectors $u$ and $v$ as we choose.) The objective function then becomes

$$\operatorname{Tr}(W) = \frac{1}{2}\sum_{x\in X} u(x) + \frac{1}{2}\sum_{y\in Y} v(y), \tag{7.17}$$

while the constraint becomes equivalent to

$$\begin{pmatrix} \operatorname{Diag}(u) & -D \\ -D^* & \operatorname{Diag}(v) \end{pmatrix} \geq 0. \tag{7.18}$$

Summarizing what we have just concluded about Optimization Problem 7.1, we arrive at the following expression of the same problem.

**Optimization Problem 7.2** (SDP for XOR game bias, simplified)

*Primal problem*

$$\begin{aligned} \text{maximize:}\quad & \langle D, M\rangle \\ \text{subject to:}\quad & \begin{pmatrix} R & M \\ M^* & S \end{pmatrix} \geq 0, \\ & R(x,x) = 1 \text{ for all } x \in X, \\ & S(y,y) = 1 \text{ for all } y \in Y, \\ & R \in \operatorname{Pos}(\mathcal{X}), \\ & S \in \operatorname{Pos}(\mathcal{Y}), \\ & M \in \mathrm{L}(\mathbb{R}^Y, \mathbb{R}^X). \end{aligned}$$

*Dual problem*

$$\begin{aligned} \text{minimize:}\quad & \frac{1}{2}\sum_{x\in X} u(x) + \frac{1}{2}\sum_{y\in Y} v(y) \\ \text{subject to:}\quad & \begin{pmatrix} \operatorname{Diag}(u) & -D \\ -D^* & \operatorname{Diag}(v) \end{pmatrix} \geq 0, \\ & u \in \mathbb{R}^X, \\ & v \in \mathbb{R}^Y. \end{aligned}$$

It will be helpful later in the lecture for us to observe at this point that for any dual-optimal choice of $u$ and $v$, it must be the case that

$$\sum_{x\in X} u(x) = \sum_{y\in Y} v(y). \tag{7.19}$$

The reason is that for any choice of $u$ and $v$, and for any $\lambda > 0$, the operator

$$\begin{pmatrix} \operatorname{Diag}(u) & -D \\ -D^* & \operatorname{Diag}(v) \end{pmatrix} \tag{7.20}$$

is positive semidefinite if and only if

$$\begin{pmatrix} \lambda \operatorname{Diag}(u) & -D \\ -D^* & \frac{1}{\lambda}\operatorname{Diag}(v) \end{pmatrix} \tag{7.21}$$

is positive semidefinite. The dual objective value obtained by the operator (7.21), assuming it is positive semidefinite, is equal to

$$\frac{\lambda}{2}\sum_{x\in X} u(x) + \frac{1}{2\lambda}\sum_{y\in Y} v(y). \tag{7.22}$$

Assuming that $D$ is nonzero, which is always the case when it arises from an XOR game $G$, it must be the case that $\sum_{x\in X} u(x)$ and $\sum_{y\in Y} v(y)$ are strictly positive, and in this case the minimum value for (7.22) occurs when

$$\lambda = \sqrt{\frac{\sum_{y\in Y} v(y)}{\sum_{x\in X} u(x)}}, \tag{7.23}$$

and this is the unique choice of $\lambda$ for which the minimum is obtained. Thus, under the assumption that $u$ and $v$ are optimal, it must be the case that $\lambda = 1$, which is equivalent to (7.19).

We can now verify that the entangled value of the CHSH game is $\cos^2(\pi/8)$, as claimed in the previous lecture.

**Example 7.3** (CHSH game entangled bias/value)**.** Recall that the CHSH game is the XOR game $G = (X, Y, \pi, f)$ with

$$\begin{gathered} X = Y = \{0,1\}, \\ \pi(0,0) = \pi(0,1) = \pi(1,0) = \pi(1,1) = \frac{1}{4}, \\ f(x,y) = x \wedge y. \end{gathered} \tag{7.24}$$

The matrix $D(x,y) = \pi(x,y)(-1)^{f(x,y)}$ is then equal to

$$D = \frac{1}{4}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{7.25}$$

We will verify that the optimal value of Optimization Problem 7.2 for the game $G$ is $\varepsilon^*(G) = 1/\sqrt{2}$.

First choose

$$M = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \tag{7.26}$$

which has spectral norm equal to 1—it is a unitary operator, representing the Hadamard transform—and observe that for $R = S = \mathbb{1}$ we have that

$$Z = \begin{pmatrix} R & M \\ M^* & S \end{pmatrix} \geq 0. \tag{7.27}$$

The diagonal entries of $R$ and $S$ are equal to one, so $Z$ is primal feasible, and it achieves the objective value $\langle D, M\rangle = 1/\sqrt{2}$.

In the dual problem, choosing

$$u = \left(\tfrac{1}{2\sqrt{2}}, \tfrac{1}{2\sqrt{2}}\right) \quad \text{and} \quad v = \left(\tfrac{1}{2\sqrt{2}}, \tfrac{1}{2\sqrt{2}}\right) \tag{7.28}$$

yields a feasible solution. The objective value is the same value just achieved in the primal:

$$\frac{1}{2}\sum_{x\in\{0,1\}} u(x) + \frac{1}{2}\sum_{y\in\{0,1\}} v(y) = \frac{1}{\sqrt{2}}. \tag{7.29}$$

Having obtained the same value in the primal and dual, we have verified that the optimal value, which is the entangled bias, is

$$\varepsilon^*(G) = \frac{1}{\sqrt{2}}. \tag{7.30}$$

This implies that the entangled value of the CHSH game is

$$\omega^*(G) = \frac{1}{2} + \frac{1}{2\sqrt{2}} = \cos^2(\pi/8). \tag{7.31}$$

## 7.2 Strong parallel repetition for XOR games

Next we will prove that XOR games obey a *strong parallel repetition* property. To explain what this means, let us first discuss the notion of parallel repetition in greater generality.

Given arbitrary nonlocal games $G_1, \ldots, G_n$, described by probability distributions

$$\begin{aligned} \pi_1 &: X_1 \times Y_1 \rightarrow [0,1], \\ &\vdots \\ \pi_n &: X_n \times Y_n \rightarrow [0,1], \end{aligned} \tag{7.32}$$

and predicates

$$\begin{aligned} V_1 &: A_1 \times B_1 \times X_1 \times Y_1 \rightarrow \{0,1\}, \\ &\vdots \\ V_n &: A_n \times B_n \times X_n \times Y_n \rightarrow \{0,1\}, \end{aligned} \tag{7.33}$$

respectively, one defines the nonlocal game $G = G_1 \wedge \cdots \wedge G_n$ by the distribution

$$\pi((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \pi_1(x_1, y_1) \cdots \pi_n(x_n, y_n) \tag{7.34}$$

and the predicate

$$\begin{aligned} V((a_1,\ldots,a_n),(b_1,\ldots,b_n)|(x_1,\ldots,x_n),(y_1,\ldots,y_n)) \\ = V_1(a_1,b_1|x_1,y_1) \wedge \cdots \wedge V_n(a_n,b_n|x_n,y_n). \end{aligned} \tag{7.35}$$

In words, the game $G$ is run as if it were independent instances of the games $G_1,\ldots,G_n$, where Alice and Bob receive $n$-tuples of questions $(x_1,\ldots,x_n)$ and $(y_1,\ldots,y_n)$ all at the same time, with each question pair $(x_k,y_k)$ being chosen according to $\pi_k$, independent of all other question pairs. They are expected to provide answers $(a_1,\ldots,a_n)$ and $(b_1,\ldots,b_n)$, respectively, and they win the game $G$ if and only if every one of the pairs $(a_k,b_k)$ is correct for the corresponding question pair $(x_k,y_k)$ in the game $G_k$.

It is important to realize, though, that Alice and Bob are not required to treat the individual games $G_1,\ldots,G_n$ independently. They may, in particular, attempt to correlate their answers in otherwise independent game instances to their advantage. The following example illustrates that it is indeed possible for them to gain an advantage along these lines.

**Example 7.4** (Parallel repetition of the FFL game)**.** The classical value of the FFL game, introduced in the previous lecture, is 2/3. Remarkably, the classical value of the two-fold repetition FFL $\wedge$ FFL of this game with itself is also 2/3. A deterministic strategy that achieves this winning probability is that Alice and Bob both respond to their question pairs by simply swapping the two binary values. That is, Alice's answers are determined by the function $f : X \times X \rightarrow A \times A$ and Bob's answers are determined by the function $g : Y \times Y \rightarrow B \times B$, where

$$f(x_1,x_2) = (x_2,x_1) \quad \text{and} \quad g(y_1,y_2) = (y_2,y_1). \tag{7.36}$$

For this strategy, the winning condition in both games is the same:

$$x_1 \vee x_2 \neq y_1 \vee y_2. \tag{7.37}$$

That is, they either win both games or lose both games, never winning just one of them. If $(x_1,y_1)$ and $(x_2,y_2)$ are independently and uniformly generated from the set $\{(0,0),(0,1),(1,0)\}$ then the above condition fails only when

$$((x_1,y_1),(x_2,y_2)) \in \{((0,0),(0,0)),\ ((1,0),(0,1)),\ ((0,1),(1,0))\}. \tag{7.38}$$

Such question pairs are selected with probability 3/9 = 1/3, so the winning probability is 2/3, as claimed.

The example of the FFL game illustrates that the classical value of a nonlocal game $G = G_1 \wedge \cdots \wedge G_n$ is not always equal to the product of the values of

$G_1, \ldots, G_n$, which is what one obtains when Alice and Bob play the games independently. That is, one always has

$$\omega(G_1 \wedge \cdots \wedge G_n) \geq \omega(G_1) \cdots \omega(G_n), \tag{7.39}$$

but in some cases the inequality is strict. A similar phenomenon occurs for the entangled value—although we did not prove it, the entangled value of the FFL game agrees with the classical value, and therefore

$$\omega^*(\text{FFL} \wedge \text{FFL}) \geq \omega(\text{FFL} \wedge \text{FFL}) = \frac{2}{3} = \omega^*(\text{FFL}) > \omega^*(\text{FFL})^2. \tag{7.40}$$

We will prove that the entangled value of XOR games forbids this type of advantage. That is, if $G_1, \ldots, G_n$ are XOR games, then

$$\omega^*(G_1 \wedge \cdots \wedge G_n) = \omega^*(G_1) \cdots \omega^*(G_n). \tag{7.41}$$

This is the property referred to as *strong parallel repetition*. In particular, if $G$ is an XOR game and we write

$$G^{\wedge n} = G \wedge \cdots \wedge G \quad (n \text{ times}), \tag{7.42}$$

then it necessarily holds that

$$\omega^*(G^{\wedge n}) = \omega^*(G)^n. \tag{7.43}$$

The first step in proving that (7.41) holds for XOR games $G_1, \ldots, G_n$ is to define the XOR of two (or more) XOR games. Suppose that $G_1$ and $G_2$ are XOR games, specified by probability distributions

$$\pi_1 : X_1 \times Y_1 \rightarrow [0,1] \quad \text{and} \quad \pi_2 : X_2 \times Y_2 \rightarrow [0,1] \tag{7.44}$$

along with functions

$$f_1 : X_1 \times Y_1 \rightarrow \{0,1\} \quad \text{and} \quad f_2 : X_2 \times Y_2 \rightarrow \{0,1\}. \tag{7.45}$$

The XOR $G_1 \oplus G_2$ of these two games is the XOR game defined by the distribution $\pi : (X_1 \times X_2) \times (Y_1 \times Y_2) \rightarrow [0,1]$ given by

$$\pi((x_1,x_2),(y_1,y_2)) = \pi_1(x_1,y_1)\pi_2(x_2,y_2) \tag{7.46}$$

and the function $f : (X_1 \times X_2) \times (Y_1 \times Y_2) \rightarrow \{0,1\}$ given by

$$f((x_1,x_2),(y_1,y_2)) = f_1(x_1,y_1) \oplus f_2(x_2,y_2). \tag{7.47}$$

In words, the question sets for the game $G_1 \oplus G_2$ are $X_1 \times X_2$ and $Y_1 \times Y_2$, and if Alice receives $(x_1, x_2)$ and Bob receives $(y_1, y_2)$, they are expected to provide answers $a, b \in \{0, 1\}$ that are consistent with the equations

$$a = a_1 \oplus a_2 \quad \text{and} \quad b = b_1 \oplus b_2 \tag{7.48}$$

for some choice of $a_1, a_2, b_1, b_2 \in \{0, 1\}$ that would cause $(a_1, b_1)$ to be correct for $(x_1, y_1)$ and $(a_2, b_2)$ to be correct for $(x_2, y_2)$. The XOR of more than two XOR games is defined similarly (or, equivalently, by applying this definition iteratively).

It is evident, for any two XOR games $G_1$ and $G_2$, that

$$\varepsilon^*(G_1 \oplus G_2) \geq \varepsilon^*(G_1)\varepsilon^*(G_2). \tag{7.49}$$

The reason is that if Alice and Bob play the game $G_1 \oplus G_2$ by playing $G_1$ and $G_2$ independently and optimally, and then answer according to the XOR of their answers in $G_1$ and $G_2$, then they will achieve the bias $\varepsilon^*(G_1)\varepsilon^*(G_2)$. We will prove that this is, in fact, the best they can do. That is, we will prove

$$\varepsilon^*(G_1 \oplus G_2) = \varepsilon^*(G_1)\varepsilon^*(G_2). \tag{7.50}$$

This will be done using semidefinite programming duality.

In particular, consider the dual form of Optimization Problem 7.2, the semidefinite program for the entangled bias of an XOR game, for the two separate XOR games $G_1$ and $G_2$. Suppose that $(u_1, v_1)$ and $(u_2, v_2)$ represent dual-optimal solutions to these semidefinite programs. As argued previously, this implies

$$\sum_{x_1} u_1(x_1) = \sum_{y_1} v_1(y_1) \quad \text{and} \quad \sum_{x_2} u_2(x_2) = \sum_{y_2} v_2(y_2). \tag{7.51}$$

The dual form of Optimization Problem 7.2 for the entangled bias of the XOR game $G_1 \oplus G_2$ has the following form:

$$\begin{aligned}
\text{minimize:}\quad & \frac{1}{2} \sum_{x \in X_1 \times X_2} u(x) + \frac{1}{2} \sum_{y \in Y_1 \times Y_2} v(y) \\
\text{subject to:}\quad & \begin{pmatrix} \operatorname{Diag}(u) & -D_1 \otimes D_2 \\ -D_1^* \otimes D_2^* & \operatorname{Diag}(v) \end{pmatrix} \geq 0, \\
& u \in \mathbb{R}^{X_1 \times X_2},\ v \in \mathbb{R}^{Y_1 \times Y_2}.
\end{aligned}$$

Here, $D_1$ and $D_2$ are the operators given by

$$\begin{aligned}
D_1(x_1, y_1) &= \pi_1(x_1, y_1)(-1)^{f_1(x_1, y_1)}, \\
D_2(x_2, y_2) &= \pi_2(x_2, y_2)(-1)^{f_2(x_2, y_2)}.
\end{aligned} \tag{7.52}$$

Next we observe that $u = u_1 \otimes u_2$ and $v = v_1 \otimes v_2$ provide a dual feasible solution to Optimization Problem 7.2 for $\varepsilon^*(G_1 \oplus G_2)$. To prove that this is so, it is helpful to observe that if

$$\begin{pmatrix} P_1 & X_1 \\ X_1^* & Q_1 \end{pmatrix} \geq 0 \quad \text{and} \quad \begin{pmatrix} P_2 & X_2 \\ X_2^* & Q_2 \end{pmatrix} \geq 0 \tag{7.53}$$

then

$$\begin{pmatrix} P_1 \otimes P_2 & X_1 \otimes X_2 \\ X_1^* \otimes X_2^* & Q_1 \otimes Q_2 \end{pmatrix} \geq 0. \tag{7.54}$$

This is so because, after some simultaneous re-ordering of rows and columns, the matrix above is a principal submatrix of the positive semidefinite matrix

$$\begin{pmatrix} P_1 & X_1 \\ X_1^* & Q_1 \end{pmatrix} \otimes \begin{pmatrix} P_2 & X_2 \\ X_2^* & Q_2 \end{pmatrix}. \tag{7.55}$$

Alternatively, by (7.53) we may conclude that

$$X_1 = \sqrt{P_1} K_1 \sqrt{Q_1} \quad \text{and} \quad X_2 = \sqrt{P_2} K_2 \sqrt{Q_2} \tag{7.56}$$

for $K_1$ and $K_2$ satisfying $\|K_1\| \leq 1$ and $\|K_2\| \leq 1$. As we therefore have

$$X_1 \otimes X_2 = \sqrt{P_1 \otimes P_2} (K_1 \otimes K_2) \sqrt{Q_1 \otimes Q_2}, \tag{7.57}$$

and $\|K_1 \otimes K_2\| = \|K_1\| \|K_2\| \leq 1$, it follows that (7.54) holds. One may therefore conclude that

$$\begin{pmatrix} \operatorname{Diag}(u_1) \otimes \operatorname{Diag}(u_2) & D_1 \otimes D_2 \\ D_1^* \otimes D_2^* & \operatorname{Diag}(v_1) \otimes \operatorname{Diag}(v_2) \end{pmatrix} \geq 0. \tag{7.58}$$

As $\operatorname{Diag}(u_1) \otimes \operatorname{Diag}(u_2) = \operatorname{Diag}(u)$ and $\operatorname{Diag}(v_1) \otimes \operatorname{Diag}(v_2) = \operatorname{Diag}(v)$, we have

$$\begin{aligned} &\begin{pmatrix} \operatorname{Diag}(u) & -D_1 \otimes D_2 \\ -D_1^* \otimes D_2^* & \operatorname{Diag}(v) \end{pmatrix} \\ &\quad = \begin{pmatrix} \mathbb{1} & 0 \\ 0 & -\mathbb{1} \end{pmatrix} \begin{pmatrix} \operatorname{Diag}(u) & D_1 \otimes D_2 \\ D_1^* \otimes D_2^* & \operatorname{Diag}(v) \end{pmatrix} \begin{pmatrix} \mathbb{1} & 0 \\ 0 & -\mathbb{1} \end{pmatrix} \geq 0. \end{aligned} \tag{7.59}$$

The objective value of the dual solution $(u, v)$ is

$$\begin{aligned} \frac{1}{2} \sum_x u(x) + \frac{1}{2} \sum_y v(y) &= \frac{1}{2} \sum_{x_1} u_1(x_1) \sum_{x_2} u_2(x_2) + \frac{1}{2} \sum_{y_1} v_1(y_1) \sum_{y_2} v_2(y_2) \\ &= \frac{1}{2} \varepsilon^*(G_1)\, \varepsilon^*(G_2) + \frac{1}{2} \varepsilon^*(G_1)\, \varepsilon^*(G_2) = \varepsilon^*(G_1)\, \varepsilon^*(G_2), \end{aligned} \tag{7.60}$$

where we have used the fact that

$$\sum_{x_1} u_1(x_1) = \sum_{y_1} v_1(y_1) = \varepsilon^*(G_1) \quad \text{and} \quad \sum_{x_2} u_2(x_2) = \sum_{y_2} v_2(y_2) = \varepsilon^*(G_2). \tag{7.61}$$

Therefore, $\varepsilon^*(G_1 \oplus G_2) \leq \varepsilon^*(G_1)\varepsilon^*(G_2)$, which implies (7.50).

We are now prepared to prove that, for XOR games $G_1, \ldots, G_n$, it holds that

$$\omega^*(G_1 \wedge \cdots \wedge G_n) = \omega^*(G_1) \cdots \omega^*(G_n). \tag{7.62}$$

As we have already observed, it holds that

$$\omega^*(G_1 \wedge \cdots \wedge G_n) \geq \omega^*(G_1) \cdots \omega^*(G_n), \tag{7.63}$$

and therefore it remains to prove

$$\omega^*(G_1 \wedge \cdots \wedge G_n) \leq \omega^*(G_1) \cdots \omega^*(G_n), \tag{7.64}$$

Assume hereafter that XOR games $G_1, \ldots, G_n$ have been fixed, and consider an arbitrary strategy for Alice and Bob in the game $G_1 \wedge \cdots \wedge G_n$, through which Alice and Bob answer question tuples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ with answer tuples $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$, respectively. We will consider how well this strategy performs for the XOR game

$$G_{k_1} \oplus \cdots \oplus G_{k_m}, \tag{7.65}$$

for various choices of a subset $S = \{k_1, \ldots, k_m\} \subseteq \{1, \ldots, n\}$, provided that we define Alice and Bob's answers as

$$a_{k_1} \oplus \cdots \oplus a_{k_m} \quad \text{and} \quad b_{k_1} \oplus \cdots \oplus b_{k_m} \tag{7.66}$$

and where we assume that they have chosen to share randomly generated question pairs $(x_k, y_k)$ for those choices of $k \notin S$.

To do this we will define binary-valued random variables $Z_1, \ldots, Z_n$ as

$$Z_k = a_k \oplus b_k \oplus f_k(x_k, y_k), \tag{7.67}$$

where we view $x_1, \ldots, x_n$, $y_1, \ldots, y_n$, $a_1, \ldots, a_n$, and $b_1, \ldots, b_n$ as random variables, with $(x_1, y_1), \ldots, (x_n, y_n)$ distributed independently according to the distributions $\pi_1, \ldots, \pi_n$ given by the games $G_1, \ldots, G_n$ and $a_1, \ldots, a_n, b_1, \ldots, b_n$ distributed and correlated with $x_1, \ldots, x_n, y_1, \ldots, y_n$ in whatever manner Alice and Bob's strategy determines. It holds that

$$Z_k = \begin{cases} 0 & \text{if Alice and Bob win } G_k \\ 1 & \text{if Alice and Bob lose } G_k \end{cases} \tag{7.68}$$

and therefore the probability of winning minus losing game $G_k$ is equal to the expectation

$$\mathrm{E}\Big((-1)^{Z_k}\Big). \tag{7.69}$$

More generally, if Alice and Bob's strategy is transformed into a strategy for the XOR game $G_{k_1} \oplus \cdots \oplus G_{k_m}$ as suggested above, we find that

$$Z_{k_1} \oplus \cdots \oplus Z_{k_m} = \begin{cases} 0 & \text{if Alice and Bob win } G_{k_1} \oplus \cdots \oplus G_{k_m} \\ 1 & \text{if Alice and Bob lose } G_{k_1} \oplus \cdots \oplus G_{k_m} \end{cases} \tag{7.70}$$

and therefore the probability they win minus the probability they lose in this XOR game is

$$\mathrm{E}\Big((-1)^{Z_{k_1}+\cdots+Z_{k_m}}\Big). \tag{7.71}$$

Now, we know that probability of winning minus the probability of losing in the game $G_{k_1} \oplus \cdots \oplus G_{k_m}$ is upper-bounded by its bias:

$$\mathrm{E}\Big((-1)^{Z_{k_1}+\cdots+Z_{k_m}}\Big) \leq \varepsilon^*(G_{k_1} \oplus \cdots \oplus G_{k_m}) = \varepsilon^*(G_{k_1}) \cdots \varepsilon^*(G_{k_m}). \tag{7.72}$$

Here we have used the fact concerning XOR game biases proved above, which is the key to making the entire argument work. The probability that Alice and Bob's strategy wins $G_1 \wedge \cdots \wedge G_n$ is therefore bounded as follows:

$$\begin{aligned} \Pr(Z_1 = 0, \ldots, Z_n = 0) &= \mathrm{E}\bigg(\bigg(\frac{1+(-1)^{Z_1}}{2}\bigg) \cdots \bigg(\frac{1+(-1)^{Z_n}}{2}\bigg)\bigg) \\ &= \frac{1}{2^n} \sum_{S \subseteq \{1,\ldots,n\}} \mathrm{E}\Big((-1)^{\sum_{k\in S} Z_k}\Big) \\ &\leq \frac{1}{2^n} \sum_{S \subseteq \{1,\ldots,n\}} \prod_{k \in S} \varepsilon^*(G_k) \\ &= \bigg(\frac{1+\varepsilon^*(G_1)}{2}\bigg) \cdots \bigg(\frac{1+\varepsilon^*(G_n)}{2}\bigg) \\ &= \omega^*(G_1) \cdots \omega^*(G_n). \end{aligned} \tag{7.73}$$

Maximizing over all possible entangled strategies for Alice and Bob yields

$$\omega^*(G_1 \wedge \cdots \wedge G_n) \leq \omega^*(G_1) \cdots \omega^*(G_n), \tag{7.74}$$

as required.

Lecture 8

# The hierarchy of Navascués, Pironio, and Acín

In this lecture, we will define and study a class of strategies for nonlocal games known as *commuting measurement strategies*, or alternatively as *commuting operator strategies*. These strategies include all entangled strategies, in a sense that will be made more precise momentarily—and it was not long ago that this inclusion was proved to be proper by Slofstra [arXiv:1606.03140]. A proof that the *values* defined by the classes of entangled and commuting measurement strategies are different, where we take the supremum winning probability over the two classes of strategies, has only recently been announced by Ji, Natarajan, Vidick, Wright, and Yuen [arXiv:2001.04383]. But be warned—the paper is over 200 pages long. This refutes the famous Connes' embedding conjecture from the subject of von Neumann algebras, so it is worth every page it needs.

We will then analyze the semidefinite programming hierarchy of Navascués, Pironio, and Acín, better known as the *NPA hierarchy*, which provides us with a uniform family of semidefinite programs that converges to the *commuting measurement value* of any nonlocal game. This result is, in fact, a necessary ingredient in Ji, Natarajan, Vidick, Wright, and Yuen's proof.

## 8.1 Representing and comparing strategies

Let us fix question sets $X$ and $Y$, and answer sets $A$ and $B$, for a nonlocal game.

It is natural to think about *strategies* for nonlocal games having these question and answer sets as being represented by operators of the form

$$M \in \mathrm{L}(\mathbb{R}^X \otimes \mathbb{R}^Y, \mathbb{R}^A \otimes \mathbb{R}^B), \tag{8.1}$$

or equivalently as matrices whose columns are indexed by question pairs and whose rows are indexed by answer pairs. To be precise, the value $M(a,b|x,y)$ represents the probability that Alice and Bob answer the questions $(x,y)$ with the answer $(a,b)$.[1]

This representation is nice because not only does $M$ store the probabilities with which Alice and Bob respond to each question pair $(x,y)$ with answers $(a,b)$, but also the action of $M$ as a linear operator has meaning. For example, assuming the referee selects question pairs according to a probability vector $\pi$, we have that Alice and Bob's answers are distributed according to the vector $M\pi$. Perhaps more useful is to consider the vector

$$v = \sum_{(x,y)\in X\times Y} \pi(x,y)\,|x\rangle \otimes |y\rangle \otimes |x\rangle \otimes |y\rangle \tag{8.2}$$

and to observe that

$$\bigl(M \otimes \mathbb{1}_{X\times Y}\bigr)v \tag{8.3}$$

represents the joint probability distribution of quadruples $(a,b,x,y)$ that arise from the selection of the questions $(x,y)$ according to $\pi$ together with Alice and Bob's answers to those questions.

Notice also that the probability for a strategy $M \in \mathrm{L}(\mathbb{R}^X \otimes \mathbb{R}^Y, \mathbb{R}^A \otimes \mathbb{R}^B)$ to win a particular nonlocal game $G = (X, Y, A, B, \pi, V)$ is equal to

$$\sum_{(x,y)\in X\times Y} \pi(x,y) \sum_{(a,b)\in A\times B} V(a,b|x,y)M(a,b|x,y), \tag{8.4}$$

which one may alternatively express as the inner product $\langle K, M\rangle$, where the operator $K \in \mathrm{L}(\mathbb{R}^X \otimes \mathbb{R}^Y, \mathbb{R}^A \otimes \mathbb{R}^B)$ is defined as

$$K(a,b|x,y) = \pi(x,y)V(a,b|x,y) \tag{8.5}$$

for all $x \in X$, $y \in Y$, $a \in A$, and $b \in B$.

Now, when we consider a particular *class* of strategies, such as the classical strategies or the entangled strategies, we are effectively defining a subset

$$\mathcal{S} \subset \mathrm{L}(\mathbb{R}^X \otimes \mathbb{R}^Y, \mathbb{R}^A \otimes \mathbb{R}^B) \tag{8.6}$$

of operators that represent strategies in the class under consideration. We then have the associated *value*

$$\omega_{\mathcal{S}}(G) = \sup_{M\in\mathcal{S}} \langle K, M\rangle \tag{8.7}$$

of a game $G$ for the class $\mathcal{S}$.

[1] Throughout the lecture, entries of operators having the form $M \in \mathrm{L}(\mathbb{R}^X \otimes \mathbb{R}^Y, \mathbb{R}^A \otimes \mathbb{R}^B)$ are expressed as $M(a,b|x,y)$ rather than $M((a,b),(x,y))$, to mirror the notation we have adopted for the referee's predicate $V(a,b|x,y)$.

$$\begin{pmatrix}1&1&1&1\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{pmatrix}\qquad\begin{pmatrix}1&0&1&0\\0&1&0&1\\0&0&0&0\\0&0&0&0\end{pmatrix}\qquad\begin{pmatrix}0&1&0&1\\1&0&1&0\\0&0&0&0\\0&0&0&0\end{pmatrix}\qquad\begin{pmatrix}0&0&0&0\\1&1&1&1\\0&0&0&0\\0&0&0&0\end{pmatrix}$$

$$\begin{pmatrix}1&1&0&0\\0&0&0&0\\0&0&1&1\\0&0&0&0\end{pmatrix}\qquad\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}\qquad\begin{pmatrix}0&1&0&0\\1&0&0&0\\0&0&0&1\\0&0&1&0\end{pmatrix}\qquad\begin{pmatrix}0&0&0&0\\1&1&0&0\\0&0&0&0\\0&0&1&1\end{pmatrix}$$

$$\begin{pmatrix}0&0&1&1\\0&0&0&0\\1&1&0&0\\0&0&0&0\end{pmatrix}\qquad\begin{pmatrix}0&0&1&0\\0&0&0&1\\1&0&0&0\\0&1&0&0\end{pmatrix}\qquad\begin{pmatrix}0&0&0&1\\0&0&1&0\\0&1&0&0\\1&0&0&0\end{pmatrix}\qquad\begin{pmatrix}0&0&0&0\\0&0&1&1\\0&0&0&0\\1&1&0&0\end{pmatrix}$$

$$\begin{pmatrix}0&0&0&0\\0&0&0&0\\1&1&1&1\\0&0&0&0\end{pmatrix}\qquad\begin{pmatrix}0&0&0&0\\0&0&0&0\\1&0&1&0\\0&1&0&1\end{pmatrix}\qquad\begin{pmatrix}0&0&0&0\\0&0&0&0\\0&1&0&1\\1&0&1&0\end{pmatrix}\qquad\begin{pmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\1&1&1&1\end{pmatrix}$$

Figure 8.1: The 16 deterministic strategies for binary question and answer pairs.

**Example 8.1.** Suppose $X = Y = A = B = \{0,1\}$. There are 16 deterministic strategies for games having these question and answer sets, and they are represented by the matrices shown in Figure 8.1.

It is natural to associate the classical strategies for these question and answer sets with the *convex hull* of these 16 deterministic strategies to account for the possibility that Alice and Bob make use of randomness.

Representing the CHSH game as an operator $K$, as described above, yields

$$K = \frac{1}{4}\begin{pmatrix}1&1&1&0\\0&0&0&1\\0&0&0&1\\1&1&1&0\end{pmatrix}. \tag{8.8}$$

It is now perhaps more clear by an inspection of this matrix together with those appearing in Figure 8.1 that the classical value of the CHSH game is $3/4$; if $M$ ranges over the matrices representing deterministic strategies, it is possible to make three of the 1s in any of these matrices, but not all four, overlap the nonzero entries of $K$.

Now, if we were to represent the set of all (probabilistic) classical strategies by

$$\mathcal{P} \subset \mathrm{L}(\mathbb{R}^X \otimes \mathbb{R}^Y, \mathbb{R}^A \otimes \mathbb{R}^B) \tag{8.9}$$

and the set of all entangled strategies as

$$\mathcal{E} \subset \mathrm{L}(\mathbb{R}^X \otimes \mathbb{R}^Y, \mathbb{R}^A \otimes \mathbb{R}^B), \tag{8.10}$$

then there are various observations we could make. For example:

1. $\mathcal{P}$ and $\mathcal{E}$ are both convex sets.
2. $\mathcal{P}$ is a polytope.
3. $\mathcal{P} \subseteq \mathcal{E}$, and the inclusion is proper so long as $X$, $Y$, $A$, and $B$ all have at least two elements.

As special cases of (8.7) we have

$$\omega(G) = \sup_{M \in \mathcal{P}} \langle K, M \rangle \quad \text{and} \quad \omega^*(G) = \sup_{M \in \mathcal{E}} \langle K, M \rangle. \tag{8.11}$$

## 8.2 Commuting measurement strategies

Now we will introduce a new set of strategies, called *commuting measurement strategies*. The basic idea is to drop the tensor product structure that is present in an entangled strategy, replacing it with the assumption that Alice and Bob's measurement operators *commute*. We also drop the requirement that the space $\mathcal{H}$ is finite dimensional.[2]

**Definition 8.2.** For a given choice of question sets $X$ and $Y$ and answer sets $A$ and $B$, we say that an operator $M \in \mathrm{L}(\mathbb{R}^X \otimes \mathbb{R}^Y, \mathbb{R}^A \otimes \mathbb{R}^B)$ represents a *commuting measurement strategy* if there exists a complex Hilbert space $\mathcal{H}$, a unit vector $u \in \mathcal{H}$, and projection operators

$$\left\{ P_a^x \,:\, x \in X,\, a \in A \right\} \quad \text{and} \quad \left\{ Q_b^y \,:\, y \in Y,\, b \in B \right\} \tag{8.12}$$

acting on $\mathcal{H}$, such that the following properties are satisfied. The projections represent measurements, in the sense that

$$\sum_{a \in A} P_a^x = \mathbb{1}_{\mathcal{H}} \quad \text{and} \quad \sum_{b \in B} Q_b^y = \mathbb{1}_{\mathcal{H}} \tag{8.13}$$

[2] If we restrict the definition to finite-dimensional $\mathcal{H}$, then we obtain precisely the entangled strategies. It is not the case, in contrast, that allowing the spaces $\mathcal{A}$ and $\mathcal{B}$ appearing in the definition of entangled strategies to be infinite dimensional makes entangled strategies equivalent to commuting measurement strategies.

for all $x \in X$ and $y \in Y$, the two collections of projections commute in pairs, meaning that $\left[P_a^x, Q_b^y\right] = 0$ for all $x \in X$ and $y \in Y$, and we have

$$M(a,b|x,y) = \langle u, P_a^x Q_b^y u \rangle \tag{8.14}$$

for all $x \in X$, $y \in Y$, $a \in A$, and $b \in B$.

Hereafter we will write

$$\mathcal{C} \subset \mathrm{L}(\mathbb{R}^X \otimes \mathbb{R}^Y, \mathbb{R}^A \otimes \mathbb{R}^B) \tag{8.15}$$

to denote the set of all commuting measurement strategies for the question and answer sets $X$, $Y$, $A$, and $B$. The *commuting measurement value* of $G$, which is denoted $\omega^c(G)$, is the supremum value of the winning probability for $G$ taken over all commuting measurement strategies for Alice and Bob:

$$\omega^c(G) = \sup_{M \in \mathcal{C}} \langle K, M \rangle, \tag{8.16}$$

assuming again that $K$ is defined from $G$ as in (8.5).

Here are a couple of known facts about the set $\mathcal{C}$ of commuting measurement strategies and its relationship to the entangled strategies $\mathcal{E}$, stated without proof:

1. $\mathcal{C}$ is compact and convex.
2. $\mathcal{E} \subseteq \mathcal{C}$.

Proving that $\mathcal{C}$ is convex and includes $\mathcal{E}$ is a good exercise that I will leave to you. The fact that $\mathcal{C}$ is compact happens to follow from what we will prove later in the lecture (as does convexity). The containment $\mathcal{E} \subseteq \mathcal{C}$ is proper, as was mentioned at the start of the lecture.

## 8.3 The NPA hierarchy

This section is devoted to a description of the semidefinite programming hierarchy of Navascués, Pironio, and Acín—the *NPA hierarchy*—which is principally concerned with the commuting measurement strategies. Its convergence, and what exactly that means, will be discussed in the section following this one.

### Basics of strings

When defining the NPA hierarchy in precise terms, it is helpful to make use of some elementary concepts concerning *strings* of symbols. In particular, we will be discussing matrices and vectors whose entries are indexed by strings of varying lengths. These concepts are elementary and quite familiar within theoretical computer science, and it will take just a moment to become familiar with them in case you are not.

Suppose that an alphabet $\Sigma$, which is a finite and nonempty set whose elements are viewed as symbols, has been fixed. A *string* over $\Sigma$ is any finite, ordered sequence of symbols from $\Sigma$. (Infinite sequences of symbols are *not* to be considered as strings.) The *length* of a string is the total number of symbols, counting all repetitions, appearing in that string.

For example, if $\Sigma = \{0,1\}$, then 0, 0110, and 1000100100011 are examples of strings over $\Sigma$; the string 0 has length 1, the string 0110 has length 4, and the string 1000100100011 has length 13. There is a special string that has length 0, and this string is called the *empty string*. We use the Greek letter $\varepsilon$ to denote this string.

For every nonnegative integer $n$, we write $\Sigma^{\leq n}$ to denote the set of all strings having length at most $n$ and we write $\Sigma^*$ to denote the set of all strings, of any (finite) length, over $\Sigma$. For every alphabet $\Sigma$, the set $\Sigma^*$ is countably infinite.

Lastly, given any string $s \in \Sigma^*$, we let $s^{\mathsf{R}}$ denote the string obtained by *reversing* the ordering of the symbols in $s$. For example, if $s = 00010$, then $s^{\mathsf{R}} = 01000$.

## Intuition behind the NPA hierarchy

Next, let us introduce the basic idea behind the NPA hierarchy.

When we think about a particular commuting measurement strategy, represented by a complex Hilbert space $\mathcal{H}$, a unit vector $u \in \mathcal{H}$, and collections of projection operators $\{P_a^x : x \in X, a \in A\}$ and $\{Q_b^y : y \in Y, b \in B\}$ acting on $\mathcal{H}$, as described above, our interest is naturally with the numbers

$$M(a,b|x,y) = \langle u, P_a^x Q_b^y u \rangle, \tag{8.17}$$

ranging over all $x \in X$, $y \in Y$, $a \in A$, and $b \in B$.

These numbers arise when we consider the *Gram matrix* of the vectors

$$\{u\} \cup \{P_a^x u : x \in X, a \in A\} \cup \{Q_b^y u : y \in Y, b \in B\}. \tag{8.18}$$

Among the entries of this matrix, one finds all of the values

$$\langle P_a^x u, Q_b^y u \rangle = \langle u, P_a^x Q_b^y u \rangle = M(a,b|x,y) \tag{8.19}$$

that we ostensibly care about, as well as others, including (for instance)

$$\langle P_a^x u, P_c^z u \rangle, \quad \langle u, Q_b^y u \rangle, \quad \text{and} \quad \langle u, u \rangle, \tag{8.20}$$

for appropriate choices of $x$, $y$, $z$, $a$, $b$, and $c$.

We may now think about the various properties that must hold for such a Gram matrix, but before we do this let us discuss how we will index the entries of such matrices. Assuming that question and answer sets $X$, $Y$, $A$, and $B$ have been fixed, we will introduce three alphabets:

$$\Sigma_A = X \times A, \quad \Sigma_B = Y \times B, \quad \text{and} \quad \Sigma = \Sigma_A \sqcup \Sigma_B. \tag{8.21}$$

Here, $\sqcup$ denotes the *disjoint union*, meaning that $\Sigma_A$ and $\Sigma_B$ are to be treated as disjoint sets when forming $\Sigma$. In words, there are $|X \times A| + |Y \times B|$ symbols in the alphabet $\Sigma$, one symbol for each pair $(x, a) \in X \times A$ and a separate symbol for each pair $(y, b) \in Y \times B$. The collection of vectors (8.18) may naturally be labeled by the set

$$\{\varepsilon\} \cup \Sigma = \Sigma^{\leq 1}, \tag{8.22}$$

and so we may consider that the Gram matrix of these vectors has rows and columns indexed by this set.

Now, supposing that

$$R \in \operatorname{Pos}\left(\mathbb{C}^{\Sigma^{\leq 1}}\right) \tag{8.23}$$

is such a Gram matrix—and naturally this is a positive semidefinite matrix, as all Gram matrices are—there are various things one may say, based on the conditions that commuting measurement strategies must meet. In particular, we observe these conditions:

1. $R(\varepsilon, \varepsilon) = 1$, given that $u$ is a unit vector.
2. For every $x \in X$ we must have

   $$\sum_{a \in A} R((x, a), s) = R(\varepsilon, s) \quad \text{and} \quad \sum_{a \in A} R(s, (x, a)) = R(s, \varepsilon), \tag{8.24}$$

   and likewise for every $y \in Y$ we must have

   $$\sum_{b \in B} R((y, b), s) = R(\varepsilon, s) \quad \text{and} \quad \sum_{b \in B} R(s, (y, b)) = R(s, \varepsilon), \tag{8.25}$$

   for every $s \in \Sigma^{\leq 1}$ (in all four equalities). These conditions reflect the fact that summing over all operators in any given measurement yields the identity operator, as in (8.13).
3. For every $x \in X$ and $a, c \in A$ satisfying $a \neq c$, we have

   $$R((x, a), (x, c)) = 0, \tag{8.26}$$

   because $P_a^x$ and $P_c^x$ must be orthogonal projection operators. Similarly, for every $y \in Y$ and $b, d \in B$ satisfying $b \neq d$, we have

   $$R((y, b), (y, d)) = 0. \tag{8.27}$$

4. For every $(z,c) \in \Sigma$, we have the equality

$$R((z,c),(z,c)) = R(\varepsilon,(z,c)) = R((z,c),\varepsilon); \tag{8.28}$$

each $P_a^x$ and $Q_b^y$ is a projection operator, thus squaring to itself.

5. For every $(x,a) \in X \times A$ and $(y,b) \in Y \times B$ we have

$$R((x,a),(y,b)) = R((y,b),(x,a)), \tag{8.29}$$

as $P_a^x$ and $Q_b^y$ commute.

**Remark 8.3.** We can actually do a bit better in the case of the fifth item, and say that every entry $R((x,a),(y,b))$ must be a nonnegative real number, given that

$$R((x,a),(y,b)) = \langle P_a^x u, Q_b^y u \rangle = \left\| P_a^x Q_b^y u \right\|^2, \tag{8.30}$$

and likewise for $R((y,b),(x,a))$. This is a stronger condition than (8.29), as the entries $R((x,a),(y,b))$ and $R((y,b),(x,a))$ must be equal if they are real given that $R$ is Hermitian. This stronger claim is not really needed though, and does not appear in the formal description of the NPA hierarchy coming up.

Let us now take $\mathcal{C}_1$ to be the subset of $\mathrm{L}(\mathbb{R}^X \otimes \mathbb{R}^Y, \mathbb{R}^A \otimes \mathbb{R}^B)$ containing all $M$ for which there exists a positive semidefinite operator $R$ satisfying items 1 through 5 above, as well as

$$M(a,b|x,y) = R((x,a),(y,b)) \tag{8.31}$$

for every $x \in X$, $y \in Y$, $a \in A$, and $b \in B$.

Observe that $\mathcal{C} \subseteq \mathcal{C}_1$, as every commuting measurement strategy defines a Gram matrix that satisfies the conditions of items 1 through 5. Thus, assuming once again that for a given nonlocal game $G = (X,Y,A,B,\pi,V)$ we have defined $K$ as in (8.5), we see that

$$\omega^c(G) = \sup_{M \in \mathcal{C}} \langle K, M \rangle \leq \sup_{M \in \mathcal{C}_1} \langle K, M \rangle. \tag{8.32}$$

The inclusion $\mathcal{C} \subseteq \mathcal{C}_1$ is proper in general, but we can still use this relationship to obtain upper bounds on the commuting operator value (and therefore on the entangled value) of nonlocal games.

Now, notice that items 1 through 5 in the list above are all *affine linear constraints* on $R$. (With the exception of $R(\varepsilon,\varepsilon) = 1$ they are all linear constraints.) If we define a Hermitian operator $H \in \mathrm{Herm}\big(\mathbb{C}^{\Sigma^{\leq 1}}\big)$ as

$$H((x,a),(y,b)) = H((y,b),(x,a)) = \frac{1}{2}\pi(x,y)V(a,b|x,y) \tag{8.33}$$

for all $x \in X$, $y \in Y$, $a \in A$, and $b \in B$, and with all other entries equal to zero, we see that $\langle K, M\rangle = \langle H, R\rangle$. Thus, the optimization of $\langle K, M\rangle$ over all $M \in \mathcal{C}_1$ is represented by a semidefinite program, where we optimize $\langle H, R\rangle$ over all positive semidefinite $R$ satisfying the affine linear constraints imposed by items 1 through 5 in the list above.

The semidefinite program just suggested is the *first* level of the NPA hierarchy. The idea behind subsequent levels is to consider not just the Gram matrix of the vectors (8.18), but of larger sets of vectors that include ones such as

$$P_a^x Q_b^y u,\ P_a^x Q_b^y P_c^z u,\ \text{etc.} \tag{8.34}$$

This will lead us to consider operators $R$ whose rows and columns are indexed not by $\Sigma^{\leq 1}$, but by $\Sigma^{\leq k}$ for larger choices of $k$. (Larger choices for $k$ yield higher levels in the hierarchy.) Although the inner products between most pairs of these vectors are not informative to the task of determining how well such a strategy performs in a given nonlocal game, the benefit comes from the introduction of additional *constraints* that reflect the same properties through which the five affine linear constraints above were derived. Indeed, the sequence of optimal values obtained from this hierarchy always converges to $\omega^c(G)$, as we will prove in the next section.

## Formal description of the hierarchy

We are now ready to formally define the NPA hierarchy. We will begin by defining an equivalence relation $\sim$ on strings over the alphabet $\Sigma = \Sigma_A \sqcup \Sigma_B$ defined above. This equivalence relation will be used to equate various entries in matrices that generalize items 4 and 5 in the list described above.

Specifically, we take this equivalence relation to be the one generated by these rules holding for every $s, t \in \Sigma^*$, $(x, a) \in \Sigma_A$, and $(y, b) \in \Sigma_B$:

1. $s(x,a)t \sim s(x,a)(x,a)t$ and $s(y,b)t \sim s(y,b)(y,b)t$.
2. $s(x,a)(y,b)t \sim s(y,b)(x,a)t$.

That is, two strings are equivalent with respect to the relation $\sim$ if and only if one can be obtained from the other by any number of applications of the above rules. These equivalences reflect the fact that projections square to themselves (the first rule) and Alice's measurements commute with Bob's measurements (the second rule).

Next, notice that the values that appear in any Gram matrix $R$ of the form suggested above always take the form

$$\left\langle u, \Pi_{c_1}^{z_1} \cdots \Pi_{c_n}^{z_n} u \right\rangle \tag{8.35}$$

for some finite sequence of projection operators $\Pi^{z_1}_{c_1}, \ldots, \Pi^{z_n}_{c_n}$ selected from Alice and Bob's projections.

With this observation in mind, we say that a function of the form $\phi : \Sigma^* \to \mathbb{C}$ is *admissible* if it satisfies these properties:

1. $\phi(\varepsilon) = 1$.
2. For all strings $s, t \in \Sigma^*$ we have

   $$\sum_{a \in A} \phi(s(x,a)t) = \phi(st) \quad \text{and} \quad \sum_{b \in B} \phi(s(y,b)t) = \phi(st) \tag{8.36}$$

   for every $x \in X$ and $y \in Y$.
3. For all strings $s, t \in \Sigma^*$ we have

   $$\phi(s(x,a)(x,c)t) = 0 \quad \text{and} \quad \phi(s(y,b)(y,d)t) = 0 \tag{8.37}$$

   for every $x \in X$ and $a, c \in A$ satisfying $a \neq c$, and every $y \in Y$ and $b, d \in B$ satisfying $b \neq d$, respectively.
4. For all strings $s, t \in \Sigma^*$ satisfying $s \sim t$ we have $\phi(s) = \phi(t)$.

We will also consider a restriction of this notion to functions of the form

$$\phi : \Sigma^{\leq k} \to \mathbb{C}, \tag{8.38}$$

which we call *admissible* if and only if the same conditions listed above hold for those strings $s$ and $t$ that are sufficiently short so that $\phi$ is defined on the arguments indicated within each condition.

Thus, if $\phi$ is a function that is defined from an actual commuting measurement strategy as

$$\phi((z_1, c_1) \cdots (z_n, c_n)) = \langle u, \Pi^{z_1}_{c_1} \cdots \Pi^{z_n}_{c_n} u \rangle \tag{8.39}$$

for every string $(z_1, c_1) \cdots (z_n, c_n)$, where each $\Pi^z_c$ denotes $P^z_c$ or $Q^z_c$ depending on whether $(z, c) \in \Sigma_A$ or $(z, c) \in \Sigma_B$, respectively, then $\phi$ is necessarily admissible. This is true for functions of both forms $\phi : \Sigma^* \to \mathbb{C}$ and $\phi : \Sigma^{\leq k} \to \mathbb{C}$.

Finally, a positive semidefinite operator

$$R \in \mathrm{Pos}\Big(\mathbb{C}^{\Sigma^{\leq k}}\Big) \tag{8.40}$$

is said to be a *k-th order admissible operator* if there exists an admissible function $\phi : \Sigma^{\leq 2k} \to \mathbb{C}$ such that

$$R(s, t) = \phi(s^{\mathsf{R}} t) \tag{8.41}$$

for every choice of strings $s, t \in \Sigma^{\leq k}$.

Observe that for each positive integer $k$, the condition that a positive semidefinite operator is $k$-th order admissible is an affine linear constraint, as there are a finite number of affine linear constraints imposed by the equation (8.41) and the condition that $\phi$ is admissible. Thus, the optimization over all $k$-th order admissible operators can be represented by a semidefinite program. This is the NPA hierarchy, where different choices of $k$ correspond to different levels of the hierarchy. (Any linear objective function may naturally be considered, but often the objective function reflects the probability to win a given nonlocal game. Alternatively, one can set up a semidefinite program that tests whether a given $M$ agrees with some $k$-th order admissible operator.)

## 8.4 Convergence of the NPA hierarchy

Define $\mathcal{C}_k$ to be the subset of $\mathrm{L}(\mathbb{R}^X \otimes \mathbb{R}^Y, \mathbb{R}^A \otimes \mathbb{R}^B)$ containing all $M$ for which there exists a $k$-th order admissible operator $R \in \mathrm{Pos}(\mathbb{C}^{\Sigma^{\leq k}})$ satisfying (8.31) for every $x \in X$, $y \in Y$, $a \in A$, and $b \in B$. One might call any such $M$ a *$k$-th order pseudo-commuting measurement strategy*.

A moment's thought reveals that

$$\mathcal{C}_1 \supseteq \mathcal{C}_2 \supseteq \mathcal{C}_3 \supseteq \cdots \tag{8.42}$$

as any $(k+1)$-st order admissible operator must yield a $k$-th order admissible operator when its rows and columns corresponding to strings longer than $k$ are deleted. It is also the case that $\mathcal{C} \subseteq \mathcal{C}_k$ for every positive integer $k$; just like in the $k = 1$ case, an actual commuting measurement strategy defines a Gram matrix $R$ that is $k$-th order admissible for every choice of $k$.

The remainder of the lecture is devoted to proving that the sequence (8.42) converges to $\mathcal{C}$ in the sense made precise by the following theorem.

**Theorem 8.4.** *Let $X$, $Y$, $A$, and $B$ be finite and nonempty sets and let $\mathcal{C}$ and $\mathcal{C}_k$, for every positive integer $k$, be as defined above. It is the case that*

$$\mathcal{C} = \bigcap_{k=1}^{\infty} \mathcal{C}_k. \tag{8.43}$$

*Equivalently, for every $M \in \mathrm{L}(\mathbb{R}^X \otimes \mathbb{R}^Y, \mathbb{R}^A \otimes \mathbb{R}^B)$ the following two statements are equivalent:*

1. *$M$ is a commuting measurement strategy.*
2. *$M$ is a $k$-th order pseudo-commuting measurement strategy for every positive integer $k$.*

## The easier implication

We have already observed the implication that statement 1 implies statement 2. In more detail, under the assumption that statement 1 holds, it must be that $M$ is defined by a commuting measurement strategy in which Alice and Bob's projective measurements are described by $\{P_a^x : a \in A\}$ for Alice and $\{Q_b^y : b \in B\}$ for Bob, with all of these projections acting on a Hilbert space $\mathcal{H}$, along with a unit vector $u \in \mathcal{H}$. As before, let $\Pi_c^z$ denote $P_c^z$ if $z \in X$ and $c \in A$, or $Q_c^z$ if $z \in Y$ and $c \in B$. With respect to this notation, one may consider the $k$-th order admissible operator $R$ defined by

$$R(s,t) = \phi(s^{\mathsf{R}}t), \tag{8.44}$$

where the function $\phi$ is defined as

$$\phi\bigl((z_1,c_1)\cdots(z_j,c_j)\bigr) = \bigl\langle u, \Pi_{c_1}^{z_1}\cdots\Pi_{c_j}^{z_j} u\bigr\rangle \tag{8.45}$$

for every string $(z_1,c_1)\cdots(z_j,c_j) \in \Sigma^{\leq 2k}$. A straightforward verification reveals that this operator is consistent with $M$, in the sense of (8.31), and therefore $M$ is a $k$-th order pseudo-commuting measurement strategy.

## A bound on entries of any $k$-th order admissible operator

Before approaching the more challenging implication of Theorem 8.4, which is that statement 2 implies statement 1, we will prove that every entry of a $k$-th order admissible operator is bounded by 1 in absolute value.

More explicitly, if $R \in \operatorname{Pos}\bigl(\mathbb{C}^{\Sigma^{\leq k}}\bigr)$ is $k$-th order admissible, then

$$\bigl|R(s,t)\bigr| \leq 1 \tag{8.46}$$

for every $s,t \in \Sigma^{\leq k}$. To see that this is so, observe first that

$$\bigl|R(s,t)\bigr| \leq \sqrt{R(s,s)}\sqrt{R(t,t)} \tag{8.47}$$

for each $s,t \in \Sigma^*$, which is a consequence of the fact that each $2 \times 2$ principal submatrix

$$\begin{pmatrix} R(s,s) & R(s,t) \\ R(t,s) & R(t,t) \end{pmatrix} \tag{8.48}$$

must be positive semidefinite. Noting that $R(s,s)$ is real and nonnegative, it therefore suffices to prove that

$$R(s,s) \leq 1 \tag{8.49}$$

for every $s \in \Sigma^{\leq k}$.

The bound (8.49) may be proved by induction on the length of $s$. For the base case, one has that $R(\varepsilon,\varepsilon) = 1$. For the general case, one has that for any string $s \in \Sigma^*$ and any choice of $(z,c) \in \Sigma$, it holds that

$$\begin{aligned} R((z,c)s,(z,c)s) \leq \sum_d R((z,d)s,(z,d)s) &= \sum_d \phi(s^{\mathsf{R}}(z,d)(z,d)s) \\ &= \sum_d \phi(s^{\mathsf{R}}(z,d)s) = \phi(s^{\mathsf{R}}s) = R(s,s), \end{aligned} \tag{8.50}$$

where $\phi : \Sigma^{\leq 2k} \to \mathbb{C}$ is the admissible function from which $R$ is defined, and the sums are over all $d \in A$ or $d \in B$, depending on whether $z \in X$ or $z \in Y$, respectively. By the hypothesis of induction the required bound (8.49) follows.

## Entry-wise convergence to an admissible matrix

Now assume statement 2 of Theorem 8.4 holds. For every $k \geq 1$, let

$$R_k \in \operatorname{Pos}\left(\mathbb{C}^{\Sigma^{\leq k}}\right) \tag{8.51}$$

be a $k$-th order admissible operator satisfying

$$M(a,b|x,y) = R_k((x,a),(y,b)) \tag{8.52}$$

for every $x \in X$, $y \in Y$, $a \in A$, and $b \in B$, and let $\phi_k : \Sigma^{\leq 2k} \to \mathbb{C}$ be the admissible function that defines $R_k$.

We will begin with the observation that there exists an infinite matrix

$$R : \Sigma^* \times \Sigma^* \to \mathbb{C} \tag{8.53}$$

having the following properties:

1. Every finite principal submatrix of $R$ is positive semidefinite.
2. $M(a,b|x,y) = R((x,a),(y,b))$ for every $x \in X$, $y \in Y$, $a \in A$, and $b \in B$.
3. There exists an admissible function $\phi : \Sigma^* \to \mathbb{C}$ such that $R(s,t) = \phi(s^{\mathsf{R}}t)$ for all $s,t \in \Sigma^*$.

Note here that we must draw a distinction between an infinite matrix of the form (8.53) and a linear operator, as these concepts are no longer equivalent in infinite dimensions.

Such an infinite matrix $R$ can, in fact, be obtained in a fairly straightforward fashion. Observe first that for every string $s \in \Sigma^*$ and every infinite, strictly increasing sequence of positive integers $(k_1, k_2, k_3, \ldots)$, the sequence

$$(\phi_{k_1}(s),\ \phi_{k_2}(s),\ \phi_{k_3}(s),\ \ldots) \tag{8.54}$$

must have at least one limit point.[3] This is because each value $\phi_k(s)$ agrees with an entry of $R_k$, and the entries of each operator $R_k$ are bounded by 1 in absolute value.

Beginning with the string $s = \varepsilon$ and the sequence $(k_1, k_2, \ldots) = (1, 2, \ldots)$, we consider the function $\phi : \Sigma^* \rightarrow \mathbb{C}$ defined by the following process:

1. Define $\phi(s)$ to be any limit point of the sequence (8.54).
2. Restrict the sequence $(k_1, k_2, \ldots)$ to any infinite subsequence for which (8.54) converges to the chosen limit point, and rename the indices forming this subsequence as $(k_1, k_2, k_3, \ldots)$.
3. Increment $s$ and return to step 1.

Here, when we say "increment $s$," we are referring to any fixed total ordering of $\Sigma^*$ for which $\varepsilon$ is the first string. For example, the strings in $\Sigma^*$ may be ordered according to their length, with strings of equal length being ordered according to the natural "dictionary ordering" induced by a fixed ordering of $\Sigma$ (which is the so-called *lexicographic ordering* of $\Sigma^*$). It must be that any function $\phi$ obtained through this process is admissible, by virtue of the fact that every $\phi_k$ is admissible.

Now we may define

$$R(s,t) = \phi(s^{\mathsf{R}}t). \tag{8.55}$$

The three properties required of $R$ follow, either directly or from the recognition that every finite submatrix of $R$ is equal to the limit of the corresponding submatrices of some convergent subsequence of the sequence

$$(R_1, R_2, R_3, \ldots) \tag{8.56}$$

together with the fact that the positive semidefinite cone is closed.

## Construction of a commuting measurement strategy

We will now make use of a fact concerning countably infinite matrices for which all finite principal submatrices are positive semidefinite—and that is that any such matrix must be the Gram matrix of a countably infinite set of vectors chosen from some Hilbert space. This is not a trivial fact to prove, but we will take it as given. In the case at hand, this implies that there must exist a Hilbert space $\mathcal{K}$ and a collection of vectors

$$\{u_s : s \in \Sigma^*\} \subset \mathcal{K}, \tag{8.57}$$

such that

$$R(s,t) = \langle u_s, u_t \rangle \tag{8.58}$$

[3] When considering the sequence (8.54), we ignore those indices $k$ for which $\phi_k(s)$ is not defined, of which there are at most finitely many.

for every $s, t \in \Sigma^*$. (The inner product is, naturally, the inner product on $\mathcal{K}$.)

Now let us take $\mathcal{H}$ to be the closure of the span of the set $\{u_s : s \in \Sigma^*\}$, which is a Hilbert space having countable dimension (and is therefore a separable Hilbert space). We will define a commuting measurement strategy for Alice and Bob, with $\mathcal{H}$ being the Hilbert space for this strategy. The unit vector associated with this strategy will be $u_\varepsilon \in \mathcal{H}$, which is indeed a unit vector given that

$$\langle u_\varepsilon, u_\varepsilon \rangle = R(\varepsilon, \varepsilon) = 1. \tag{8.59}$$

Next we define a collection of projections on $\mathcal{H}$. For each choice of $x \in X$ and $a \in A$, define $P_a^x$ to be the projection onto the orthogonal complement of the set

$$\big\{u_{(x,c)s} \,:\, c \in A \backslash \{a\},\, s \in \Sigma^*\big\}, \tag{8.60}$$

and along similar lines, for each choice of $y \in Y$ and $b \in B$, define $Q_b^y$ to be the projection onto the orthogonal complement of the set

$$\big\{u_{(y,d)s} \,:\, d \in B \backslash \{b\},\, s \in \Sigma^*\big\}. \tag{8.61}$$

(The orthogonal complement of any collection of vectors is closed, so these are well-defined projection operators.)

In order to verify that the objects just defined induce a valid commuting measurement strategy that agrees with $M$, we will first prove the following useful fact. For all $(x, a) \in X \times A$, $(y, b) \in Y \times B$, and $s \in \Sigma^*$, it is the case that

$$P_a^x u_s = u_{(x,a)s} \quad \text{and} \quad Q_b^y u_s = u_{(y,b)s}. \tag{8.62}$$

Note first that for any choice of $x \in X$, $a, c \in A$ with $a \neq c$, and $s, t \in \Sigma^*$, we have

$$\big\langle u_{(x,a)s}, u_{(x,c)t} \big\rangle = R((x,a)s, (x,c)t) = \phi\big(s^{\mathsf{R}}(x,a)(x,c)t\big) = 0, \tag{8.63}$$

and similarly for any choice of $y \in Y$, $b, d \in B$ with $b \neq d$, and $s, t \in \Sigma^*$, we have

$$\big\langle u_{(y,b)s}, u_{(y,d)t} \big\rangle = 0. \tag{8.64}$$

This implies that

$$P_a^x u_{(x,a)s} = u_{(x,a)s} \quad \text{and} \quad Q_b^y u_{(y,b)s} = u_{(y,b)s}. \tag{8.65}$$

We also see that for any choice of $x \in X$ and $s, t \in \Sigma^*$,

$$\begin{aligned} \sum_{a \in A} \big\langle u_t, u_{(x,a)s} \big\rangle &= \sum_{a \in A} R(t, (x,a)s) \\ &= \sum_{a \in A} \phi(t^{\mathsf{R}}(x,a)s) = \phi(t^{\mathsf{R}}s) = R(t,s) = \langle u_t, u_s \rangle, \end{aligned} \tag{8.66}$$

from which it follows that

$$u_s = \sum_{a\in A} u_{(x,a)s}. \tag{8.67}$$

Similarly, for any choice of $y \in Y$ and $s \in \Sigma^*$,

$$u_s = \sum_{b\in B} u_{(y,b)s}. \tag{8.68}$$

Consequently,

$$P_a^x u_s = \sum_{c\in A} P_a^x u_{(x,c)s} = P_a^x u_{(x,a)s} = u_{(x,a)s}, \tag{8.69}$$

and by similar reasoning

$$Q_b^y u_s = u_{(y,b)s}, \tag{8.70}$$

as claimed.

With the formulas (8.62) in hand, we can verify the required properties of $\mathcal{H}$, $u_\varepsilon$, $\{P_a^x\}$, and $\{Q_b^y\}$. First, see that

$$\begin{aligned} \langle u_s, P_a^x Q_b^y u_t \rangle &= \langle u_s, u_{(x,a)(y,b)t} \rangle = \phi\bigl(s^{\mathsf{R}}(x,a)(y,b)t\bigr) \\ &= \phi\bigl(s^{\mathsf{R}}(y,b)(x,a)t\bigr) = \langle u_s, u_{(y,b)(x,a)t} \rangle = \langle u_s, Q_b^y P_a^x u_t \rangle \end{aligned} \tag{8.71}$$

for every $s, t \in \Sigma^*$. This implies that $P_a^x$ and $Q_b^y$ commute on the span of the vectors $\{u_s : s \in \Sigma^*\}$, and it follows that they commute on all of $\mathcal{H}$ by continuity. Second, for every $x \in X$ and $s \in \Sigma^*$ we find that

$$\sum_{a\in A} P_a^x u_s = \sum_{a\in A} u_{(x,a)s} = u_s, \tag{8.72}$$

from which it follows (again by continuity) that

$$\sum_{a\in A} P_a^x = \mathbb{1}_{\mathcal{H}} \tag{8.73}$$

and similarly

$$\sum_{b\in B} Q_b^y = \mathbb{1}_{\mathcal{H}}. \tag{8.74}$$

To complete the proof, it remains to observe that the strategy represented by the unit vector $u_\varepsilon$ and the projections $\{P_a^x\}$ and $\{Q_b^y\}$ yields the strategy $M$. This is also evident from the formulas (8.62), as one has

$$R((x,a),(y,b)) = \langle u_{(x,a)}, u_{(y,b)} \rangle = \langle u_\varepsilon, P_a^x Q_b^y u_\varepsilon \rangle \tag{8.75}$$

and therefore

$$M(a,b|x,y) = \langle u_\varepsilon, P_a^x Q_b^y u_\varepsilon \rangle \tag{8.76}$$

for every choice of $x \in X$, $y \in Y$, $a \in A$, and $b \in B$.

## Two implications

We will conclude the lecture by briefly observing two facts that follow from Theorem 8.4. The first is that the set $\mathcal{C}$ of commuting measurement strategies is closed, as it is the intersection of the closed sets $\mathcal{C}_1, \mathcal{C}_2, \ldots$ (and, by similar reasoning, it is convex). The second fact is that there is no loss of generality in restricting one's attention to separable Hilbert spaces in the definition of commuting measurement strategies, for this is so for the strategy constructed in the proof.