DRAFT

# Convex optimization in quantum information theory

Instituto de Ciencias Matemáticas (ICMAT), Madrid March 2023

Assoc. Univ.-Prof. Dr. Richard Kueng

# Contents

# 1. Introduction to convex geometry

**Date:** 27 February 2023

**Agenda:**

1 motivation
2 convex optimization
  1.1 convex sets
  1.2 convex functions
  1.3 convex opt.
3 linear programs
4 semidefinite programs

## 1.1 Motivation

Many problems that humans, certain animals and, more recently, machines face can be recast as a constrained optimization problem. These problems can take many shapes and forms, but often have a finite number of variables (or degrees of freedom). We succinctly collect them into an array, e.g. a $D$-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_D)^T \in \mathbb{R}^D$ (1D array) or a (Hermitian) $D \times D$ matrix $\boldsymbol{X} \in \mathbb{H}^D$ (2D array).

The overall objective of an optimization problem is to maximize (or minimize) a real-valued function $f$ in the parameters $\boldsymbol{x}$. This function is called the *objective function*. However, it is often the case that not all parameter configurations are feasible. Additional constraints enforce membership in a subset $A \subseteq \mathsf{V}$ of all possible vectors/matrices. The set $A$ is called the set of feasible solutions, or *feasible set*. We now have all the ingredients in place to introduce a general optimization problem:

general optimization problem

$$\begin{aligned} \underset{\boldsymbol{x} \in \mathsf{V}}{\text{minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in A. \end{aligned} \tag{1.1}$$

Here, $\mathsf{V}$ is a placeholder for a finite-dimensional, real-valued inner product space. We will mostly focus on $\left(\mathbb{R}^D, \langle \cdot, \cdot \rangle\right)$ and $\left(\mathbb{H}^D, (\cdot, \cdot)\right)$.

This framework is very general and encompasses many well-known scientific problems and objectives. What is more, such optimization problems can either be easy or they can be hard. Let us gather some intuition by means of three examples.

**Example 1.1 (linear system of equation).** Fix a matrix $\boldsymbol{A} \in \mathbb{R}^{D' \times D}$, a vector $\boldsymbol{b} \in \mathbb{R}^{D'}$, define the following subset $A = \left\{ \boldsymbol{x} \in \mathbb{R}^D : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \right\}$ and introduce a trivial objective function $f(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{R}^D$. Then, the following optimization problem

example: linear systems

$$\begin{aligned} \underset{\boldsymbol{x} \in \mathsf{V}}{\text{minimize}} \quad & 0 \\ \text{subject to} \quad & \boldsymbol{x} \in A. \end{aligned}$$

is equivalent to solving the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. ■

Linear systems are very important, but not (really) difficult. Standard techniques, like Gaussian elimination, solve a linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ in a number of arithmetic operations that scales cubically in $D$. Such a polynomial scaling is not great, but far from terrible.

**Example 1.2 (training a binary classification model).** Fix a dimension $D$ and a class of functions $\mathcal{F}$ from $\mathbb{R}^D \to \mathbb{R}$. Let $(\boldsymbol{x}_t, y_t)$ with $\boldsymbol{x}_t \in \mathbb{R}^D$ and $y_t \in \{\pm 1\}$ be a collection of $T$ labeled training data points (e.g. pictures of cats and dogs with the correct label). Then, the following optimization problem

example: training for binary classification

$$\begin{aligned} \underset{f : \mathbb{R}^D \to \mathbb{R}}{\text{minimize}} \quad & \sum\nolimits_{t=1}^{T} \left( f(\boldsymbol{x}_t) - y_t \right)^2 \\ \text{subject to} \quad & f \in \mathcal{F} \end{aligned} \tag{1.2}$$

attempts to identify the function $f$ within the function class $\mathcal{F}$ that performs best in the sense that it minimizes the mean squared error (MSE) over the labeled training data. ■

Binary classification is a good introductory example for machine learning and artificial intelligence. There, the overarching vision is to use copious amounts of training data in order to extract (or learn) interesting functional dependencies. This training stage corresponds to a complicated and very high-dimensional optimization problem, see Eq. (1.2). And it is by no means clear how to solve these problems in an optimal and resource-efficient fashion. Indeed, the training stage is very resource-intensive for most state-of-the-art ML models. Fortunately, empirical results highlight that it is not necessary to find the global optimal solution to Eq. (1.2). Local optima, e.g. those identified by stochastic gradient descent, already tend to perform verty well in practice.

**Example 1.3 (quadratic binary optimization (QUBO)).** Let $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ be a symmetric $D \times D$ matrix. Then, the following optimization problem constitutes an (unconstrained) *quadratic binary optimization problem* (QUBO):

example: quadratic binary optimization (QUBO)

$$\begin{aligned} \underset{\boldsymbol{x} \in \mathbb{R}^D}{\text{maximize}} \quad & \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{x} \in \{\pm 1\}^D \end{aligned} \tag{1.3}$$

The name binary stems from the fact that every entry of $\boldsymbol{x}$ can only assume binary values $\pm 1$. Some QUBOs are easy, while others are suspected to be very hard. ■

QUBOs encompass many important problems in computer science, graph theory and combinatorial optimization. Examples include MaxCUT and 3-Sat (which requires additional constraints) and many more.Interestingly, there is also an intimate connection between QUBOs and the Ising problem in quantum many-body physics: interpret $\boldsymbol{x} \in \{\pm 1\}^n$ as a spin configuration and $\boldsymbol{A}$ as a Hamiltonian. This correspondence is the original motivation for adiabatic quantum computation (think D-wave), as well as the Quantum Approximate Optimization Algorithm (QAOA).

QUBO $\approx$ Ising model

## 1.2 Convex optimization

The three motivating examples highlight that optimization problems of the form (1.1) can sometimes be easy (think: linear systems) and can sometimes be hard (think: training and general QUBO). Whether a given optimization problem is easy or hard depends on the underlying structure. Certain desirable features of both the feasible set and the objective function can have a huge impact on the feasibility of the underlying optimization problem. One such structural property is *convexity*. It is defined for both sets and functions and gives rise to entire families of well-behaved optimization problems.

### 1.2.1 Convex sets

Let V be a finite-dimensional, real-valued vector space.

**Definition 1.4 (convex set).** A set $X \subseteq \mathsf{V}$ is a *convex set* if

convex set

$$p\boldsymbol{x} + (1-p)\boldsymbol{y} \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in X \text{ and } p \in [0,1].$$

*In words:* if $A$ contains two points, then it must also contain the line that connects these two points.

Convexity is well-defined for any dimension $D$. Here are a couple of example sets in different dimensions:

examples of convex sets

- the space V itself is a convex set;
- all 2-dimensional polygons are convex sets;
- all 3-dimensional platonic solids are convex sets;
- the $D$-dimensional unit ball $\left\{\boldsymbol{x} \in \mathbb{R}^D : \langle \boldsymbol{x}, \boldsymbol{x} \rangle \leq 1\right\}$ is a convex set;
- the $(D-1)$-dimensional unit sphere $\left\{\boldsymbol{x} \in \mathbb{R}^D : \langle \boldsymbol{x}, \boldsymbol{x} \rangle = 1\right\}$ is not convex;
- general point sets $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathsf{V}$ are not convex, unless $N = 1$ (a single point);
- the union of two unit balls with different origins is never convex.

These examples already hint at some desirable structural properties of convex sets. First and foremost, convex sets are very interconnected: every two points can be connected by a straight line. Moreover, the perimeter cannot contain any dents. These desirable features are preserved by many geometric operations. In particular,

convexity-preserving operations

- *affine shifts:* if $X$ is a convex set, then $\boldsymbol{A}X+\boldsymbol{b}=\{\boldsymbol{Ax}+\boldsymbol{b}: \ \boldsymbol{x}\in X\}$ is also a convex set;
- *intersection:* if $X,Y$ are convex set, then $X\cup Y=\{\boldsymbol{x}\in \mathsf{V}: \ \boldsymbol{x}\in X \text{ and } x\in Y\}$ is also a convex set.
- *convex hull:* if $X\subseteq \mathsf{V}$ is a set (not necessarily convex), then
$$\mathrm{conv}\left(X\right)=\left\{\sum_{i=1}^{N}p_i\boldsymbol{x}_i: \ N\in\mathbb{N}, p_i\geq 0, \sum_{i=1}^{N}p_i=1, \ \boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\in X\right\}$$
is a convex set.

These examples tell us that convexity plays nicely with certain types of operations. But convexity doesn't play nicely with others. For instance, the union of two convex sets $X\cup Y$ is almost never convex.

Convex sets have a well-defined boundary that must not contain any dents. This boundary constitutes of *extreme points* $\boldsymbol{x}\in X$ which have the particular property that they cannot be represented as a probabilistic average of two other points in the set:

extreme points

$$\boldsymbol{x}=t\boldsymbol{y}+(1-t)\boldsymbol{z} \quad \text{with } t\in(0,1) \quad \Rightarrow \quad \boldsymbol{y}=\boldsymbol{z}=\boldsymbol{x}.$$

We conclude this subsection with two basic, but powerful, insights into the geometry of convex sets. We don't have time to provide a proof and refer to standard textbooks instead.

**Fact 1.5** **(Krein-Millman theorem).** Under some mild regularity conditions (boundedness and closedness), every convex set $X\subseteq\mathsf{V}$ is the convex hull of its extreme points. Equivalently, we can decompose every $\boldsymbol{x}\in X$ as

convex sets are convex hull of the boundary (Krein-Millman)

$$\boldsymbol{x}=\sum_{i=1}^{N}p_i\boldsymbol{x}_i^\star,$$

where $p_i\geq 0$, $\sum_{i=1}^{N}p_i=1$ and each $\boldsymbol{x}_i^\star$ is an extreme point of $X$. ■

**Fact 1.6** **(Separating hyperplane theorem).** Let $X\subseteq\mathsf{V}$ be a compact and closed convex set and let $\boldsymbol{y}\in\mathsf{V}$ be outside the set, i.e. $\boldsymbol{y}\notin X$. Then, there exists an affine hyperplane $A_{(\boldsymbol{a},b)}=\{\boldsymbol{x}\in\mathsf{V}: \ \langle\boldsymbol{a},\boldsymbol{x}\rangle=b\}\subseteq\mathsf{V}$ that strictly separates the point from the convex set:

convex sets can be separated with hyperplanes

$$\langle\boldsymbol{a},\boldsymbol{y}\rangle\geq b \quad \text{while} \quad \langle\boldsymbol{a},\boldsymbol{x}\rangle<b \text{ for all } \boldsymbol{x}\in X.$$

■

### 1.2.2 Convex functions

Let $\mathsf{V}$ be a finite-dimensional, real-valued vector space.

**Definition 1.7** Let $X\subseteq\mathsf{V}$ be convex and let $f:X\to\mathbb{R}$ be a continuous function. Then, this function is convex (on the domain $X$) if

convex function

$$f\left(p\boldsymbol{x}+(1-p)\boldsymbol{y}\right)\leq pf(\boldsymbol{x})+(1-p)f(\boldsymbol{y}) \tag{1.4}$$

for all $\boldsymbol{x},\boldsymbol{y}\in X$ and $p\in[0,1]$. *In words:* probabilistic averages of function values exceed function values of the probabilistic average.

The 'reverse' of a convex function is a *concave function*. A function $f : X\mathbb{R}$ is concave if and only if $-f$ is a convex function. Here are a couple of simple examples:

$f$ concave $\Leftrightarrow -f$ convex

- The *exponential function* $\exp : \mathbb{R} \longrightarrow \mathbb{R}$ is convex.
- The *logarithm* $\log : \mathbb{R}_+ \longrightarrow \mathbb{R}$ is concave.
- The *trigonometric functions* $\sin, \cos : \mathbb{R} \longrightarrow \mathbb{R}$ are neither convex nor concave. They can be either if we restrict attention to certain subintervals of the real line.
- The *absolute value function* $x \mapsto |x|$ and the *ReLu function* $f(x) = \max\{0, x\}$ are both convex functions.
- The squared $\ell_2$-norm function $\boldsymbol{x} \mapsto \langle \boldsymbol{x}, \boldsymbol{x} \rangle$ is a convex function from $\mathbb{R}^D$ to $\mathbb{R}_+$.

**Exercise 1.8** Confirm all of these statements by establishing the defining property for convexity/concavity for each of these functions.

Note that the convex functions mentioned above have one thing in common: they have at most one minimum. This is, in fact, a general feature of convex functions that is very desirable for optimization.

**Theorem 1.9 (convex functions don't have local minima).** Let $F : \mathsf{V} \longrightarrow \mathbb{R}$ be a convex function. Then, every local minimum is also a global minimum.

convex functions don't have local minima

*Proof.* Let $\boldsymbol{x}_\star$ be a local minimum of $f$ and assume that there exists another point $\boldsymbol{x}_\sharp$ that achieves a strictly smaller function value (e.g. a global minimum), i.e. $f(\boldsymbol{x}_\sharp) < f(\boldsymbol{x}_\star)$. Set $\boldsymbol{x}(p) = (1-p)\boldsymbol{x}_\star + p\boldsymbol{x}_\sharp$ which traverses a line segment from $\boldsymbol{x}(0) = \boldsymbol{x}_\star$ to $\boldsymbol{x}(1) = \boldsymbol{x}_\sharp$. Then, for every $p > 0$, convexity of $f$ ensures

$$f\left(\boldsymbol{x}(p)\right) = f\left((1-p)\boldsymbol{x}_\star + p\boldsymbol{x}_\sharp\right) \leq (1-p)f(\boldsymbol{x}_\star) + pf(\boldsymbol{x}_\sharp) < f(\boldsymbol{x}_\star). \quad (1.5)$$

This, however, contradicts our assumption that $\boldsymbol{x}_\star$ is a local minimum. Rel. (1.5) tells us that we can further decrease the function value by making an arbitrarily small step away from $\boldsymbol{x}_\star$. The only resolution is to concede that $f(\boldsymbol{x}_\sharp) < f(\boldsymbol{x}_\star)$ cannot be true to begin with. ■

### 1.2.3 Convex optimization

We have now introduced and analyzed the concept of convexity for both sets and functions. In each context, they are responsible for very desirable properties.

**Definition 1.10 (general convex optimization problem).** Let $X \subseteq \mathsf{V}$ be a *convex set* and let $f : X \longrightarrow \mathbb{R}$ be a *concave function*. Then,

general convex optimization

$$\begin{aligned} \underset{\boldsymbol{x} \in \mathsf{V}}{\text{maximize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in X, \end{aligned} \quad (1.6)$$

is a *convex optimization problem*.

This definition combines several desirable properties of an optimization problem. Let us start with the objective function: concavity ensures that every local maximum is also a global maximum. This ensures that iterative optimization techniques, like gradient descent, cannot get stuck in local optima. Next, note that the feasible set is a convex sets. This ensures that the set has a well-defined boundary and plenty of space within to navigate. In particular, the feasible set cannot contain any bottlenecks. Even better: every pair of feasible points is connected by a straight line that is also feasible.

**Example 1.11** The linear system problem from Example 1.1 is a convex optimization problem. Indeed, the trivial function $f(\boldsymbol{x}) = 0$ is both convex and concave and the feasible set $\{\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ is an affine subspace and thus also conves. In contrast, the training problem from Example 1.2 is typically not a convex optimization problem. ■

As a rule of thumb, convex optimization problems tend to be tractable optimization problems. In fact, we have efficient algorithms for several important subclasses of convex optimization.

It is worthwhile, however, to emphasize that this really is just a rule of thumb. It is possible to come up with convex optimization problems that come with complexity-theoretic obstacles. You will see one such example in the problem section below.

**Warning 1.12** Not all convex optimization problems are tractable. ■

## 1.3 Linear programming

We set V to be $\mathbb{R}^D$ endowed with the standard inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\dagger \boldsymbol{y} = \sum_{i=1}^{D} x_i y_i \in \mathbb{R}.$$

This inner product readily allows us to write down a linear function as

linear objective function

$$f(\boldsymbol{x}) = \langle \boldsymbol{c}, \boldsymbol{x} \rangle = \sum_{i=1}^{D} a_i x_i \quad \text{with } \boldsymbol{c} \in \mathbb{R}^D.$$

Every such function is both convex and concave. And, what is more, every linear function from $\mathbb{R}^D$ to $\mathbb{R}$ can be represented in this fashion.

Let us now turn our attention to the feasible set. It comprises $m$ *linear equality constraints*

$m$ linear equality constraints

$$\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle = b_i \quad \text{for } 1 \leq i \leq m$$

with $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{R}^D$ and $b_1, \ldots, b_m \in \mathbb{R}$. Geometrically, each constraint of this form defines a $(D-1)$-dimensional hyperplane

$$H((\boldsymbol{a}_i, b_i)) = \left\{ \boldsymbol{x} \in \mathbb{R}^D : \langle \boldsymbol{a}, \boldsymbol{x} \rangle = b \right\} \subset \mathbb{R}^D.$$

Hyperlpanes are convex sets. Constraint (1.3) demands that every feasible $\boldsymbol{x} \in \mathbb{R}^D$ must be contained in each of these hyperplanes. This is equivalent to demanding

$$\boldsymbol{x} \in H(\boldsymbol{a}_1, b_1) \cap \cdots \cap H(\boldsymbol{a}_m, b_m).$$

Note that this is the intersection of $m$ convex sets (hyperplanes) and therefore also convex. Here is a succinct representation of this intersection space:

$$\begin{aligned} H(\boldsymbol{a}_1, b_1) \cap \cdots \cap H(\boldsymbol{a}_m, b_m) &= \left\{\boldsymbol{x} \in \mathbb{R}^D : \langle \boldsymbol{a}_1, \boldsymbol{x}\rangle = b_1, \ldots, \langle \boldsymbol{a}_m, \boldsymbol{x}\rangle = b_m\right\} \\ &= \left\{\boldsymbol{x} \in \mathbb{R}^D : \begin{pmatrix} \langle \boldsymbol{a}_1, \boldsymbol{x}\rangle \\ \vdots \\ \langle \boldsymbol{a}_m, \boldsymbol{x}\rangle \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}\right\} \\ &= \left\{\boldsymbol{x} \in \mathbb{R}^D : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\right\} \end{aligned} \tag{1.7}$$

with $\boldsymbol{A} = \sum_{i=1}^m \boldsymbol{e}_i \boldsymbol{a}_i^\dagger \in \mathbb{R}^{m\times D}$ and $\boldsymbol{b} = \sum_{i=1}^m b_i \boldsymbol{e}_i$. Here, $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m$ denotes the standard basis of $\mathbb{R}^D$.

Optimizing a linear objective function over a feasible set constrained by linear equality constraints produces a linear optimization problem

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{maximize}} \quad & \langle \boldsymbol{c}, \boldsymbol{x}\rangle \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \end{aligned} \tag{1.8}$$

This is a promising start, but in itself not very exciting yet. To get a linear program, we include one more convex object that has a qualitatively different character. The *nonnegative orthant* in $\mathbb{R}^D$ is defined as

convex cone constraint: nonnegative orthant

$$\mathbb{R}_+^D = \left\{\boldsymbol{x} = (x_1, \ldots, x_D)^\dagger \in \mathbb{R}^D : x_1, \ldots, x_D \geq 0\right\} \subseteq \mathbb{R}^D. \tag{1.9}$$

In words: $\boldsymbol{x} \in \mathbb{R}_+^D$ if and only if every vector coordinate is nonnegative ($x_i \geq 0$). It is easy to check that $\mathbb{R}_+^D$ is a *convex set*. We leave this as a simple exercise. What is more, $\mathbb{R}_+^D$ is actually a convex cone[1].

**Exercise 1.13 (nonnegative orthant).** Show that the nonnegative orthant $\mathbb{R}_+^D$ defined in Eq. (1.9) is both a *convex set* (i.e. $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^D \Rightarrow p\boldsymbol{x} + (1-p)\boldsymbol{y} \in \mathbb{R}_+^D$ for all $p \in [0,1]$) and a *cone* (i.e. $\boldsymbol{x} \in \mathbb{R}^D \Rightarrow \alpha\boldsymbol{x} \in \mathbb{R}_+^D$ for all $\alpha \geq 0$).

Convex cones can be used to define a partial ordering on the underlying space. The nonnegative orthant in $\mathbb{R}^D$ leads to

$$\boldsymbol{x} \geq \boldsymbol{y} \quad \text{if and only if} \quad \boldsymbol{x} - \boldsymbol{y} \in \mathbb{R}_+^D.$$

Note that this ordering relation admits a very simple explanation: $\boldsymbol{x} - \boldsymbol{y} \in \mathbb{R}_+^D$ if and only if $x_i - y_i \geq 0$ for all $1 \leq i \leq D$, or equivalently:

partial order induced by nonnegative orthant

$$\boldsymbol{x} \geq \boldsymbol{y} \quad \text{if and only if} \quad x_i \geq y_i \quad \text{for all } 1 \leq i \leq D.$$

In words: every entry of $\boldsymbol{x}$ is at least as large as the corresponding entry of $\boldsymbol{y}$. We use this observation to succinctly abbreviate the nonnegative orthant condition itself as $\boldsymbol{x} \geq \boldsymbol{0}$, where $\boldsymbol{0} = (0, \ldots, 0)^\dagger$ is the all-zeroes vector.

[1] Think of a pixelated ice cream cone whose tip is in the origin.

A linear program is obtained by intersecting the feasible set in Eq. (1.8) with the nonnegative orthant $\mathbb{R}_+^D = \left\{ \boldsymbol{x} \in \mathbb{R}^D : \boldsymbol{x} \geq \mathbf{0} \right\}$. Since both sets are convex, this intersection is convex as well.

linear program (LP)

**Definition 1.14 (linear program (LP)).** A *linear program* (LP) is a vector-valued optimization problem of the form

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^D}{\text{maximize}} \quad & \langle \boldsymbol{c}, \boldsymbol{x} \rangle \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ & \boldsymbol{x} \geq \mathbf{0}, \end{aligned}$$

where $\boldsymbol{c} \in \mathbb{R}^D$, $\boldsymbol{b} \in \mathbb{R}^M$ (vectors) and $\boldsymbol{A} \in \mathbb{R}^{m\times D}$ (linear map). This is a convex optimization problem with $D$ optimization variables, $m$ equality constraints and one conic constraint.

LPs are very tractable

> **Computational Primitive (Linear programming).** Linear programs admit tractable solutions whose runtime and memory scale (at most) polynomially in $D$ (problem dimension) and $m$ (number of constraints).

You will hear more about different types of solution strategies in other lectures.

## 1.4 Semidefinite programming

We set V to be $\mathbb{H}^D = \left\{ \boldsymbol{X} \in \mathbb{C}^{D\times D} : \boldsymbol{X}^\dagger = \boldsymbol{X} \right\}$ – the space of Hermitian $D \times D$ matrices – endowed with with the trace (or Frobenius) inner product:

$$(\boldsymbol{X}, \boldsymbol{Y}) = \operatorname{tr}(\boldsymbol{X}\boldsymbol{Y}) = \sum_{i,j=1}^D \boldsymbol{X}_{i,j}\boldsymbol{Y}_{j,i}.$$

Note that $\mathbb{H}^D$ is a real-valued vector space with $D^2$ degrees of freedom. The objective function is again a linear function

linear objective function

$$f(\boldsymbol{X}) = (\boldsymbol{C}, \boldsymbol{X}) = \operatorname{tr}(\boldsymbol{C}\boldsymbol{X}) \quad \text{with } \boldsymbol{C} \in \mathbb{H}^D.$$

This linear function is both convex and concave. Let us now turn our attention to the feasible set which is a subset of $\mathbb{H}^D$. The first part is very similar to linear programming and involves *$m$ linear equality constraints*

$m$ linear equality constraints

$$(\boldsymbol{A}_i, \boldsymbol{X}) = b_i \quad \text{for } 1 \leq i \leq m$$

with $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m \in \mathbb{H}^D$ and $b_1, \ldots, b_m \in \mathbb{R}$. Geometrically, each of these constraints defines a hyperplane of dimension $(D^2 - 1)$ (co-dimension 1) in $\mathbb{H}^D$. The intersection of $m$ such hyperplanes again forms an affine intersection space

$$\boldsymbol{X} \in H(\boldsymbol{A}_1, b_1) \cap \cdots \cap H(\boldsymbol{A}_m, b_m).$$

We can succinctly represent this intersection by defining

$$\begin{aligned}\mathcal{A} : \mathbb{H}^D \to& \mathbb{H}^m \\ \boldsymbol{X} \mapsto& \operatorname{diag}\left( (\boldsymbol{A}_1, \boldsymbol{X}), \ldots, (\boldsymbol{A}_m, \boldsymbol{X}) \right),\end{aligned}$$

and $\boldsymbol{B} = \operatorname{diag}(b_1, \ldots, b_m)$. Here, $\operatorname{diag}(b_1, \ldots, b_m)$ is the diagonal matrix with $b_1$ on the first diagonal entry, $b_2$ on the second diagonal entry and so on. So, in summary all affine constraints can be subsumed as

$$\left\{ \boldsymbol{X} \in \mathbb{H}^D : \mathcal{A}(\boldsymbol{X}) = \boldsymbol{B} \right\}, \tag{1.10}$$

in complete analogy to Eq. (1.7) for linear programs. The difference between linear programs (for vectors) and semidefinite programs (for matrices) instead arises from considering a different type of cone and therefore, by extension, a different type of partial ordering. The *cone of positive semidefinite (psd) matrices* is defined as

convex cone constraint: positive-semidefinite (psd) matrices

$$\mathbb{H}_+^D = \left\{ \boldsymbol{X} \in \mathbb{H}^D : \langle \boldsymbol{y}, \boldsymbol{X}\boldsymbol{y} \rangle = \boldsymbol{y}^\dagger \boldsymbol{X} \boldsymbol{y} \geq 0 \text{ for all } \boldsymbol{y} \in \mathbb{C}^D \right\}. \tag{1.11}$$

Equivalently, $\boldsymbol{X} \in \mathbb{H}_+^D$ if and only if every eigenvalue of $\boldsymbol{X}$ is nonnegative. It is easy to check that the cone of psd matrices is a *convex set*, more precisely: a convex cone.

**Exercise 1.15 (cone of psd matrices).** Show that the set of psd matrices $\mathbb{H}_+^D$ defined in Eq. (1.11) is both a *convex set* (i.e. $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{H}_+^D \Rightarrow p\boldsymbol{X} + (1-p)\boldsymbol{Y} \in \mathbb{R}_+^D$ for all $p \in [0,1]$) and a *cone* (i.e. $\boldsymbol{X} \in \mathbb{H}^D \Rightarrow \alpha \boldsymbol{X} \in \mathbb{H}_+^D$ for all $\alpha \geq 0$).

The cone of psd matrices induces the following partial ordering relation among Hermitian $D \times D$ matrices:

partial order induced by psd cone

$$\boldsymbol{X} \geq \boldsymbol{Y} \quad \text{if and only if} \quad \boldsymbol{X} - \boldsymbol{Y} \in \mathbb{H}_+^D.$$

In words: a matrix $\boldsymbol{X}$ is at least as big as another matrix $\boldsymbol{Y}$ if their matrix difference $\boldsymbol{X} - \boldsymbol{Y}$ has exclusively nonnegative eigenvalues. We use this notation convention to succinctly abbreviate the psd condition itself as $\boldsymbol{X} \geq \boldsymbol{O}$, where $\boldsymbol{O} \in \mathbb{H}^D$ is the all-zeroes matrix.

A semidefinite program is obtained by intersecting the feasible set in Eq. (1.10) with the convex cone of psd matrices $\mathbb{H}_+^D = \left\{ \boldsymbol{X} \in \mathbb{H}^D : \boldsymbol{X} \geq \boldsymbol{O} \right\}$. Since both sets are convex, the intersection is convex as well.

**Definition 1.16 (semidefinite program (SDP)).** A *semidefinite program (SDP)* is an matrix-valued optimization problem of the form

semidefinite program (SDP)

$$\begin{aligned} \underset{\boldsymbol{X} \in \mathbb{H}^D}{\text{maximize}} \quad & (\boldsymbol{C}, \boldsymbol{X}) \\ \text{subject to} \quad & \mathcal{A}(\boldsymbol{X}) = \boldsymbol{B}, \\ & \boldsymbol{X} \geq \boldsymbol{O}, \end{aligned}$$

where $\boldsymbol{C} \in \mathbb{H}^D$, $\boldsymbol{B} \in \mathbb{H}^m$ (matrices) and $\mathcal{A} : \mathbb{H}^D \to \mathbb{H}^m$ (linear map). This is a convex optimization problem with $D^2$ optimization variables, $m$ equality constraints and one conic constraint.

**Computational Primitive (semidefinite programming).** Semidefinite programs admit tractable solutions whose runtime and memory scale (at most) polynomially in $D$ (matrix dimension) and $m$ (number of linear constraints).

LPs are kind of tractable

You will hear more about different types of SDP solvers in other lectures. For now, we content ourselves with stating that the memory and runtime demands of SPD solvers grow much faster with problem dimension than LP solvers do. This, in practice, restricts the use of SDPs to $D \lesssim 1000$ even with dedicated hardware.

## 1.5 Problems

**Problem 1.17 (Variance of random variables).** Consider a discrete random variable, i.e. a weighted distribution of real-valued numbers: $X = (p_i, X_i)$ with $X_i \in \mathbb{R}$, $p_i \geq 1$ and $\sum_{i=1}^N p_i = 1$. The first two (uncentered) moments are

$$\mu_1(X) = \mathbb{E}\left[X\right] = \sum\nolimits_{i=1}^N p_i X_i \quad \text{and} \quad \mu_2(X) = \mathbb{E}\left[X^2\right] = \sum\nolimits_{i=1}^N p_i X_i^2.$$

Prove that the *variance* is always a nonnegative number, i.e,

$$\sigma^2(X) = \mu_2(X) - \mu_1(X)^2 \geq 0.$$

**Problem 1.18 (convex functions achieve maximum at boundary of convex sets).** Let $X \subseteq \mathsf{V}$ be a convex set and let $f : X \longrightarrow \mathbb{R}$ be a convex function. Prove the following helpful equality:

$$\max_{\boldsymbol{y} \in X} f(\boldsymbol{y}) = \max_{\boldsymbol{x}:\text{ extreme point of } X} f(\boldsymbol{x}).$$

In words: a convex function achieves its maximum value at an extreme point of the underlying convex set. **Hint:** use the Krein-Millman theorem: every $\boldsymbol{y} \in X$ can be decomposed as $\sum_i p_i \boldsymbol{x}_i$, where each $\boldsymbol{x}_i \in X$ is an extreme point.

**Problem 1.19 (some convex optimization problems are hard).** Fix a matrix $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ and consider the following convex optimization problem over symmetric $D \times D$ matrices:

$$\begin{aligned} \underset{\boldsymbol{X} \in \mathbb{H}^D}{\text{maximize}} \quad & (\boldsymbol{A}, \boldsymbol{X}) = \operatorname{tr}(\boldsymbol{A}\boldsymbol{X}) \\ \text{subject to} \quad & \boldsymbol{X} \in S = \operatorname{conv}\left\{\boldsymbol{s}\boldsymbol{s}^\dagger : \boldsymbol{s} \in \{\pm 1\}^D\right\} \end{aligned}$$

1. Verify that this is a convex optimization problem, i.e. $S \subset \mathbb{H}^D$ is a convex set and $f(\boldsymbol{X}) = (\boldsymbol{A}, \boldsymbol{X})$ is a concave (and convex) function.
2. Show that this convex optimization problem is equivalent (in the sense that it yields the same optimal function value) as the QUBO problem introduced in Eq. (1.3).
3. Conclude that convexity alone is not enough to ensure that the underlying problem can be solved efficiently.

In words: not all convex optimization problems are simple.

**Problem 1.20 (two platonic solids in $D$ dimensions).** Consider $\mathsf{V} = \left(\mathbb{R}^D, \langle \cdot, \cdot \rangle\right)$ and let $\|\boldsymbol{x}\|_{\ell_1} = \sum_{i=1}^D |\boldsymbol{x}_i|$ and $\|\boldsymbol{x}\|_{\ell_\infty}$ denote the $\ell_1$ and $\ell_\infty$-norm respectively. Consider the following two subsets of $\mathbb{R}^D$:

$$B_{\ell_1} = \left\{\boldsymbol{x} \in \mathbb{R}^D : \|\boldsymbol{x}\|_{\ell_1} \leq 1\right\} \quad \text{and} \quad B_{\ell_\infty} = \left\{\boldsymbol{x} \in \mathbb{R}^D : \|\boldsymbol{x}\|_{\ell_\infty} \leq 1\right\}.$$

1. Show that both $B_{\ell_1}$ and $B_{\ell_\infty}$ are convex sets.
2. What do these sets look like in $D = 2$ and $D = 3$ dimensions?
3. Determine all extreme points for $B_{\ell_1}$ and $B_{\ell_\infty}$. How many are there?
4. Show that both convex sets are exactly equal to the convex hull of the extreme points you identified in (3).

Context: convex sets that are described as the convex hull of finitely many extreme points are called *(convex) polytopes*.

# 2. Distinguishing quantum states

**Date:** 28 February 2023

Today, we will show that the difference between classical probability theory and quantum mechanics is a direct analogue of the difference between linear programming (LP) and semidefinite programming (SDP).

**Agenda:**

1. Recap.: LPs & SDPs
2. Probability theory
3. Quantum mechanics
4. Distinguishing distributions:
   - 4.1 maximum likelihood rule (LP)
   - 4.2 Holevo-Helstrom theorem (SDP)

## 2.1 Recapitulation: Linear and semidefinite programming

Let us start by recapitulating these two concepts.

### Linear programming (LP)

We endow the space $\mathbb{R}^D$ with the standard inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{d} x_i y_i$$

and define the non-negative orthant

nonnegative orthant

$$\mathbb{R}^d_+ = \{\boldsymbol{x} \in \mathbb{R}^D : x_i \geq 0,\ 1 \leq i \leq d\}.$$

This induces a partial order on $\mathbb{R}^D$ given by

$$\boldsymbol{x} \geq \boldsymbol{y} \ \Leftrightarrow\ \boldsymbol{x} - \boldsymbol{y} \in \mathbb{R}^d_+ \ \Leftrightarrow\ x_i \geq y_i,\ 1 \leq i \leq d.$$

A *linear program* (LP) is an optimization problem of the following form:

linear program (LP)

$$\begin{aligned} \underset{\boldsymbol{z} \in \mathbb{R}^D}{\text{maximize}} \quad & \langle \boldsymbol{c}, \boldsymbol{z} \rangle \\ \text{subject to} \quad & \langle \boldsymbol{a}_i, \boldsymbol{z} \rangle = b_i \quad 1 \leq i \leq m, \\ & \boldsymbol{z} \geq \boldsymbol{0}. \end{aligned}$$

The vectors $\boldsymbol{c} \in \mathbb{R}^D$, $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \in \mathbb{R}^m$ and numbers $b_1, \ldots, b_m \in \mathbb{R}$ completely specify the problem. Problems of this form can be solved efficiently. Linear programming is a powerful technique from both an analytical and computational point of view.

### Semidefinite programming

We denote the space of $D \times D$ hermitian matrices as $\mathbb{H}^D = \{\boldsymbol{X} \in \mathbb{C}^{D \times D} : \boldsymbol{X}^* = \boldsymbol{X}\}$ and endow it with the Frobenius (or Hilbert-Schmidt) inner product

$$(\boldsymbol{X}, \boldsymbol{Y}) = \operatorname{tr}(\boldsymbol{X}\boldsymbol{Y}).$$

**Remark 2.1** We note that while members of $\mathbb{H}^D$ can have complex entries, $\mathbb{H}^D$ is not closed under multiplication with complex numbers and thus forms a $d^2$-dimensional vector space over the real numbers.

A matrix $\boldsymbol{X} \in \mathbb{H}^D$ is positive semidefinite (psd), if $\langle \boldsymbol{x}, \boldsymbol{X}\boldsymbol{x} \rangle \geq 0$ for all $\boldsymbol{x} \in \mathbb{C}^D$. The set of psd matrices $\mathbb{H}^D_+ \subset \mathbb{H}^D$ forms a convex cone. This cone induces the following partial ordering on $\mathbb{H}^D$:

psd cone

$$\boldsymbol{X} \geq \boldsymbol{Y} \Leftrightarrow \boldsymbol{X} - \boldsymbol{Y} \in \mathbb{H}^D_+.$$

We succinctly write $\boldsymbol{X} \geq \boldsymbol{O}$ to indicate that $\boldsymbol{X} \in \mathbb{H}^D$ is psd.

A *semidefinite program* (SDP) is an optimization program of the following form

semidefinite program (SDP)

$$\begin{aligned} \underset{\boldsymbol{X} \in \mathbb{H}^D}{\text{maximize}} \quad & (\boldsymbol{C}, \boldsymbol{X}) \\ \text{subject to} \quad & (\boldsymbol{A}_i, \boldsymbol{X}) = b_i \quad 1 \leq i \leq m, \\ & \boldsymbol{X} \geq \boldsymbol{O}. \end{aligned}$$

This optimization is completely specified by the matrices $\boldsymbol{C}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_m \in \mathbb{H}^D$ and $m$ numbers $b_1, \ldots, b_m \in \mathbb{R}$.

Like LPs, SDPs are very useful both in theory and practice. We note that LPs and SDPs arose in totally analogous ways from the triples $(\mathbb{R}^D, \langle \cdot, \cdot \rangle, \geq)$ and $(\mathbb{H}^D, (\cdot, \cdot), \geq)$. We will now show that the difference between classical probability theory and quantum mechanics can equally be understood as replacing the former, with the latter triple.

## 2.2 (Discrete) probability theory

Probability theory is modeled by *probability triples* consisting of a sample space (which contains all potential outcomes), a set of events (to which we might want to assign probabilities), and a probability rule (assigning a probability to each and every event). In the setting of *discrete* probability theory, the set of all possible outcomes is finite ($|\Omega| = D$). In this case, we can simply choose the power set of $\Omega$ as the set of events and correspondingly, the probability triple is fully characterized by a *probability density vector* that assigns a probability to each outcome in $\Omega$. Let $\boldsymbol{e} = (1, \ldots, 1)^T$ denote the all-ones vector in $\mathbb{R}^D$

(probability) density vector

**Definition 2.2 (probability density).** A *probability density vector* is a vector

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_D \end{pmatrix} \in \mathbb{R}^D : \quad \boldsymbol{p} \geq \boldsymbol{0}, \quad \langle \boldsymbol{e}, \boldsymbol{p} \rangle = \sum_{i=1}^{D} p_i = 1.$$

Probability theory is concerned with characterizing the likelihood of events or, equivalently, the distribution of *measurement* outcomes.

events/measurements

**Definition 2.3 (measurement).** *Measurements* are resolutions of the identity (vector):

$$\{\boldsymbol{h}_a : \ a \in A\} \subset \mathbb{R}^D : \quad \boldsymbol{h}_a \geq \boldsymbol{0}, \ a \in A \quad \text{and} \quad \sum\nolimits_{a \in A} \boldsymbol{h}_a = \boldsymbol{1}.$$

Here, $A$ is a (finite) set of potential measurement outcomes.

We still need a final ingredient to describe how probability densities (as vectors in $\mathbb{R}^D$) and measurements $\{\boldsymbol{h}_a : \ a \in A\}$ relate to the probability of different measurement outcomes.

probability rule

**Definition 2.4 (probability rule).** For a probability density $\boldsymbol{p} \in \mathbb{R}^D$ and a measurement $\{\boldsymbol{h}_a : a \in A\} \subset \mathbb{R}^D$ define the *probability rule*

$$\Pr\left[a | \boldsymbol{p}\right] = \langle \boldsymbol{h}_a, \boldsymbol{p} \rangle, \quad \text{for all} \quad a \in A.$$

This assigns a probability to each possible outcome $a$ of the measurement.

example: dice rolling

**Example 2.5 (Fair dice roll).** The probability density of a fair dice roll is a flat distribution over 6 potential events: $\boldsymbol{p} = \frac{1}{6}\boldsymbol{1} \in \mathbb{R}^6$. Suppose that we wish to test whether a single dice roll results in either, $\{1,2\}$, $\{3,4\}$, or $\{5,6\}$. This measurement may be associated with the following resolution of identity:

$$h_{\{1,2\}} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad h_{\{3,4\}} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad h_{\{5,6\}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix},$$

The probability rule then readily implies:

$$\Pr\left[\{1,2\} | \boldsymbol{p}\right] = \Pr\left[\{3,4\} | \boldsymbol{p}\right] = \Pr\left[\{5,6\} | \boldsymbol{p}\right] = \frac{1}{3}.$$

■

probability simplex

We introduce the *probability simplex* in $\mathbb{R}^D$,

$$\Delta_{D-1} := \left\{ x \in \mathbb{R}^D : \boldsymbol{x} \geq \boldsymbol{0}, \langle \boldsymbol{e}, \boldsymbol{x} \rangle = 1 \right\},$$

and observe that it equal to the convex hull of the standard basis vectors $\boldsymbol{e}_1 = (1,0,\ldots,0)^T, \ldots, \boldsymbol{e}_D = (0,\ldots,0,1)^T$:

$$\Delta_{D-1} = \operatorname{conv}\left\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\right\}.$$

pure probability distribution

**Definition 2.6 (pure probability distribution).** A probability distribution $\boldsymbol{p} \in \Delta_{D-1}$ is called *pure*, if it is an extreme point of $\Delta_{D-1}$. This is the case if and only if the probability distribution is deterministic.

The essential concepts of classical probability theory are summarized in Table 2.1

| Concept | Explanation | Mathematical formulation |
|---|---|---|
| probability density | normalized, non-negative vectors $\boldsymbol{p} \in \mathbb{R}^D$ | $\boldsymbol{p} \geq \mathbf{0}$, $\langle \mathbf{1}, \boldsymbol{p} \rangle = 1$ |
| measurement | resolution of the identity $\{\boldsymbol{h}_a : a \in A\}$ | $\boldsymbol{h}_a \geq \mathbf{0}$, $\sum_{a \in A} \boldsymbol{h}_a = \mathbf{1}$ |
| probability rule | standard inner product | $\Pr\left[a \vert \boldsymbol{p}\right] = \langle \boldsymbol{h}_a, \boldsymbol{p} \rangle$ |

**Table 2.1** *Axioms for classical probability theory:* The structure of discrete probability theory is captured by the following geometric configuration: $\mathbb{R}^D$ endowed with the partial order $\geq$ and the identity element $\mathbf{1} = (1, \ldots, 1)^T$. This closely resembles *linear programming*.

## 2.3 (Finite-dimensional) quantum mechanics

The postulates of quantum mechanics naturally arise from an extension of classical probability theory. Replace the triple $\left(\mathbb{R}^D, \geq, \mathbf{1}\right)$, by the triple $\left(\mathbb{H}^D, \geq, \mathbf{I}\right)$. Here, $\mathbf{I}$ is the identity matrix, i.e. $[\mathbf{I}]_{i,j} = \delta_{i,j}$. The analogous object to a probability density vector is a (probability) *density matrix*.

(probability) density matrix

**Definition 2.7 (density matrix).** The state of a $D$-dimensional quantum mechanical system is fully described by a *density matrix*

$$\boldsymbol{\rho} \in \mathbb{H}^D : \quad \boldsymbol{\rho} \geq \boldsymbol{O}, \quad (\mathbf{I}, \boldsymbol{\rho}) = \operatorname{tr}(\boldsymbol{\rho}) = 1.$$

In analogy to measurements in classical probability theory, we define a quantum measurement as follows.

measurement

**Definition 2.8 (measurement).** A *measurement* is a resolution of the identity (matrix):

$$\{\boldsymbol{H}_a : a \in A\} : \quad \boldsymbol{H}_a \geq \boldsymbol{O}, \; a \in A, \quad \sum_{a \in A} \boldsymbol{H}_a = \mathbf{I}.$$

Once more, we need a way to describe how density matrices and measurements can be combined to tell us something about quantum measurements and their outcomes. This is captured by the following rule.

Born's (probability) rule

**Definition 2.9 (Born's rule).** For a density matrix $\boldsymbol{\rho} \in \mathbb{H}^D$ and a measurement $\{\boldsymbol{H}_a : a \in A\} \subset \mathbb{H}^D$, we have the following probability rule:

$$\Pr[a \vert \rho] = (\boldsymbol{H}_a, \boldsymbol{\rho}) \quad \text{for } a \in A. \tag{2.1}$$

If a measurement is performed on a quantum mechanical density matrix, two things happen:

1 we obtain a random measurement outcome $a \in A$ that is distributed according to Eq. (2.1) ('god does play dice').
2 the quantum system described by $\boldsymbol{\rho}$ ceases to exist ('wavefunction collapse').

| Concept | Explanation | Mathematical formulation |
|---|---|---|
| Probability density | normalized, psd matrix $\boldsymbol{\rho} \in \mathbb{H}^D$ | $\boldsymbol{\rho} \geq \boldsymbol{O}$, $(\mathbf{I}, \boldsymbol{\rho}) = 1$ |
| measurement | resolution of the identity $\{\boldsymbol{H}_a : a \in A\}$ | $\boldsymbol{H}_a \geq \boldsymbol{O}$, $\sum_{a \in A} \boldsymbol{H}_a = \mathbf{I}$ |
| probability rule | standard inner product | $\Pr[a \vert \boldsymbol{\rho}] = (\boldsymbol{H}_a, \boldsymbol{\rho})$ |

**Table 2.2** *Axioms for quantum mechanics:* The structure of quantum mechanics is captured by the following geometric configuration: $\mathbb{H}^D$ endowed with the psd order $\geq$ and the identity matrix $\mathbf{I}$. This closely resembles *semidefinite programming*.

The fundamental axioms of quantum mechanics are a straightforward generalization of classical probability theory, see Table 2.2. The transition from classical to quantum probability theory resembles a transition from linear programming to semidefinite programming.

example: Stern-Gerlach

**Example 2.10 (Stern-Gerlach experiment).** Fix $D = 2$ (single "spin") and consider the density matrix

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

and two distinct potential measurements:

$$\begin{aligned} \left\{ \boldsymbol{H}_{\pm}^{(z)} \right\} &= \left\{ \frac{1}{2}\mathbf{I} \pm \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \right\} = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}, \\ \left\{ \boldsymbol{H}_{\pm}^{(x)} \right\} &= \left\{ \frac{1}{2}\mathbf{I} \pm \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\} = \left\{ \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \right\}. \end{aligned}$$

The resulting probabilities are then given by

$$\begin{aligned} \Pr[+, (z) \vert \boldsymbol{\rho}] &= \left( \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \right) = 1, \\ \Pr[-, (z) \vert \boldsymbol{\rho}] &= 0, \\ \Pr[+, (x) \vert \boldsymbol{\rho}] &= \left( \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \right) = \frac{1}{2}, \\ \Pr[-, (x) \vert \boldsymbol{\rho}] &= \frac{1}{2}. \end{aligned}$$

This may seem surprising. The state $\boldsymbol{\rho}$ provides completely deterministic measurement outcomes for $\left\{ \boldsymbol{H}_{\pm}^{(z)} \right\}$. Yet, the outcomes for $\left\{ \boldsymbol{H}_{\pm}^{(x)} \right\}$ are completely random. This interesting feature of quantum mechanics is the basis of the famous Stern-Gerlach experiment (1923). ■

The union of all possible quantum states form a convex set in $\mathbb{H}^D$ which we call *quantum state space*:

quantum state space

$$\mathsf{S}\left(\mathbb{H}^D\right) = \left\{\boldsymbol{X} \in \mathbb{H}^D : \ \boldsymbol{X} \geq \boldsymbol{O}, \ (\mathbb{I}, \boldsymbol{X}) = \operatorname{tr}(\boldsymbol{X}) = 1\right\}.$$

This is the quantum analogue of the classical probability simplex.

**Definition 2.11 (pure density matrix).** A density matrix $\boldsymbol{\rho} \in \mathsf{S}(\mathbb{H}^D)$ is called *pure* if it has rank-one, i.e. $\boldsymbol{\rho} = \boldsymbol{x}\boldsymbol{x}^*$ with $\boldsymbol{x} \in \mathbb{C}^d$ normalized to unit Euclidean length.

pure density matrix

Pure quantum states correspond to extreme points of the convex set $\mathsf{S}(\mathbb{H}^D)$ and one can show

$$\mathsf{S}^D = \operatorname{conv}\left\{\boldsymbol{x}\boldsymbol{x}^* : \ \boldsymbol{x} \in \mathbb{C}^d, \ \langle \boldsymbol{x}, \boldsymbol{x}\rangle = 1\right\}.$$

This is the quantum version of the decomposition of the standard simplex into the convex hull of its extreme points: $\Delta_{d-1} = \operatorname{conv}\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$. Classical density vectors are extreme if and only if they are one-sparse, i.e. only one component is different from zero. Quantum density matrices are extreme if and only if they have rank-one. This is the natural matrix generalization of sparsity: a rank-one matrix is one-sparse in its eigenbasis.

In contrast to pure density vectors (classical), pure density matrices (quantum) are not necessarily deterministic. We have encountered this feature in Example 2.10.

## 2.4 Distinguishing quantum and classical probability distributions

In the last two sections we have illustrated the common structure of classical probability theory and quantum mechanics. Extending these parallels, we will now show the optimality of the maximum likelihood rule, and the Holevo-Helstrom theorem. Both address the task of distinguishing two probability densities in the single-shot limit.

### 2.4.1 Classical: the maximum likelihood rule

Suppose that we perfectly know descriptions of two probability distributions $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^D$ and choose to play the following game: a referee chooses one of these distributions uniformly at random and hands it to us. We are allowed to perform a single measurement and – based on its outcome – we must guess which probability distribution was handed to us. We win the game if the guess was correct, otherwise we lose.

challenge: distinguish classical distributions

Let us now try to come up with an optimal guessing strategy. Since we are faced with a binary question, our decision should take the form of a binary measurement:

$$\left\{\boldsymbol{h}_p, \boldsymbol{h}_q\right\} : \ \boldsymbol{h}_q = \mathbf{1} - \boldsymbol{h}_p \quad \text{and} \quad \mathbf{1} \geq \boldsymbol{h}_p \geq \mathbf{0}.$$

A brief computation yields the following probability of guessing the distribution correctly, based on this binary measurement:

$$\begin{aligned}
p_{\text{succ}} =& \frac{1}{2}\Pr\left[p|\boldsymbol{p}\right] + \frac{1}{2}\Pr\left[q|\boldsymbol{q}\right] = \frac{1}{2}\left(\langle \boldsymbol{h}_p, \boldsymbol{p}\rangle + \langle \boldsymbol{h}_q, \boldsymbol{q}\rangle\right)\\
=& \frac{1}{2}\left(\langle \boldsymbol{h}_p, \boldsymbol{p}\rangle + \langle \boldsymbol{e}, \boldsymbol{q}\rangle - \langle \boldsymbol{h}_p, \boldsymbol{q}\rangle\right)\\
=& \frac{1}{2} + \frac{1}{2}\langle \boldsymbol{h}_{\boldsymbol{p}}, \boldsymbol{p} - \boldsymbol{q}\rangle
\end{aligned}$$

We may rewrite the inner-product in the last line as $\sum_{i=1}^{d}\left[\boldsymbol{h}_p\right]_i\left([\boldsymbol{p}]_i - [\boldsymbol{q}]_I\right)$. The factor $1/2$ in front of the expression should not be surprising: we can always achieve a success probability of $1/2$ by mere guessing. Optimizing over measurements $\left\{\boldsymbol{h}_p, \boldsymbol{h}_q\right\}$ allows us to further improve upon this basic strategy. This optimization problem assumes the form of a linear program:

optimal distinguishing strategy (LP)

$$\begin{aligned}
\underset{\boldsymbol{h}_p \in \mathbb{R}^D}{\text{maximize}} \quad & \frac{1}{2} + \frac{1}{2}\langle \boldsymbol{p} - \boldsymbol{q}, \boldsymbol{h}_p\rangle\\
\text{subject to} \quad & \mathbf{1} \geq \boldsymbol{h}_p \geq \mathbf{0}.
\end{aligned}$$

This linear program is simple enough to solve it analytically. The optimal measurement is

Maximum Likelihood rule

$$\left[\boldsymbol{h}_p^{\sharp}\right]_i = \begin{cases} 1, & \text{if } p_i > q_i \\ 0, & \text{else.} \end{cases} \qquad \text{for} \quad 1 \leq i \leq d.$$

The associated guessing strategy is called the *maximum likelihood rule*: upon observing measurement outcome $i$, we guess $\boldsymbol{p}$ if $[\boldsymbol{p}]_i \geq [\boldsymbol{q}]_i$ and otherwise $\boldsymbol{q}$. In words: we choose the distribution that is most likely to provide the outcome that we observed. The associated optimal success probability is

total variational distance

$$p_{\text{succ}}^{\sharp} = \frac{1}{2} + \frac{1}{2}\langle \boldsymbol{h}_{\boldsymbol{p}}^{\sharp}, \boldsymbol{p} - \boldsymbol{q}\rangle = \frac{1}{2} + \frac{1}{4}\sum_{i=1}^{d}|p_i - q_i| = \frac{1}{2} + \frac{1}{4}\|\boldsymbol{p} - \boldsymbol{q}\|_{\ell_1}$$

and the bias – the amount by which we improve over the naive guessing strategy – is proportional to the *total variational distance* $\frac{1}{2}\|\boldsymbol{p} - \boldsymbol{q}\|_{\ell_1}$ of the distributions.

### 2.4.2 Quantum: the Holevo-Helstrom Theorem

Let us now consider the analogous problem in the quantum setting. A referee hands us a black box that contains one of two quantum states: $\boldsymbol{\rho}$ or $\boldsymbol{\sigma}$. Assume that we know the density matrices associated with both states and the referee chooses each of them with equal probability.

Similarly to before, we are allowed to perform a single quantum measurement to guess which state we obtained. Note that this single-shot limit is very appropriate here. A quantum measurement necessarily destroys the quantum state.

Again, we can base our guessing rule on a two-outcome measurement (the question is binary):

$$\boldsymbol{H}_\rho, \boldsymbol{H}_\sigma = \mathbf{I} - \boldsymbol{H}_\rho.$$

If we observe $\rho$, we guess $\boldsymbol{\rho}$, otherwise we guess $\boldsymbol{\sigma}$. In analogy to the last section, we compute the success probability associated with such a guessing strategy:

$$\begin{aligned} p_{\mathrm{succ}} =& \frac{1}{2}\mathrm{Pr}[\boldsymbol{H}_\rho|\boldsymbol{\rho}] + \frac{1}{2}\mathrm{Pr}[\boldsymbol{H}_\sigma|\boldsymbol{\sigma}] = \frac{1}{2}\left(\boldsymbol{H}_\rho, \boldsymbol{\rho}\right) + \frac{1}{2}\left(\boldsymbol{H}_\sigma, \boldsymbol{\sigma}\right) \\ =& \frac{1}{2}\left(\left(\boldsymbol{H}_\rho, \boldsymbol{\rho}\right) + (\mathbf{I}, \boldsymbol{\rho}) - \left(\boldsymbol{H}_\rho, \boldsymbol{\sigma}\right)\right) \\ =& \frac{1}{2} + \frac{1}{2}\left(\boldsymbol{H}_{\boldsymbol{\rho}}, \boldsymbol{\rho} - \boldsymbol{\sigma}\right) \end{aligned}$$

Next, we optimize this expression over all possible choices of measurements:

optimal distinguishing strategy (SDP)

$$\begin{aligned} \underset{\boldsymbol{H}_\rho \in \mathbb{H}^D}{\text{maximize}} \quad & \frac{1}{2} + \frac{1}{2}\left(\boldsymbol{H}_\rho, \boldsymbol{\rho} - \boldsymbol{\sigma}\right) \\ \text{subject to} \quad & \mathbf{I} \succeq \boldsymbol{H}_\rho \succeq \boldsymbol{O}. \end{aligned}$$

Holevo-Helstrom rule

This is a semidefinite program that is simple enough to solve analytically. Apply an eigenvalue decomposition to $\boldsymbol{X} = \boldsymbol{\rho} - \boldsymbol{\sigma} = \sum_{i=1}^d \xi_i \boldsymbol{x}_i \boldsymbol{x}_i^*$. Set $\boldsymbol{P}_+ = \sum_{i=1}^d \mathbb{I}\left\{\xi_i > 0\right\} \boldsymbol{x}_i \boldsymbol{x}_i^*$ and $\boldsymbol{P}_- = \sum_{i=1}^d \mathbb{I}\left\{\xi < 0\right\} \boldsymbol{x}_i \boldsymbol{x}_i^*$. These are orthogonal projectors onto the positive- and negative ranges of $\boldsymbol{X} = \boldsymbol{\rho} - \boldsymbol{\sigma}$. They are the natural generalizations of the maximum likelihood rule to the quantum setting and are called Holevo-Helstrom measurements (or rule). In particular, the choice $\boldsymbol{H}_p^\sharp = \boldsymbol{P}_+$ is optimal and results in the following optimal success probability:

trace distance

$$p_{\mathrm{succ}}^\sharp = \frac{1}{2} + \frac{1}{4}\|\boldsymbol{\rho} - \boldsymbol{\sigma}\|_1$$

Here, $\|\boldsymbol{A}\|_1 = \mathrm{tr}\left(|\boldsymbol{X}|\right)$ quad with $|\boldsymbol{X}| = \sqrt{\boldsymbol{X}^2}$ denotes the nuclear (or trace) norm. It is the natural quantum generalization of the total variational distance.

**Theorem 2.12 (Holevo-Helstrom).** The optimal success probability for distinguishing two quantum states $\boldsymbol{\rho}, \boldsymbol{\sigma} \in \mathbb{H}^D$ with a single measurement is

$$p_{\mathrm{succ}}^\sharp = \frac{1}{2} + \frac{1}{4}\|\boldsymbol{\rho} - \boldsymbol{\sigma}\|_1.$$

The optimal measurement is the projector onto the positive range of $\boldsymbol{\rho} - \boldsymbol{\sigma}$ and depends on the states in question.

Holevo-Helstrom theorem

This observation dates back to Holevo[1] [Hol73] and Helstrom [Hel69] and plays a prominent role in modern quantum information theory. For instance, when estimating density matrices from experimental observations, error bars are typically reported in the nuclear norm.

[1] Alexander Holevo received the *Claude E. Shannon Award* in 2016 for his outstanding contributions to quantum information theory.

## 2.5 Problems

**Problem 2.13 (distinguishing two types of dices).** Consider the following two classical probability distributions for $D = 6$: $\boldsymbol{p} = (0, 1/3, 0, 1/3, 0, 1/3, 0, 1/3)^\dagger \in \mathbb{R}^6$ and $\boldsymbol{q} = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)^\dagger \in \mathbb{R}^6$. These describe a die that only ever yields even numbers and a die that is fair. How would you attempt to distinguish the two possibilities with a single coin toss?

**Problem 2.14 (distinguishing two types of quantum states).** Consider the following two quantum states in a $D$ dimensions: $\boldsymbol{\rho} = \boldsymbol{\psi}\boldsymbol{\psi}^\dagger$ (arbitrary pure state) and $\boldsymbol{\sigma} = \mathbf{I}_D/D$ (maximally mixed state). How would you attempt to distinguish the two possibilities with a single quantum measurement?

**Problem 2.15 (Linear programs for $\ell_1$-norm).** Set $\boldsymbol{e} = (1, \ldots, 1)^T \in \mathbb{R}^D$, fix a vector $\boldsymbol{a} \in \mathbb{R}^D$ and consider the following two linear programs:

$$\begin{aligned} \underset{\boldsymbol{z} \in \mathbb{R}^D}{\text{maximize}} \quad & \langle \boldsymbol{a}, \boldsymbol{z} \rangle, \\ \text{subject to} \quad & \boldsymbol{e} \geq \boldsymbol{z} \geq -\boldsymbol{e} \end{aligned}$$

and

$$\begin{aligned} \underset{\boldsymbol{y} \in \mathbb{R}^D}{\text{minimize}} \quad & \langle \boldsymbol{e}, \boldsymbol{y} \rangle, \\ \text{subject to} \quad & \boldsymbol{y} \geq \boldsymbol{a}, \ \boldsymbol{y} \geq -\boldsymbol{a}. \end{aligned}$$

1 Show that both compute the $\ell_1$-norm of $\boldsymbol{a}$, i.e. $\|\boldsymbol{a}\|_{\ell_1} = \sum_{i=1}^D |[\boldsymbol{a}]_i|$.

2 Let $\|\boldsymbol{a}\|_{\ell_2} = \sqrt{\langle \boldsymbol{a}, \boldsymbol{a} \rangle} = \sqrt{\sum_{i=1}^D [\boldsymbol{a}]_i^2}$ and $\|\boldsymbol{a}\|_{\ell_\infty} = \max_{1 \leq i \leq D} |[\boldsymbol{a}]_i|$ denote the $\ell_2$ and $\ell_\infty$-norms of $\boldsymbol{a}$. Use these LPs to show

$$\|\boldsymbol{a}\|_{\ell_2} \leq \|\boldsymbol{a}\|_{\ell_1} \leq D\|\boldsymbol{a}\|_\infty \quad \text{for all } \boldsymbol{a} \in \mathbb{R}^D.$$

**Problem 2.16 (Simpler LPs for TV distance).** Set $\boldsymbol{e} = (1, \ldots, 1)^T \in \mathbb{R}^D$ and let $\boldsymbol{a} = \boldsymbol{p} - \boldsymbol{q}$ be the difference of two probability vectors ($\boldsymbol{p}_i, \boldsymbol{q}_i \geq \mathbf{0}$ and $\langle \boldsymbol{e}, \boldsymbol{p} \rangle = \langle \boldsymbol{e}, \boldsymbol{q} \rangle = 1$). Show that the following two simplified linear programs compute (two times) the total variational distance $\|\boldsymbol{p} - \boldsymbol{q}\|_{\ell_1}$ between these distributions:

$$\begin{aligned} \underset{\boldsymbol{z} \in \mathbb{R}^D}{\text{maximize}} \quad & 2\langle \boldsymbol{a}, \boldsymbol{z} \rangle, \\ \text{subject to} \quad & \boldsymbol{e} \geq \boldsymbol{z} \geq \mathbf{0} \end{aligned}$$

and

$$\begin{aligned} \underset{\boldsymbol{y} \in \mathbb{R}^D}{\text{minimize}} \quad & \langle \boldsymbol{e}, \boldsymbol{y} \rangle, \\ \text{subject to} \quad & \boldsymbol{y} \geq \boldsymbol{a}. \end{aligned}$$

**Problem 2.17 (SDPs for trace norm and trace distance).** Consider a Hermitian matrix $\boldsymbol{A} \in \mathbb{H}^D$ with eigenvalue decomposition $\boldsymbol{A} = \sum_{i=1}^D \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^\dagger$. The trace (Schatten-1) norm is defined as $\|\boldsymbol{A}\|_1 = \operatorname{tr}(|\boldsymbol{A}|)$, where $|\boldsymbol{A}| = \sqrt{\boldsymbol{A}^2} = \sum_{i=1}^D |\lambda_i| \boldsymbol{a}_i \boldsymbol{a}_i^\dagger$ is the matrix absolute value. In other words, $\|\boldsymbol{A}\|_1 = \sum_{i=1}^D |\lambda_i|$ is the $\ell_1$-norm of the vector of eigenvalues $(\lambda_1, \ldots, \lambda_D) \in \mathbb{R}^D$.

1 Write down a pair of semidefinite programs that compute the trace norm of $\boldsymbol{A}$.
2 Can you simplify these SDPs if you are promised that $\boldsymbol{A} = \boldsymbol{\rho} - \boldsymbol{\sigma}$ is the difference of two density matrices?

**Hint:** take inspiration from Problem 2.15 and Problem 2.16 which are conceptually very similar.

## Lecture bibliography

[Hel69] C. W. Helstrom. “Quantum detection and estimation theory”. In: *J. Statist. Phys.* 1 (1969), pages 231–252. ISSN: 0022-4715. DOI: 10.1007/BF01007479. URL: https://doi.org/10.1007/BF01007479.

[Hol73] A. S. Holevo. “Optimal quantum measurements”. In: *Teoret. Mat. Fiz.* 17 (1973), pages 319–326. ISSN: 0564-6162.

# 3. Distinguishing quantum channels

**Date:** 01 March 2023

## 3.1 Motivation and background

**Agenda:**

1. Background: quantum channels
2. induced trace distance
3. diamond distance
4. SDP for diamond distance

In the last lecture, we have asked ourselves a well-motivated question from quantum information: what is the best strategy to distinguish two (known) quantum states?

Today, we will continue this line of thought by moving from quantum states to *quantum channels* that describe evolutions of quantum mechanical systems. Formally speaking, quantum channels are linear maps $\mathcal{F} : \mathbb{H}^{D} \to \mathbb{H}^{D'}$ that map density matrices to density matrices in a strong sense.

quantum channel

**Definition 3.1 (quantum channel).** A linear map $\mathcal{F} : \mathbb{H}^{D} \longrightarrow \mathbb{H}^{D'}$ is called a *quantum channel if*

1. $\mathcal{F} \otimes \mathcal{I}_{\tilde{D}}(\boldsymbol{Y}) \succeq \mathbf{0}$ for all $\boldsymbol{Y} \in \mathbb{H}^{D} \otimes \mathbb{H}^{\tilde{D}}$ (complete positivity);
2. $(\mathbf{I}, \mathcal{A}(\boldsymbol{X})) = (\mathbf{I}, \boldsymbol{X})$ for all $\boldsymbol{X} \in \mathbb{H}^{D}$ (trace preserving).

Here, $\mathcal{I}(\boldsymbol{X}) = \boldsymbol{X}$ denotes the identity operation (do nothing) on $\mathbb{H}^{\tilde{D}}$.

Quantum channels model physical evolutions in both open and closed systems. In the context of quantum computation/simulation this includes ideal unitary evolutions as well as practical implementation errors.

**Example 3.2 (Some prominent single-qubit channels).** Fix $D = 2$, i.e. a single qubit and consider the four Pauli matrices

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \ \boldsymbol{Y} = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \ \boldsymbol{Z} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \in \mathbb{H}^{2}.$$

These matrices are Hermitian and unitary ($\boldsymbol{W}\boldsymbol{W}^{\dagger} = \boldsymbol{W}^{\dagger}\boldsymbol{W} = \boldsymbol{W}^{2} = \mathbf{I}$). Many prominent single-qubit channels are defined in terms of these Pauli matrices:

1. *identity*: $\mathcal{I}(\boldsymbol{\rho}) = \mathbf{I}\boldsymbol{\rho}\mathbf{I} = \boldsymbol{\rho}$,
2. *bit flip:* $\mathcal{X}(\boldsymbol{\rho}) = \boldsymbol{X}\boldsymbol{\rho}\boldsymbol{X}$,
3. *dephasing:* $\mathcal{Z}(\boldsymbol{\rho}) = \boldsymbol{Z}\boldsymbol{\rho}\boldsymbol{Z}$,
4. *depolarizing:* $\mathcal{D}(\boldsymbol{\rho}) = \frac{1}{4}\left(\mathbf{I}\boldsymbol{\rho}\mathbf{I} + \boldsymbol{X}\boldsymbol{\rho}\boldsymbol{X} + \boldsymbol{Y}\boldsymbol{\rho}\boldsymbol{Y} + \boldsymbol{Z}\boldsymbol{\rho}\boldsymbol{Z}\right) = \frac{(\mathbf{I},\boldsymbol{\rho})}{2}\mathbf{I}$.

These channels are all extreme in the sense that their effect is strong as it can be. We obtain less severe versions of these channels by considering probabilistic mixtures between $\mathcal{F}$ (channel) and $\mathcal{I}$ (do nothing):

$$\mathcal{F}_p(\boldsymbol{\rho}) = (1-p)\mathcal{I}(\boldsymbol{\rho}) + p\mathcal{F}(\boldsymbol{\rho}) \quad \text{with } p \in [0,1].$$

∎

The objective of today's lecture is the *channel distinguishability problem*: let $\mathcal{F}_1, \mathcal{F}_2$ be two known quantum channels. One of them is chosen uniformly at random (with probability $1/2$ each) and handed to you. You get one channel invocation – i.e. choose an input state, feed it into the channel to obtain an output state – and have to decide which channel it is. What is the best you can do?

objective: distinguish two quantum channels

## 3.2 Induced trace distance

One straightforward solution strategy is to reduce the problem of distinguishing channels to the problem of distinguishing quantum states. This can be achieved by fixing an input state (at will) $\boldsymbol{\rho} \in \mathbb{H}^D$ and feed it into the unknown channel:

$$\boldsymbol{\rho}_{\text{out}} = \mathcal{F}_1(\boldsymbol{\rho}) \quad \text{or} \quad \boldsymbol{\rho}_{\text{out}} = \mathcal{F}_2(\boldsymbol{\rho}).$$

This reduces the task at hand to a problem we already know. With a 2-outcome measurement $(\boldsymbol{H}, \mathbf{I} - \boldsymbol{H})$, we obtain

$$\begin{aligned} p_{\text{succ}} =& \frac{1}{2}\Pr\left[\boldsymbol{H}|\mathcal{F}_1(\boldsymbol{\rho})\right] + \frac{1}{2}\Pr\left[\mathbf{I} - \boldsymbol{H}|\mathcal{F}_2(\boldsymbol{\rho})\right] \\ =& \frac{1}{2} + \frac{1}{2}\left(\boldsymbol{H}, \mathcal{F}_1(\boldsymbol{\rho}) - \mathcal{F}_2(\boldsymbol{\rho})\right). \end{aligned}$$

This formula is valid for any measurement $\mathbf{I} \succeq \boldsymbol{H} \succeq \boldsymbol{O}$ and any input state $\boldsymbol{\rho} \succeq \mathbf{0}, (\mathbf{I}, \boldsymbol{\rho}) = 1$. We can take inspiration from the previous lecture and optimize over the measurement $\boldsymbol{H}$ to obtain

$$\beta_{\boldsymbol{\rho}}\left(\mathcal{F}_1, \mathcal{F}_2\right) = \max_{\mathbf{I} \succeq \boldsymbol{H} \succeq \mathbf{0}} \left(\boldsymbol{H}, \mathcal{F}_1(\boldsymbol{\rho}) - \mathcal{F}_2(\boldsymbol{\rho})\right) = \frac{1}{2}\left\|\mathcal{F}_1(\boldsymbol{\rho}) - \mathcal{F}_2(\boldsymbol{\rho})\right\|_1.$$

Doing so produces the trace distance between the two output states. But now, more is possible. We can also optimize the input state:

$$\beta\left(\mathcal{F}_1, \mathcal{F}_1\right) = \frac{1}{2} \max_{\boldsymbol{\rho} \succeq \mathbf{0}, (\mathbf{I}, \boldsymbol{\rho}) = 1} \left\|\mathcal{F}_1(\boldsymbol{\rho}) - \mathcal{F}_2(\boldsymbol{\rho})\right\|_1.$$

This is now a distance measure (norm) between the channel $\mathcal{F}_1$ and the channel $\mathcal{F}_2$. We delineate it explicitly by writing

$$\|\mathcal{F}_1 - \mathcal{F}_2\|_{1\to 1} := 2\beta\left(\mathcal{F}_1, \mathcal{F}_2\right) = \max_{\boldsymbol{\rho}\geq \mathbf{0}, (\mathbf{I},\boldsymbol{\rho})=1} \|\left(\mathcal{F}_1 - \mathcal{F}_2\right)(\boldsymbol{\rho})\|_1 . \tag{3.1}$$

Viewed as a norm, this distance measure is called the *induced trace norm*. We can use insights from convex geometry to streamline this optimization problem somewhat.

**Lemma 3.3 (induced trace distance).** Let $\mathcal{F}_1, \mathcal{F}_2 : \mathbb{H}^D \to \mathbb{H}^{D'}$ be two quantum channels. Then, the *induced trace distance* from Eq. (3.1) is equivalent to the following optimization problem over pure input states:

induced trace distance

$$\|\mathcal{F}_1 - \mathcal{F}_2\|_{1\to 1} = \max_{\boldsymbol{u}\in\mathbb{C}^D, \langle \boldsymbol{u},\boldsymbol{u}\rangle=1} \left\|\mathcal{F}_1\left(\boldsymbol{u}\boldsymbol{u}^\dagger\right) - \mathcal{F}_2\left(\boldsymbol{u}\boldsymbol{u}^\dagger\right)\right\|_1 .$$

*Proof.* Let $\boldsymbol{\rho}_\sharp$ be the density matrix that achieves the optimal value on the left hand side. Apply an eigenvalue decomposition to decompose it as $\boldsymbol{\rho}_\sharp = \sum_{i=1}^D p_i \boldsymbol{u}_i\boldsymbol{u}_i^\dagger$ with $p_i \geq 0$ and $\sum_{i=1}^D p_i = 1$. Now, note that the function $f(\boldsymbol{\rho}) = \|(\mathcal{F}_1 - \mathcal{F}_2)(\boldsymbol{\rho})\|_1$ is a convex function on the space of all density matrices. Optimality of $\boldsymbol{\rho}_\sharp$ and the definition of convexity now ensure

$$\begin{aligned}\max_{\boldsymbol{\rho}\geq \mathbf{0},(\boldsymbol{I},\boldsymbol{\rho})=1} f(\boldsymbol{\rho}) =& f\left(\boldsymbol{\rho}_\sharp\right) = f\left(\sum_{i=1}^D p_i \boldsymbol{u}_i\boldsymbol{u}_i^\dagger\right) \\ &\leq \sum_{i=1}^D p_i f\left(\boldsymbol{u}_i\boldsymbol{u}_i^\dagger\right) \leq \max_{1\leq i\leq D} f\left(\boldsymbol{u}_i\boldsymbol{u}_i^\dagger\right) \\ &\leq \max_{\boldsymbol{u}\in\mathbb{C}^D,\langle \boldsymbol{u},\boldsymbol{u}\rangle=1} f\left(\boldsymbol{u}\boldsymbol{u}^\dagger\right).\end{aligned}$$

An inequality in the converse direction readily follows from noting that $\boldsymbol{\rho} = \boldsymbol{u}\boldsymbol{u}^\dagger$ is a strict subset of all possible density matrices. ■

## 3.3 Diamond distance

The induced trace distance is the product of a good candidate solution for distinguishing two quantum channels. But, it is not the end of the story. We can do even more by taking the laws of quantum information theory seriously. Entanglement, in particular, allows us to strongly correlate our input state with an additional quantum system that does not participate in the channel evolution.

entanglement-assisted channel distinguishability

More formally, this is achieved by considering an input state $\boldsymbol{\rho} \in \mathbb{H}^D \otimes \mathbb{H}^{\tilde{D}} \simeq \mathbb{H}^{D\tilde{D}}$, where $\tilde{D} \geq 0$ is the dimension of the auxiliar system. We then apply the unknown channel to the first $D$-level system while leaving the second $\tilde{D}$-level system untouched. Note that the dimension $\tilde{D}$ of the auxiliar system is now also a parameter that can be optimized. Optimizing this parameter, as well as the input state $\boldsymbol{\rho}$ and a 2-outcome measurement at the very end produces a distance measure that really reflects the best quantum strategy conceivable for distinguishing two known quantum channels. This distance measure (norm) is called the *diamond distance* (or completely bounded trace distance):

$$\|\mathcal{F}_1-\mathcal{F}_2\|_\diamond = \sup_{\tilde{D}\in\mathbb{N}} \max_{\boldsymbol{u}\in\mathbb{C}^{\tilde{D}},\langle \boldsymbol{u},\boldsymbol{u}\rangle=1} \left\|(\mathcal{F}_1-\mathcal{F}_2)\otimes\mathcal{I}_{\tilde{D}}\left(\boldsymbol{u}\boldsymbol{u}^*\right)\right\|_1 \tag{3.2}$$

Here, we have re-used our observation from Lemma 3.3 to restrict our optimization over input states to exclusively pure states. This formula looks daunting – especially the supremum over all possible auxiliar dimension $\tilde{D}$ should command respect. Fortunately, it is possible to show that we do not really need to consider all possible auxiliar dimensions $\tilde{D}\in\mathbb{N}$.

**Fact 3.4** The supremum over $\tilde{D}\in\mathbb{N}$ in Eq. (3.2) is achieved at $\tilde{D}=D$. ▪

We refer to Ref. [Wat11] for an elegant and rigorous proof. Instead, we capitalize on this observation to deliver a formal and (somewhat) readable definition of the diamond distance.

diamond distance

**Definition 3.5 (diamond distance).** Let $\mathcal{F}_1,\mathcal{F}_2:\mathbb{H}^D\to\mathbb{H}^D$ be two quantum channels. Then, their *diamond distance* is defined as

$$\|\mathcal{F}_1-\mathcal{F}_2\|_\diamond = \max_{\boldsymbol{u}\in\mathbb{C}^D\otimes\mathbb{C}^D,\langle \boldsymbol{u},\boldsymbol{u}\rangle=1} \|(\mathcal{F}_1-\mathcal{F}_2)\otimes\mathcal{I}_D\left(\boldsymbol{u}\boldsymbol{u}^*\right)\|_1 . \tag{3.3}$$

More generally, we can define the diamond norm for any type of map $\Phi:\mathbb{H}^D\to\mathbb{H}^{D'}$, not only channel differences $\Phi=\mathcal{F}_1-\mathcal{F}_2$. The diamond norm has several appealing properties:

diamond norm is sub-multiplicative

1 The diamond norm is *sub-multiplicative under the composition of maps*:

$$\|\Phi_1\circ\Phi_2\|_\diamond \leq \|\Phi_1\|_\diamond\|\Phi_2\|_\diamond . \tag{3.4}$$

2 The diamond norm is *sub-multiplicative under taking tensor products of maps*:

$$\|\Phi_1\otimes\Phi_2\|_\diamond \leq \|\Phi_1\|_\diamond\|\Phi_2\|_\diamond . \tag{3.5}$$

**Exercise 3.6** Prove Rel. (3.4) and Rel. (3.5).

These two properties are important and very desirable. The strong optimization over input states and output measurements also endow the diamond distance with a 'worst case' character. It is one of the largest (and most pessimistic) distance measures conceivable, because it uses every trick in the quantum information toolbox. This worst-case character has awarded the diamond distance a prominent role in the analysis of error channels and how their effect propagates. Understanding this is central for quantum error correction and fault tolerance. Virtually all rigorous threshold theorems that exist do require gate errors that are bounded in diamond distance.

## 3.4 SDP for diamond distance

We will now show that it is possible to compute the diamond distance between two channels $\mathcal{F}_1,\mathcal{F}_2:\mathbb{H}^D\to\mathbb{H}^{D'}$ by evaluating a semidefinite program (SDP).

As a first step, we need a convenient representation of the channels that is compatible with the SDP paradigm (linear objective function and constraints in matrix space). The *Choi matrix* is one such channel representation. It corresponds to the quantum state that arises from preparing a (pure) maximally entangled state on $\mathbb{H}^D \otimes \mathbb{H}^D$ and inputting one half of it to the channel $\mathcal{F}_i$ while leaving the other half unchanged (do nothing). More formally, let

$$\boldsymbol{\Omega} = \boldsymbol{\omega}\boldsymbol{\omega}^\dagger \in \mathbb{H}^D \otimes \mathbb{H}^D \quad \text{with} \quad \boldsymbol{\omega} = \frac{1}{\sqrt{D}} \sum_{i=1}^{D} \boldsymbol{e}_i \otimes \boldsymbol{e}_i \in \mathbb{C}^D \otimes \mathbb{C}^D$$

be a maximally entangled (Bell) state. We then define the *Choi matrix* of channel $\mathcal{F}_i : \mathbb{H}^D \to \mathbb{H}^{D'}$ as

Choi matrix

$$\boldsymbol{J}(\mathcal{F}_i) = \mathcal{F}_i \otimes \mathcal{I}_D(\Omega) \in \mathbb{H}^{D'} \otimes \mathbb{H}^D \simeq \mathbb{H}^{DD'}. \tag{3.6}$$

Choi matrices enjoy a prominent role in the study of quantum channels. This is largely due to the following fact.

**Fact 3.7 (Choi-Jamiolkowski isomorphism).** The Choi matrix (3.6) establishes a linear one-to-one correspondence (isomorphism) between Hermicity preserving maps $\mathcal{F} : \mathbb{H}^D \otimes \mathbb{H}^{D'}$ and Hermitian matrices in $\mathbb{H}^{D'} \otimes \mathbb{H}^D$. Moreover, a map $\mathcal{F}$ is a quantum channel if and only if its Choi matrix $\boldsymbol{J}(\mathcal{F})$ is a quantum state. ■

We refer to standard quantum information sources, like Watrous' book, for a more detailed context and proof. For us it suffices to note that the Choi matrix is linear. And, therefore, we can readily extend this context to differences of quantum channels:

$$\boldsymbol{J}(\mathcal{F}_1 - \mathcal{F}_2) = \boldsymbol{J}(\mathcal{F}_1) - \boldsymbol{J}(\mathcal{F}_2) \in \mathbb{H}^{D'} \otimes \mathbb{H}^D.$$

This representation allows us to state the main result of today's lecture.

**Theorem 3.8 (SDP for diamond distance).** Let $\mathcal{F}_1, \mathcal{F}_2 : \mathbb{H}^D \to \mathbb{H}^{D'}$ be two quantum channels. Then, the following SDP only depends on the Choi matrix $\boldsymbol{J}(\mathcal{F}_1 - \mathcal{F}_2) \in \mathbb{H}^{D'} \otimes \mathbb{H}^D$ and computes the diamond distance $\|\mathcal{F}_1 - \mathcal{F}_2\|_\diamond$:

SDP for diamond distance

$$\begin{aligned} \underset{\boldsymbol{W} \in \mathbb{H}^{D'} \otimes \mathbb{H}^D, \boldsymbol{\rho} \in \mathbb{H}^D}{\text{maximize}} \quad & 2D\left(\boldsymbol{J}(\mathcal{F}_1 - \mathcal{F}_2), \boldsymbol{W}\right) \\ \text{subject to} \quad & \mathbf{I}_{D'} \otimes \boldsymbol{\rho} \succeq \boldsymbol{W}, \operatorname{tr}(\boldsymbol{\rho}) = 1, \\ & \boldsymbol{W}, \boldsymbol{\rho} \succeq \mathbf{0}. \end{aligned} \tag{3.7}$$

*Proof.* We start with the definition of the diamond distance from Definition 3.5:

$$\|\mathcal{F}_1 - \mathcal{F}_2\|_\diamond = \max_{\boldsymbol{u} \in \mathbb{C}^D \otimes \mathbb{C}^D, \langle \boldsymbol{u}, \boldsymbol{u} \rangle = 1} \|(\mathcal{F}_1 - \mathcal{F}_2) \otimes \mathcal{I}_D(\boldsymbol{u}\boldsymbol{u}^*)\|_1. \tag{3.8}$$

Next, we use the following mathematical trick to rewrite any pure bipartite input state:

$$\boldsymbol{u} = \sqrt{D}\left(\mathbf{I}_D \otimes \boldsymbol{B}\right)\boldsymbol{\omega} \quad \text{with} \quad \boldsymbol{B} \in \mathbb{C}^{D \times D}, (\boldsymbol{B}, \boldsymbol{B}) = 1.$$

To see why this is true, note that $\boldsymbol{B}\boldsymbol{e}_j = \sum_{i=1}^{D} \boldsymbol{B}_{i,j}\boldsymbol{e}_j$ and, therefore

$$\begin{aligned}\sqrt{D}\left(\mathrm{I}_D \otimes \boldsymbol{B}\right)\boldsymbol{\omega} &= \sum_{j=1}^{D} (\mathrm{I}_D \boldsymbol{e}_j) \otimes (\boldsymbol{B}\boldsymbol{e}_j) \\ &= \sum_{j=1}^{D} \boldsymbol{e}_j \otimes \left(\sum_{i=1}^{D} \boldsymbol{B}_{i,j}\boldsymbol{e}_i\right) \\ &= \sum_{i,j=1}^{D} \boldsymbol{B}_{i,j}\boldsymbol{e}_j \otimes \boldsymbol{e}_i.\end{aligned}$$

This general formula is expressive enough to represent any pure bipartite state $\boldsymbol{u} \in \mathbb{C}^D \otimes \mathbb{C}^D$. Moreover, the normalization constraint translates to

$$1 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle = \sum_{i,j=1}^{D} \left|\boldsymbol{B}_{j,i}\right|^2 = (\boldsymbol{B}, \boldsymbol{B}).$$

We can now use this reparametrization of $\boldsymbol{u}$ (and also $\boldsymbol{u}^\dagger$) to rewrite Eq. (3.8) as

$$\begin{aligned}\|\mathscr{F}_1 - \mathscr{F}_2\|_\diamond &= \max_{\boldsymbol{B} \in \mathbb{C}^{D\times D}, \|\boldsymbol{B}\|_2 = 1} \left\|(\mathscr{F}_1 - \mathscr{F}_2) \otimes \mathcal{I}_D \left(\sqrt{D}(\mathrm{I}_D \otimes \boldsymbol{B})\boldsymbol{\omega}\boldsymbol{\omega}^*(\mathrm{I}_D \otimes \boldsymbol{B}^\dagger)\sqrt{D}\right)\right\|_1 \\ &= \max_{\boldsymbol{B} \in \mathbb{C}^{D\times D}, \|\boldsymbol{B}\|_2 = 1} D \left\|(\mathscr{F}_1 - \mathscr{F}_2) \otimes \mathcal{I}_D \left((\mathrm{I}_D \otimes \boldsymbol{B})\boldsymbol{\Omega}(\mathrm{I}_D \otimes \boldsymbol{B}^\dagger)\right)\right\|_1 \\ &= \max_{\boldsymbol{B} \in \mathbb{C}^{D\times D}, \|\boldsymbol{B}\|_2 = 1} D \left\|(\mathrm{I}_D \otimes \boldsymbol{B})\,(\mathscr{F}_1 - \mathscr{F}_2) \otimes \mathcal{I}_D\,(\boldsymbol{\Omega})\,(\mathrm{I}_D \otimes \boldsymbol{B}^\dagger)\right\|_1 \\ &= \max_{\boldsymbol{B} \in \mathbb{C}^{D\times D}, \|\boldsymbol{B}\|_2 = 1} D \left\|(\mathrm{I}_D \otimes \boldsymbol{B})\boldsymbol{J}\,(\mathscr{F}_1 - \mathscr{F}_2)\,(\mathrm{I}_D \otimes \boldsymbol{B}^\dagger)\right\|_1\end{aligned}$$

Here, we have used the critical fact that $(\mathscr{F}_1 - \mathscr{F}_2) \otimes \mathrm{I}_D$ acts like the identity on the second tensor factor. This has allowed us to commute $(\mathrm{I}_D \otimes \boldsymbol{B})$ and $(\mathrm{I}_D \otimes \boldsymbol{B}^\dagger)$ through the channel action and recognize the Choi matrix $\boldsymbol{J}(\mathscr{F}_1 - \mathscr{F}_2)$ at the center.

Next, we note that this Choi matrix is Hermitian and also traceless and the adjungation with $\mathrm{I}_D \otimes \boldsymbol{B}$ does not change that. This allows us to use our insights from last lecture (Helstrom's theorem) to express the trace distance as a maximization over 2-outcome measurements $(\boldsymbol{K}, \mathrm{I}_{D'} \otimes \mathrm{I}_D - \boldsymbol{K})$ on the space $\mathbb{H}^{D'} \otimes \mathbb{H}^D$. This produces the following optimization problem for $\|\mathscr{F}_1 - \mathscr{F}_2\|_\diamond$:

$$\begin{aligned}\underset{\boldsymbol{K} \in \mathbb{H}^{D'} \otimes \mathbb{H}^D, \boldsymbol{B} \in \mathbb{C}^{D\times D}}{\text{maximize}} \quad & 2D\,\left((\mathrm{I} \otimes \boldsymbol{B}^\dagger)\boldsymbol{K}(\mathrm{I} \otimes \boldsymbol{B}), \boldsymbol{J}\,(\mathscr{F}_1 - \mathscr{F}_2)\right) \\ \text{subject to} \quad & \mathrm{I}_{D'} \otimes \mathrm{I}_D \succeq \boldsymbol{K} \succeq \boldsymbol{0},\ (\boldsymbol{B}, \boldsymbol{B}) = 1,\end{aligned}$$

This now almost looks like a SDP, but the constraint on $\boldsymbol{B}$ is not quite convex yet. We can resolve this final issue with a bit of reformulation and pattern recognition. Introduce

$$\boldsymbol{W} = \left(\mathrm{I}_{D'} \otimes \boldsymbol{B}^\dagger\right)\boldsymbol{K}\,(\mathrm{I}_{D'} \otimes \boldsymbol{B}) \in \mathbb{H}^{D'} \otimes \mathbb{H}^D$$

and note that $\mathrm{I}_{D'} \otimes \mathrm{I}_D \succeq \boldsymbol{K} \succeq \boldsymbol{0}$ is equivalent to demanding $\mathrm{I}_{D'} \otimes \boldsymbol{B}^\dagger\boldsymbol{B} \succeq \boldsymbol{W} \succeq \boldsymbol{0}$. Finally, we set $\boldsymbol{\rho} = \boldsymbol{B}^\dagger\boldsymbol{B}$ which must obey $\boldsymbol{\rho} \succeq \boldsymbol{O}$ (think: Cholesky decomposition) and the normalization constraint of $\boldsymbol{B}$ translates to

$$\mathrm{tr}(\boldsymbol{\rho}) = \mathrm{tr}\left(\boldsymbol{B}^\dagger\boldsymbol{B}\right) = (\boldsymbol{B}, \boldsymbol{B}) = 1.$$

This is now a nice linear constraint and we are done. ■

## 3.5 Problems

**Problem 3.9** **(Sub-multiplicativity of the diamond norm).** Let $\Phi_1 : \mathbb{H}^D \to \mathbb{H}^{D'}$ and $\Phi_2 : \mathbb{H}^{D'} \to \mathbb{H}^{D''}$ be two linear maps (e.g. two channel differences). Show that the diamond norm (3.3) obeys

$$\|\Phi_1 \circ \Phi_2\|_\diamond \leq \|\Phi_1\|_\diamond \|\Phi_2\|_\diamond \quad \text{and} \quad \|\Phi_1 \otimes \Phi_2\|_\diamond \leq \|\Phi_1\|_\diamond \|\Phi_2\|_\diamond .$$

**Problem 3.10** **(diamond distance for single-qubit Pauli channels).** Fix $D = 2$ and consider the single-qubit Pauli channel $\mathcal{W}(\boldsymbol{\rho}) = \boldsymbol{W}\boldsymbol{\rho}\boldsymbol{W}$ with $\boldsymbol{W} \in \{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}\}$.

1 Compute the diamond distance between $\mathcal{W}$ and $\mathcal{I}$ (do nothing).
2 Fix $p \in [0,1]$ and set $\mathcal{W}_p(\boldsymbol{\rho}) = (1-p)\mathcal{I}(\boldsymbol{\rho}) + p\mathcal{W}(\boldsymbol{\rho})$. Compute the diamond distance between $\mathcal{W}_p$ and $\mathcal{I}$ (do nothing).

**Problem 3.11** **(diamond distance between identity and Hadamard).** Consider the following unitary single-qubit channels: $\mathcal{I}(\boldsymbol{\rho}) = \mathbf{I}\boldsymbol{\rho}\mathbf{I}$ and $\mathcal{H}(\boldsymbol{\rho}) = \boldsymbol{H}\boldsymbol{\rho}\boldsymbol{H}$, where

$$\boldsymbol{H} = \boldsymbol{H}^\dagger = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \in \mathbb{H}^2.$$

Show (by whatever means you see fit) that $\|\mathcal{I} - \mathcal{H}\|_\diamond = 2$, i.e. the diamond distance between these two channels is maximal.

**Problem 3.12** **(diamond distance between identity and depolarizing channel).** Fix $D \geq 2$ and let $\mathcal{I}, \mathcal{D}_p : \mathbb{H}^D \to \mathbb{H}^D$ be identity and depolarizing channel, i.e. $\mathcal{I}(\boldsymbol{\rho}) = \boldsymbol{\rho}$ and $\mathcal{D}_p(\boldsymbol{\rho}) = (1-p)\boldsymbol{\rho} + p\frac{(\mathbf{I},\boldsymbol{\rho})}{D}\mathbf{I}$ for some $p \in [0,1]$. Show that

$$\|\mathcal{I} - \mathcal{D}\|_\diamond \geq 2p\left(1 - \frac{1}{D^2}\right).$$

Do you think that this lower bound on the diamond distance is optimal?

**Problem 3.13** **(More general diamond norm SDP).** The SDP from Theorem 3.8 only applies to linear maps $\Phi : \mathbb{H}^D \to \mathbb{H}^{D'}$ that are trace-annihilating ($\mathrm{tr}(\Phi(\boldsymbol{X})) = 0$ for all $\boldsymbol{X}$), e.g. differences between two quantum channels. Modify the diamond distance SDP such that it applies to general Hermicity-preserving maps.

## Lecture bibliography

[Wat11] J. Watrous. *Theory of Quantum Information (lecture notes)*. 2011.

# 4. Analytic formulas for certain diamond distances

**Date:** 28 February 2023

**Agenda:**

1. motivation: SDPs as proof techniques
2. setting: diamond distance of certain channels
3. implications for fault-tolerant quantum computation
4. synopsis

## 4.1 Motivation: SDPs as proof technique

Today we will use the complete structure of semidefinite programs and use it as a mathematical proof technique. We will show by means of an example on how to use SDPs for structure recognition.

The starting point is a more thorough treatment of SDPs and their structural properties. We refer to standard references for a thorough treatment. Formally, a SDP is specified by a triple $(\mathscr{A}, \boldsymbol{C}\boldsymbol{B})$ where $\mathscr{A} : \mathbb{H}^D \to \mathbb{H}^m$ is a linear map and $\boldsymbol{C} \in \mathbb{H}^D$, as well as $\boldsymbol{B} \in \mathbb{H}^m$ are matrices. A *primal SDP in standard form* then corresponds to

$$\begin{aligned} \underset{\boldsymbol{Z} \in \mathbb{H}^D}{\text{maximize}} \quad & (\boldsymbol{C}, \boldsymbol{Z}) \qquad \text{(primal SDP)} \\ \text{subject to} \quad & \mathscr{A}(\boldsymbol{Z}) = \boldsymbol{B}, \\ & \boldsymbol{Z} \succeq \boldsymbol{0} \end{aligned}$$

primal-dual SDP pair in standard form

The associated *dual problem in standard form* is

$$\begin{aligned} \underset{\boldsymbol{Y} \in \mathbb{H}^m}{\text{minimize}} \quad & (\boldsymbol{Y}, \boldsymbol{B}) \qquad \text{(dual SDP)} \\ \text{subject to} \quad & \mathscr{A}^* (\boldsymbol{Y}) \succeq \boldsymbol{C}. \end{aligned}$$

Here, $\mathscr{A}^* : \mathbb{H}^m \to \mathbb{H}^D$ is the adjoint of $\mathscr{A} : \mathbb{H}^D \to \mathbb{H}^m$. It is defined as the (unique) map that obeys

$$(\boldsymbol{Y}, \mathscr{A}(\boldsymbol{Z})) = (\mathscr{A}^*(\boldsymbol{Y}), \boldsymbol{Z}) \quad \text{for all } \boldsymbol{Z} \in \mathbb{H}^D \text{ and } \boldsymbol{Y} \in \mathbb{H}^m.$$

Note that this definition is very similar to the formal definition of the adjoint matrix[1].

weak duality

As the name suggests, primal and dual SDP are dual versions of the same problem. *Weak duality* states that primal objective values are always be bounded by dual objective values:

$$(\boldsymbol{C}, \boldsymbol{Z}) \leq (\mathscr{A}^*(\boldsymbol{Y}), \boldsymbol{Z}) = (\boldsymbol{Y}, \mathscr{A}(\boldsymbol{Z})) = (\boldsymbol{Y}, \boldsymbol{B})$$

for every $\boldsymbol{Z} \in \mathbb{H}^D$ with $\mathscr{A}(\boldsymbol{Z}) = \boldsymbol{B}$, $\boldsymbol{Z} \succeq \boldsymbol{O}$ (primal feasible point) and $\boldsymbol{Y} \in \mathbb{H}^m$ with $\mathscr{A}^*(\boldsymbol{Y}) \succeq \boldsymbol{C}$ (dual feasible).

strong duality

**Definition 4.1 (strong duality).** A primal-dual SDP pair satisfies *strong duality* if there exists a primal feasible point $\boldsymbol{Z}_\sharp$ and a dual feasible point $\boldsymbol{Y}_\sharp$ such that $\left(\boldsymbol{C}, \boldsymbol{Z}_\sharp\right) = \left(\boldsymbol{Y}_\sharp, \boldsymbol{B}\right)$. In other words: primal and dual SDP produce the same optimal function value.

Most SDP pairs do satisfy strong duality. Standard results, like Slater's conditions, allow to quickly verify strong duality in concrete instances. All SDPs discussed in these lecture have this feature. The following technical statement follows from strong duality (without proof).

complementary slackness

**Theorem 4.2 (complementary slackness).** Suppose that $(\mathscr{A}, \boldsymbol{C}, \boldsymbol{D})$ characterizes an SDP that obeys strong duality and let $\boldsymbol{Z}_\sharp \in \mathbb{H}^D$ and $\boldsymbol{Y}_\sharp \in \mathbb{H}^m$ be optimal primal and dual feasible points (i.e. $(\boldsymbol{C}, \boldsymbol{Z}_\sharp) = (\boldsymbol{D}, \boldsymbol{Y}_\sharp)$). Then,

$$\mathscr{A}^*\left(\boldsymbol{Y}_\sharp\right)\boldsymbol{Z}_\sharp = \boldsymbol{C}\boldsymbol{Z}_\sharp \quad \text{and} \quad \mathscr{A}(\boldsymbol{Z}_\sharp)\boldsymbol{Y}_\sharp = \boldsymbol{D}\boldsymbol{Y}_\sharp.$$

Note that the second condition is trivial. It merely restates the linear constraints of the primal SDP. The first condition, however, is surprising. It provides a relation between optimal primal and dual solutions in terms of an equality between two matrix products. This matrix-valued equality subsumes many scalar equalities (for each matrix entry) and can unravel a lot of structure.

Sometimes, these structural relations provide enough guidance to 'guess' the optimal solution of an SDP (in standard form) without having to actually run any numerical solvers. This mindset uses SDPs as a mathematical proof technique and we will do one such example today. It addresses the diamond distance between two quantum channels which we introduced and discussed in the last lecture.

## 4.2 Setting and main result

diamond distance

Let $\mathscr{F}_1, \mathscr{F}_2 : \mathbb{H}^D \to \mathbb{H}^{D'}$ be two quantum channels. Last lecture, we introduced the *diamond distance* between these channels:

$$\|\mathscr{F}_1 - \mathscr{F}_2\|_\diamond = \max_{\boldsymbol{\rho} \succeq \boldsymbol{0}, \mathrm{tr}(\boldsymbol{\rho})=1} \left\|(\mathscr{F}_1 - \mathscr{F}_2) \otimes \mathscr{I}_D\left(\boldsymbol{\rho}\right)\right\|_1.$$

[1] For $\boldsymbol{A} \in \mathbb{C}^{D' \times D}$, the adjoint $\boldsymbol{A}^\dagger \in \mathbb{C}^{D \times D'}$ is the unique matrix that obeys $\langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{z}\rangle = \langle \boldsymbol{A}^\dagger \boldsymbol{y}, \boldsymbol{z}\rangle$ for all $\boldsymbol{z} \in \mathbb{C}^D$ and $\boldsymbol{y} \in \mathbb{C}^{D'}$.

The diamond distance quantifies the best achievable success probability when it comes to distinguishing $\mathcal{F}_1$ from $\mathcal{F}_2$ with a single channel use. Mathematically speaking it corresponds to maximizing a convex function (a shifted trace distance) over a convex set (the set of all quantum states in $\mathbb{H}^D \otimes \mathbb{H}^D$). Convexity tells us that this maximum is achieved at the boundary of the convex feasible set:

$$\|\mathcal{F}_1 - \mathcal{F}_2\|_\diamond = \max_{\boldsymbol{u} \in \mathbb{C}^D \otimes \mathbb{C}^D, \langle \boldsymbol{u}, \boldsymbol{u}\rangle = 1} \left\| (\mathcal{F}_1 - \mathcal{F}_2) \otimes \mathcal{I}_D \left( \boldsymbol{u}\boldsymbol{u}^\dagger \right) \right\|_1 .$$

This optimization problem does not look very convex. Somewhat surprisingly, it is nonetheless possible to reformulate this maximization of a convex function over a (non-)convex set as a semidefinite program:

primal SDP for diamond distance

$$\begin{aligned} \underset{\boldsymbol{W} \in \mathbb{H}^{D'} \otimes \mathbb{H}^D, \boldsymbol{\rho} \in \mathbb{H}^D}{\text{maximize}} \quad & 2D \ \left( \boldsymbol{J}\left(\mathcal{F}_1 - \mathcal{F}_2\right), \boldsymbol{W} \right) \\ \text{subject to} \quad & \mathbf{I}_{D'} \otimes \boldsymbol{\rho} \geq \boldsymbol{W}, \operatorname{tr}(\boldsymbol{\rho}) = 1, \\ & \boldsymbol{W}, \boldsymbol{\rho} \geq \mathbf{0}. \end{aligned} \tag{4.1}$$

Here, $\boldsymbol{J}(\mathcal{F}_1 - \mathcal{F}_2) = \boldsymbol{J}(\mathcal{F}_1) - \boldsymbol{J}(\mathcal{F}_2) \in \mathbb{H}^{D'} \otimes \mathbb{H}^D$ denotes the difference of Choi matrices. The SDP reformulation (4.1) has been the main insight from last lecture and today we further explore this SDP reformulation.

As a first step, we convert this SDP into (primal) standard form. This can be achieved by subsuming both optimization matrices into a single larger matrix variable

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{\rho} & \cdot & \cdot \\ \cdot & \boldsymbol{W} & \cdot \\ \cdot & \cdot & \boldsymbol{S} \end{pmatrix} \in \mathbb{H}^{D + D'D + D'D}.$$

Here, the dots on the off-diagonal blocks are placeholders for arbitrary matrix blocks. We need not constrain them, because optimization of the objective function will implicitly force these matrix blocks to vanish identically. The final block matrix $\boldsymbol{S}$ is a so-called slack variable. It allows us to reformulate the inequality constraint $\mathbf{I}_{D'} \otimes \boldsymbol{\rho} \geq \boldsymbol{W}$ as an equality constraint: $\mathbf{I}_{D'} \otimes \boldsymbol{\rho} = \boldsymbol{W} + \boldsymbol{S}$ for some $\boldsymbol{S} \geq \mathbf{0}$. Doing so allows us to recast the SDP in Eq. (4.1) as a primal SDP in standard form.

**Proposition 4.3 (diamond distance SDP in primal standard form).** The following triple captures the diamond distance SDP in standard form:

diamond distance SDP in standard form

$$\boldsymbol{C} = \begin{array}{cc} \begin{array}{ccc} \overset{D}{\longleftrightarrow} & \overset{DD'}{\longleftrightarrow} & \overset{DD'}{\longleftrightarrow} \end{array} & \\ \begin{pmatrix} \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & 2D\boldsymbol{J} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \end{pmatrix} & \begin{array}{l} \updownarrow D \\ \updownarrow DD' \\ \updownarrow DD' \end{array} \end{array} \qquad \boldsymbol{C} = \begin{array}{cc} \begin{array}{cc} \overset{1}{\longleftrightarrow} & \overset{DD'}{\longleftrightarrow} \end{array} & \\ \begin{pmatrix} 1 & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{pmatrix} & \begin{array}{l} \updownarrow 1 \\ \updownarrow DD' \end{array} \end{array},$$

and the linear map $\mathcal{A} : \mathbb{H}^{D + DD' + DD'} \to \mathbb{H}^{1 + DD'}$ acts like

$$\mathcal{A}\begin{pmatrix} \boldsymbol{\rho} & \cdot & \cdot \\ \cdot & \boldsymbol{W} & \cdot \\ \cdot & \cdot & \boldsymbol{S} \end{pmatrix} = \begin{pmatrix} \operatorname{tr}(\boldsymbol{\rho}) & \boldsymbol{O}_{1 \times DD'} \\ \boldsymbol{O}_{DD' \times 1} & \mathbf{I}_D \otimes \boldsymbol{\rho} - \boldsymbol{W} - \boldsymbol{S} \end{pmatrix}.$$

It is a relatively straightforward if tedious exercise to show that the adjoint map is

$$\mathcal{A}^*\begin{pmatrix} \lambda & \cdot \\ \cdot & \boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} \lambda \mathbf{I} + \mathrm{tr}_1(\boldsymbol{Y}) & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & -\boldsymbol{Y} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & -\boldsymbol{Y} \end{pmatrix}. \tag{4.2}$$

This allows us to readily infer the dual formulation of our diamond distance SDP.

**Corollary 4.4 (diamond distance SDP in simplified dual form).** The following SDP is the dual problem for computing $\|\mathcal{F}_1 - \mathcal{F}_2\|_\diamond$:

dual SDP for diamond distance

$$\begin{aligned} \underset{\boldsymbol{Y} \in \mathbb{H}^{DD'}}{\text{minimize}} \quad & \lambda_{\max}(\mathrm{tr}_1(\boldsymbol{Y})) \\ \text{subject to} \quad & \boldsymbol{Y} \geq 2D\boldsymbol{J}(\mathcal{F}_1 - \mathcal{F}_2), \\ & \boldsymbol{Y} \geq \boldsymbol{0}, \end{aligned} \tag{4.3}$$

where $\lambda_{\max}(\mathrm{tr}_1(\boldsymbol{Y}))$ is the largest eigenvalue of the partial trace of $\boldsymbol{Y}$.

**Exercise 4.5 (Proof of Corollary 4.4).** Verify Eq. (4.2), formulate the dual SDP in standard form and simplify it to obtain Corollary 4.4.

We are now ready to present the main result of today's lecture. In the following, we write $|\boldsymbol{A}|$ to denote the absolute value of a Hermitian matrix $\boldsymbol{A}$. This matrix is defined as $|\boldsymbol{A}| = \sqrt{\boldsymbol{A}^2}$. If $\boldsymbol{A} = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\dagger$ is an eigenvalue decomposition of $\boldsymbol{A}$, then $|\boldsymbol{A}| = \sum_i |\lambda_i| \boldsymbol{u}_i \boldsymbol{u}_i^\dagger$. This matrix is always positive semidefinite. We will also need the partial trace (over the first tensor factor):

$$\mathrm{tr}_1(\boldsymbol{S} \otimes \boldsymbol{T}) = \mathrm{tr}(\boldsymbol{S})\,\boldsymbol{T}$$

and linearly extended to all of $\mathbb{H}^{D'} \otimes \mathbb{H}^{D}$.

**Theorem 4.6 (simple diamond distance formula for certain channel differences).** Let $\boldsymbol{J}(\mathcal{F}_1 - \mathcal{F}_2) \in \mathbb{H}^{D'} \otimes \mathbb{H}^{D}$ be the Choi matrix difference of two quantum channels $\mathcal{F}_1, \mathcal{F}_2 : \mathbb{H}^{D} \to \mathbb{H}^{D'}$. The following two statements are equivalent:

analytic diamond distance for certain channel differences

$\mathrm{tr}_1(|\boldsymbol{J}(\mathcal{F}_1 - \mathcal{F}_2)|) = c\mathbf{I}_D$ for some constant $c \geq 0$.

**(2)** $\|\mathcal{F}_1 - \mathcal{F}_2\|_\diamond = \|\boldsymbol{J}(\mathcal{F}_1 - \mathcal{F}_2)\|_1$

This equivalence relation has first been established in Ref. [Kli+16] and combines two statements that are qualitatively very different. (1) is a property of the Choi matrix that resembles the notion of extremely mixed subsystems (think: maximal entanglement). This condition can be checked for entire families of quantum channels.

Condition (2), on the other hand, tells us that the diamond distance is much easier to compute than one might think. In particular, we do not need to solve an SDP at all. The trace norm of the Choi matrix can be computed with a single eigenvalue decomposition. Conceptually, (2) also tells us something remarkable about the channel distinguishability protocol. There is no need to optimize over bipartite input states at all. The maximally entangled (Bell) state is guaranteed to achieve optimal distinguishability.

## 4.3 Proof of Theorem 4.6

Let us now present the proof of Theorem 4.6. We will proceed in two steps: show that (1) implies (2) and then show that (2) also implies (1). These steps are technical, but use a standard template that is relatively straightforward to execute. This standard template is very versatile and can be used to get geometrical insights into many problems that are convex in nature. Sometimes, these insights can subsequently be converted into a much shorter and more elegant proof argument. We refer to Ref. [Mic+18] for a much shorter argument that does not require duality or SDPs.

### 4.3.1 Part A: Property (1) implies Property (2)

This is the easy direction. Let us use $\boldsymbol{J}$ as shorthand notation for $\boldsymbol{J}(\mathscr{F}_1 - \mathscr{F}_2)$. Property (1) ensures that this matrix obeys $\mathrm{tr}_1\left(|\boldsymbol{J}|\right) = c\mathbf{I}_D$ for some constant $c \geq 0$. Note that we can explicitly compute this constant by taking the normalized trace:

$$c = \frac{1}{D}\mathrm{tr}\left(c\mathbf{I}_D\right) = \frac{1}{D}\mathrm{tr}\left(\mathrm{tr}_1\left(|\boldsymbol{J}|\right)\right) = \frac{1}{D}\mathrm{tr}\left(|\boldsymbol{J}|\right) = \frac{1}{D}\|\boldsymbol{J}\|_1. \tag{4.4}$$

We can now use this structural assumption to guess a good solution for the (simplified) dual SDP (4.3):

$$\boldsymbol{Y}_{\text{guess}} = D\left(|\boldsymbol{J}| + \boldsymbol{J}\right)$$

A moment of thought reveals that $\boldsymbol{Y}_{\text{guess}} \geq 2D\boldsymbol{J}$ (because $|\boldsymbol{J}| \geq \boldsymbol{J}$) and $\boldsymbol{Y}_{\text{guess}} \geq \boldsymbol{O}$ (because $|\boldsymbol{J}| + \boldsymbol{J} = \sum_i \left(|\lambda_i| + \lambda_i\right)\boldsymbol{u}_i\boldsymbol{u}_i^\dagger$ and $|\lambda_i| + \lambda_i \geq 0$ for any $\lambda_i \in \mathbb{R}$). So, $\boldsymbol{Y}_{\text{guess}}$ is dual feasible. Moreover, Property (1) ensures

$$\mathrm{tr}_1\left(\boldsymbol{Y}_{\text{guess}}\right) = D\mathrm{tr}_1\left(|\boldsymbol{J}|\right) + D\mathrm{tr}_1\left(\boldsymbol{J}\right) = cD\mathbf{I}_D + \boldsymbol{O} = \|\boldsymbol{J}\|_1\mathbf{I}_D. \tag{4.5}$$

Here we have used our explicit computation of the constant $c$ from Eq. (4.4) and an important feature of Choi matrices: $\boldsymbol{J}\left(\mathscr{F}_i\right) = \mathbf{I}_D/D$, because $\mathscr{F}_i$ is trace preserving. This in turn demands

$$\mathrm{tr}_1(\boldsymbol{J}) = \mathrm{tr}_1\left(\boldsymbol{J}(\mathscr{F}_1)\right) - \mathrm{tr}_1\left(\boldsymbol{J}(\mathscr{F}_2)\right) = \frac{1}{D}\mathbf{I}_D - \frac{1}{D}\mathbf{I}_D = \boldsymbol{O}. \tag{4.6}$$

Eq. (4.5) is really the main insight, because it ensures that the optimal dual SDP function value must obey

$$\|\mathscr{F}_1 - \mathscr{F}_2\|_\diamond = \left\|\mathrm{tr}_1\left(\boldsymbol{Y}_\sharp\right)\right\|_\infty \leq \left\|\mathrm{tr}_1\left(\boldsymbol{Y}_{\text{guess}}\right)\right\|_\infty = \|\boldsymbol{J}\|_1\,\|\mathbf{I}_D\|_\infty = \|\boldsymbol{J}\|_1\,.$$

In turn, this produces an upper bound on the actual value of the diamond distance: An inequality in the converse direction readily follows from the definition of the diamond distance:

$$\begin{aligned}\|\mathscr{F}_1 - \mathscr{F}_2\|_\diamond &= \max_{\boldsymbol{u}\in\mathbb{C}^D\otimes\mathbb{C}^D,\langle\boldsymbol{u},\boldsymbol{u}\rangle=1}\left\|(\mathscr{F}_1 - \mathscr{F}_2)\otimes\mathcal{I}_D\left(\boldsymbol{u}\boldsymbol{u}^\dagger\right)\right\|_1\\ &\geq \left\|(\mathscr{F}_1 - \mathscr{F}_2)\otimes\mathcal{I}_D\left(\boldsymbol{\omega}\boldsymbol{\omega}^\dagger\right)\right\|_1\\ &= \|\boldsymbol{J}\left(\mathscr{F}_1 - \mathscr{F}_2\right)\|_1 = \|\boldsymbol{J}\|_1\,.\end{aligned}$$

### 4.3.2 Part B: Property (2) implies Property (1)

Now, we assume $\|\mathcal{F}_1 - \mathcal{F}_2\|_\diamond = \|\boldsymbol{J}(\mathcal{F}_1 - \mathcal{F}_2)\|_1$ and must use this information to deduce structural insights about the absolute value of the Choi matrix. Again, we use the shorthand notation $\boldsymbol{J} = \boldsymbol{J}(\mathcal{F}_1 - \mathcal{F}_2)$ and guess a good solution for the (simplified) primal SDP:

$$\boldsymbol{\rho}_{\text{guess}} = \frac{1}{D}\mathbf{I}_D \quad \text{and} \quad \boldsymbol{W}_{\text{guess}} = \frac{1}{2D}\left(\mathbf{I}_{D'} \otimes \mathbf{I}_D + \text{sign}(\boldsymbol{J})\right),$$

where $\text{sign}(\boldsymbol{J}) = \sum_i \text{sign}(\lambda_i)\boldsymbol{u}_i\boldsymbol{u}_i^\dagger$ if $\boldsymbol{J} = \sum_i \lambda_i \boldsymbol{u}_i\boldsymbol{u}_i^\dagger$. The matrix sign generalizes the scalar sign function to matrices. The resulting eigenvalues take on $+1$, $0$ and $-1$ which ensures that $\boldsymbol{W}_{\text{guess}} \geq \boldsymbol{0}$. In addition, $\boldsymbol{\rho}_{\text{guess}}$ is the maximally mixed state and therefore feasible. Moreover,

$$\mathbf{I}_{D'} \otimes \boldsymbol{\rho}_{\text{guess}} = \frac{1}{2D}\left(\mathbf{I}_{D'} \otimes \mathbf{I}_D + \mathbf{I}_{D'} \otimes \mathbf{I}_D\right) \geq \frac{1}{2D}\left(\mathbf{I}_{D'} \otimes \mathbf{I}_D + \text{sign}(\boldsymbol{J})\right) = \boldsymbol{W}_{\text{guess}}.$$

This guess achieves a (simplified) primal SDP value of

$$2D\left(\boldsymbol{J}, \boldsymbol{W}_{\text{guess}}\right) = \left(\boldsymbol{J}, \mathbf{I}_{D'} \otimes \mathbf{I}_D\right) + \left(\boldsymbol{J}, \text{sign}(\boldsymbol{J})\right) = \text{tr}\left(\boldsymbol{J}\right) + \|\boldsymbol{J}\|_1 = \|\boldsymbol{J}\|_1,$$

because $\boldsymbol{J}$ is traceless (recall from Eq. (4.6) that already the partial trace must vanish). Property (2) ensures that this objective value is actually optimal and so this (simplified) primal feasible pair is actually an optimal feasible point. We can use it to readily extract a primal optimal point for the diamond distance SDP in standard form from Proposition 4.3:

$$\begin{aligned}\boldsymbol{Z}_\sharp &= \begin{pmatrix} \boldsymbol{\rho}_{\text{guess}} & \cdot & \cdot \\ \cdot & \boldsymbol{W}_{\text{guess}} & \cdot \\ \cdot & \cdot & \mathbf{I}_{D'} \otimes \boldsymbol{\rho}_{\text{guess}} - \boldsymbol{W}_{\text{guess}} \end{pmatrix} \\ &= \frac{1}{2D}\begin{pmatrix} 2\mathbf{I}_D & \cdot & \cdot \\ \cdot & \mathbf{I}_{D'} \otimes \mathbf{I}_D + \text{sign}(\boldsymbol{J}) & \cdot \\ \cdot & \cdot & \mathbf{I}_{D'} \otimes \mathbf{I}_D - \text{sign}(\boldsymbol{J}) \end{pmatrix}.\end{aligned}$$

This is our starting point for applying Theorem 4.2 (complementary slackness). We make the following ansatz for the dual optimal point

$$\boldsymbol{Y}_\sharp = \begin{pmatrix} \lambda_\sharp & \cdot \\ \cdot & \tilde{\boldsymbol{Y}}_\sharp \end{pmatrix}.$$

Complementary slackness the provides the following nontrivial relation between the primal optimal point (which we know) and the dual optimal point (which we don't yet know): $\mathcal{A}^*\left(\boldsymbol{Y}_\sharp\right)\boldsymbol{Z}_\sharp = \boldsymbol{C}\boldsymbol{Z}_\sharp$, where $\mathcal{A}^*$ is defined in Eq. (4.2) and $\boldsymbol{C}$ is defined in Proposition 4.3. Relatively straightforward block matrix multiplications yield

$$\begin{aligned}\mathcal{A}^*\left(\boldsymbol{Y}_\sharp\right)\boldsymbol{Z}_\sharp &= \begin{pmatrix} \lambda_\sharp \mathbf{I}_D + \text{tr}_1\left(\tilde{\boldsymbol{Y}}_\sharp\right) & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & -\tilde{\boldsymbol{Y}}_\sharp & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & -\tilde{\boldsymbol{Y}}_\sharp \end{pmatrix} \frac{1}{2D}\begin{pmatrix} 2\mathbf{I}_D & \cdot & \cdot \\ \cdot & \mathbf{I}_{D'} \otimes \mathbf{I}_D + \text{sign}(\boldsymbol{J}) & \cdot \\ \cdot & \cdot & \mathbf{I}_{D'} \otimes \mathbf{I}_D - \text{sign}(\boldsymbol{J}) \end{pmatrix} \\ &= \frac{1}{2D}\begin{pmatrix} 2\lambda_\sharp \mathbf{I}_D + 2\text{tr}_1\left(\tilde{\boldsymbol{Y}}_\sharp\right) & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & -\tilde{\boldsymbol{Y}}_\sharp - \tilde{\boldsymbol{Y}}_\sharp\text{sign}(\boldsymbol{J}) & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & -\tilde{\boldsymbol{Y}}_\sharp + \tilde{\boldsymbol{Y}}_\sharp\text{sign}(\boldsymbol{J}) \end{pmatrix},\end{aligned}$$

as well as

$$\begin{aligned}
\boldsymbol{C}\boldsymbol{Z}_\sharp =& \begin{pmatrix} \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & 2D\boldsymbol{J} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \end{pmatrix} \frac{1}{2D} \begin{pmatrix} 2\mathbf{I}_D & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \mathbf{I}_{DD'} + \mathrm{sign}(\boldsymbol{J}) & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \mathbf{I}_{DD'} - \mathrm{sign}(\boldsymbol{J}) \end{pmatrix} \\
=& \begin{pmatrix} \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & (\boldsymbol{J} + |\boldsymbol{J}|) & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \end{pmatrix}.
\end{aligned}$$

Complementary slackness (Theorem 4.2) equates these two matrix products which really produces three different matrix-valued equations:

$$\begin{aligned}
\frac{1}{2D}\left(2\lambda_\sharp \mathbf{I}_D + 2\mathrm{tr}_1\left(\tilde{\boldsymbol{Y}}_\sharp\right)\right) =& \boldsymbol{O}, \\
-\frac{1}{2D}\left(\tilde{\boldsymbol{Y}}_\sharp + \tilde{\boldsymbol{Y}}_\sharp \mathrm{sign}\left(\boldsymbol{J}\right)\right) =& \boldsymbol{J} + |\boldsymbol{J}|, \\
-\tilde{\boldsymbol{Y}}_\sharp + \tilde{\boldsymbol{Y}}_\sharp \mathrm{sign}(\boldsymbol{J}) =& \boldsymbol{O},
\end{aligned}$$

or equivalently:

$$\begin{aligned}
\lambda_\sharp \mathbf{I}_D =& -\mathrm{tr}\left(\tilde{\boldsymbol{Y}}_\sharp\right), \\
\tilde{\boldsymbol{Y}}_\sharp + \tilde{\boldsymbol{Y}}_\sharp \mathrm{sign}(\boldsymbol{J}) =& -2D\left(\boldsymbol{J} + |\boldsymbol{J}|\right), \\
\tilde{\boldsymbol{Y}}_\sharp \mathrm{sign}(\boldsymbol{J}) =& \tilde{\boldsymbol{Y}}_\sharp.
\end{aligned}$$

Inserting the third equality into the second one yields $\tilde{\boldsymbol{Y}}_\sharp = -D\left(\boldsymbol{J} + |\boldsymbol{J}|\right)$. Note that this is exactly the guess $\boldsymbol{Y}_{\mathrm{guess}}$ we made in the first part of the proof. But this time it is a rigorous consequence of knowing the primal optimal solution and using complementary slackness. We can insert this observation into the remaining equation to conclude

$$\lambda_\sharp \mathbf{I}_D = -\mathrm{tr}\left(\tilde{\boldsymbol{Y}}_\sharp\right) = +D\mathrm{tr}\left(\boldsymbol{J} + \left|\tilde{\boldsymbol{J}}\right|\right) = \boldsymbol{O} + D\mathrm{tr}_1\left(|\boldsymbol{J}|\right) = D\mathrm{tr}_1\left(|\boldsymbol{J}|\right),$$

because $\mathrm{tr}\left(\boldsymbol{J}\right) = \boldsymbol{O}$ according to Eq. (4.6). This, however, is just Property (1) if we divide by $D$ and set $c = \lambda_\sharp / D \in \mathbb{R}$.

## 4.4 Consequences: diamond distance between Pauli channels

Let us now briefly move closer to standard quantum information and computation and consider channels on $n$-qubit systems, i.e. $D = 2^n$. A $n$-qubit Pauli matrix/operator is a $n$-fold tensor product of single-qubit Paulis:

$n$-qubit Pauli operator

$$\begin{aligned}
\boldsymbol{P} =& \boldsymbol{P}_1 \otimes \cdots \otimes \boldsymbol{P}_n \quad \text{with} \quad \boldsymbol{P}_i \in \left\{\mathbf{I}, \boldsymbol{W}_x, \boldsymbol{W}_y, \boldsymbol{W}_z\right\}, \\
\boldsymbol{W}_x =& \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \ \boldsymbol{W}_y = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \ \boldsymbol{W}_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}
\end{aligned}$$

There are in total $4^n$ different $n$-qubit Pauli matrices and we can label them by integers and use the convention $\boldsymbol{P}_1 = \mathbf{I} \otimes \cdots \otimes \mathbf{I} = \mathbf{I}_D$. Each $n$-qubit Pauli matrix is Hermitian, i.e. $\boldsymbol{P} \in (\mathbb{H}^2)^{\otimes n} \simeq \mathbb{H}^D$ and obeys $\boldsymbol{P}^2 = \mathbf{I}_2 \otimes \cdots \otimes \mathbf{I}_2 = \mathbf{I}_D$.

Different Pauli operators are mutually orthogonal: $\langle \boldsymbol{P}, \boldsymbol{P}' \rangle = \mathrm{tr}\left(\boldsymbol{P}\boldsymbol{P}'\right) = 0$ if $\boldsymbol{P} \neq \boldsymbol{P}'$. We say that a channel $\mathscr{P} : \mathbb{H}^D \to \mathbb{H}^D$ is a ($n$-qubit) *Pauli channel* if we can write it as

$n$-qubit Pauli channel

$$\mathscr{P}\left(\boldsymbol{\rho}\right) = \sum\nolimits_{i=1}^{4^n} p_i \boldsymbol{P}_i \rho \boldsymbol{P}_i \quad \text{with } p_i \geq 0 \text{ and } \sum\nolimits_{i=1}^{4^n} p_i = 1.$$

Intuitively: we apply the $n$-qubit Pauli matrix $\boldsymbol{P}_i$ (viewed as a unitary) with probability $p_i$. Pauli channels are very important when modeling noise in quantum computation. Examples include bit flip, phase flip, depolarizing noise on single and multiple qubits. In this context, the most important Pauli channel is $\mathscr{I}(\boldsymbol{\rho}) = \boldsymbol{P}_1 \boldsymbol{\rho} \boldsymbol{P}_1$, aka do nothing or perfect execution of a certain functionality. We can use our findings to derive an easy formula for diamond distances between Pauli channels.

**Proposition 4.7** Fix $D = 2^n$ ($n$ qubits) and let $\mathscr{P}_1(\boldsymbol{\rho}) = \sum_{i=1}^{4^n} \boldsymbol{P}_i \boldsymbol{\rho} \boldsymbol{P}_i$ and $\mathscr{P}_2(\boldsymbol{\rho}) = \sum_{i=1}^{4^n} q_i \boldsymbol{P}_i \boldsymbol{\rho} \boldsymbol{P}_i$ be two Pauli channels. Then,

easy formula for diamond distance of Pauli channels

$$\left\|\mathscr{P}_1 - \mathscr{P}_2\right\|_\diamond = \sum\nolimits_{i=1}^{4^n} \left|p_i - q_i\right|.$$

In words: the diamond distance between two Pauli channels is exactly given by the total variational distance of the accompanying weight distributions.

*Proof.* This claim is almost an immediate consequence of Theorem 4.6. We 'only' have to recognize the relevant structure. Let us start by writing down the Choi matrix of the difference:

$$\begin{aligned} \boldsymbol{J} =& \boldsymbol{J}(\mathscr{P}_1) - \boldsymbol{J}(\mathscr{P}_2) = \sum\nolimits_{i=1}^{4^n} \left(p_i - q_i\right) \left(\boldsymbol{P}_i \otimes \mathbf{I}_D\right) \boldsymbol{\omega}\boldsymbol{\omega}^\dagger \left(\boldsymbol{P}_i \otimes \mathbf{I}_D\right) \\ =& \sum\nolimits_{i=1}^{4^n} (p_i - q_i) \boldsymbol{\omega}_i \boldsymbol{\omega}_i^\dagger \end{aligned} \tag{4.7}$$

with $\boldsymbol{w}_i = \left(\boldsymbol{P}_i \otimes \mathbf{I}\right) \boldsymbol{\omega}$ and $\boldsymbol{\omega} = (1/D) \sum_{i=1}^{D} \boldsymbol{e}_i \otimes \boldsymbol{e}_i$. Now, note that all these vectors are orthonormal and maximally entangled (think Bell basis, but more general). Orthonormality is implied by

$$\begin{aligned} \langle \boldsymbol{\omega}_i, \boldsymbol{\omega}_j \rangle =& \boldsymbol{\omega}_i^\dagger \boldsymbol{\omega}_j = \boldsymbol{\omega}^\dagger (\boldsymbol{P}_i \otimes \mathbf{I}(\boldsymbol{P}_j \otimes \mathbf{I})\boldsymbol{\omega} \\ =& \boldsymbol{\omega}^* \left(\boldsymbol{P}_i \boldsymbol{P}_j \otimes \mathbf{I}_D\right) \boldsymbol{w} = \frac{1}{D} \mathrm{tr}\left(\boldsymbol{P}_i \boldsymbol{P}_j\right) = \delta_{i,j}, \end{aligned}$$

while maximum entanglement follows from

$$\begin{aligned} \mathrm{tr}_1\left(\boldsymbol{\omega}_i \boldsymbol{\omega}_i^\dagger\right) =& \mathrm{tr}_1\left((\boldsymbol{P}_i \otimes \mathbf{I})\boldsymbol{\omega}\boldsymbol{\omega}^\dagger(\boldsymbol{P}_i \otimes \mathbf{I})\right) \\ =& \mathrm{tr}_1\left((\boldsymbol{P}_i^2 \otimes \mathbf{I})\boldsymbol{\omega}\boldsymbol{\omega}^\dagger\right) = \mathrm{tr}_1\left(\boldsymbol{\omega}_1 \boldsymbol{\omega}_1^\dagger\right) = \frac{1}{D}\mathbf{I}_D. \end{aligned}$$

Orthnormality of the $\boldsymbol{\omega}_i$s ensures that our formula from Eq. (4.7) is actually an eigenvalue decomposition. This makes it very easy to compute the matrix absolute value, its partial trace and its trace norm:

$$
\begin{aligned}
|\boldsymbol{J}| &= \sum\nolimits_{i=1}^{4^n} |p_i - q_i| \, \boldsymbol{\omega}_i \boldsymbol{\omega}_i^*, \\
\mathrm{tr}_1\left(|\boldsymbol{J}|\right) &= \sum\nolimits_{i=1}^{4^n} |p_i - q_i| \, \mathrm{tr}_1\left(\boldsymbol{\omega}_1 \boldsymbol{\omega}_1^\dagger\right) = \left(\sum\nolimits_{i=1}^{4^n} |p_i - q_i|\right) \frac{1}{D}\mathbf{I}_D = c\mathbf{I}_D, \\
\|\boldsymbol{J}\|_1 &= \mathrm{tr}\left(|\boldsymbol{J}|\right) = \mathrm{tr}\left(\mathrm{tr}_1\left(|\boldsymbol{J}|\right)\right) = \left(\sum\nolimits_{i=1}^{4^n} |p_i - q_i|\right) \mathrm{tr}\left(\frac{1}{D}\mathbf{I}_D\right) \\
&= \sum\nolimits_{i=1}^{4^n} |p_i - q_i| \, .
\end{aligned}
$$

Here, we have also used the observation that each $\boldsymbol{\omega}_i$ is a maximally entangled state. The second computation ensures that diamond distance and trace distance are the same, the third display computes the latter. ■

## 4.5 Problems

**Problem 4.8 (dual SDP for diamond distance).** Derive the dual SDP for the diamond distance provided in Corollaty 4.4.

**Problem 4.9 (Problems 3.10 and 3.12 with benefit of hindsight).**

1. Fix $D = 2$ and compute the diamond distance between $\mathcal{I}(\boldsymbol{\rho})$ (do nothing) and $\mathcal{W}_p(\boldsymbol{\rho}) = (1-p)\mathcal{I}(\boldsymbol{\rho}) + p\mathcal{W}(\boldsymbol{\rho})$ (Pauli channel) with $p \in [0,1]$ and $\mathcal{W}(\boldsymbol{\rho}) = \boldsymbol{W}\boldsymbol{\rho}\boldsymbol{W}$ with $\boldsymbol{W} \in \{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}\}$.
2. Fix $D \geq 2$ and compute the diamond distance between $\mathcal{I}(\boldsymbol{\rho})$ (do nothing) and $\mathcal{D}_p(\boldsymbol{\rho}) = (1-p)\boldsymbol{\rho} + p\frac{(\mathbf{I},\boldsymbol{\rho})}{D}\mathbf{I}$.

**Problem 4.10 (Diamond distance for tensor products of depolarizing channels).** Fix $D = 2$, $n \in \mathbb{N}$ and let $\mathcal{D}_{p_i}(\boldsymbol{\rho}) = (1-p_i)\boldsymbol{\rho} + p_i\frac{(\mathbf{I},\boldsymbol{\rho})}{D}\mathbf{I}$ with $1 \leq i \leq n$ be different single-qubit depolarizing channels. Compute the diamond distance between the the $n$-fold tensor product of the identity $\mathcal{I}_1 \otimes \cdots \otimes \mathcal{I}_n$ (do nothing on $n$-qubits) and tensor product $\mathcal{D}_{p_1} \otimes \cdots \otimes \mathcal{D}_{p_n}$ of these depolarizing channels. In formulas, compute

$$
\left\|\mathcal{I} \otimes \cdots \otimes \mathcal{I} - \mathcal{D}_{p_1} \otimes \cdots \otimes \mathcal{D}_{p_n}\right\|_\diamond = ?.
$$

# 5. Complexity by design

**Date:** 03 March 2023

The quantum complexity of a unitary transformation or quantum state is defined as the size of the shortest quantum computation that executes the unitary or prepares the state. It is reasonable to expect that the complexity of a quantum state governed by a chaotic many-body Hamiltonian grows linearly with time for a time that is exponential in the system size; however, because it is hard to rule out a shortcut that improves the efficiency of a computation, it is notoriously difficult to derive lower bounds on quantum complexity for particular unitaries or states without making additional assumptions. To go further, one may study more generic models of complexity growth. We provide a rigorous connection between complexity growth and unitary $K$-designs, ensembles that capture the randomness of the unitary group. This connection allows us to leverage existing results about design growth to draw conclusions about the growth of complexity.

This chapter discusses joint work with *Nicholas Hunter Jones*, *Wissam Chemissany*, *Fernando G.S.L. Brandão* and *John Preskill* [Bra+21].

**Agenda:**

1. Motivation and statement of results
2. Proof part 1: almost all states have high complexity (concentration+counting)
3. Proof part 2: complexity by design (partial derandomization)
4. Generalizations and follow-up work

**Warning 5.1 (different notational conventions).** This final lecture has a dedicated quantum computing focus. We underscore this by switching notational conventions: today, we use bra-ket notation $|\psi\rangle$ (column vector), $\langle\psi|$ (adjoint row vector) and normal text $U, V$ for matrices. We refer to Table 5.1 for additional notation conventions. ▪

| symbol | meaning |
|---|---|
| $d$ | local dimension (qu*d*its) |
| $n$ | number of qu*d*its |
| $D = d^n$ | total Hilbert space dimension |
| $\mathsf{G}$ | universal 2-qu*d*it gate set |
| $T, R$ | circuit sizes (i.e. nr. of gates) |
| $\mathsf{V}_R$ | set of all local size-$R$ circuits<br>(with 2-local gates chosen from $\mathsf{G}$) |

**Table 5.1** Summary of the notation conventions used in this talk.

## 5.1 Motivation and statement of results

The complexity of a *circuit* (quantum or classical) is defined as the minimal number of elementary steps needed to evaluate the function. This depends on the choice of model ('gate set'), but only in a mild way. It allows us to assert whether a given computational task is 'easy' (small complexity) or 'hard' (high complexity).

In quantum physics, the notion of complexity extends meaningfully to quantum states as well. State complexity measures the effort/time required to produce $|\psi\rangle$ from a simple starting state $|\psi_0\rangle$, e.g. a product state $|0\rangle$.

Here are two basic facts about the analysis of complexity:

- *upper bounds are 'easy'*, because every circuit decomposition yields one for free. Certain circuit families also come with universal upper bounds, e.g. $2^{O(n)}$ for $n$-qubit quantum circuits and $O(n^2/\log(n))$ for $n$-qubit Clifford circuits.
- *lower bounds are 'hard'*, because it requires us to rule out potential shortcuts. In classical Boolean logic, complexity captures the notion of optimal compilation. This problem sits in the second level of the polynomial hierarchy. Quantum circuit compilation is even harder.

Circuit complexity has long been a prominent foundational concept in (classical and quantum) computer science. Recently, *state complexity* has been identified as a useful concept in quantum physics. Here are a couple of examples:

**a1** *topological phases of matter (at zero temperature)* can be classified using the complexity of the ground state wave function;

**a2** *chaotic Hamiltonians* produce long-time quantum evolutions that generate highly complex states;

**a3** the *AdS/CFT-correspondence* posits that the complexity of a quantum state of the boundary theory corresponds to the volume in the bulk geometry, which is hidden behind the event horizon of a black hole.

It is extremely difficult to study complexity growth for concrete Hamiltonian evolutions. An alternative approach is to consider ensembles of circuits, and

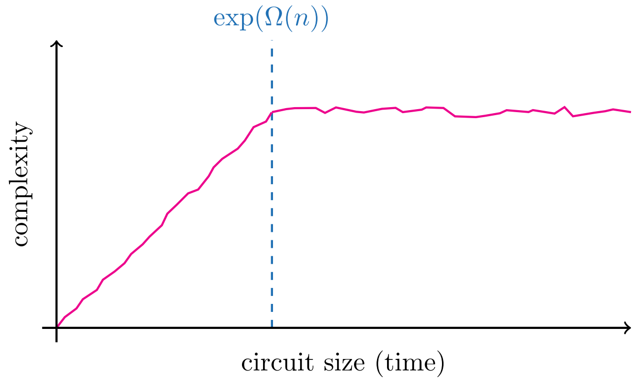

**Figure 5.1** *Expected complexity growth in random circuits.* Conjecture 5.2 states that, for random quantum circuits acting on $n$ qu$d$its, the circuit complexity grows linearly with circuit size (time) until it saturates at a value exponentially large in $n$.

to derive lower bounds on complexity, which hold with high probability when samples are selected from these ensembles. This together with the AdS/CFT conjecture gave rise to the following conjecture:

**Conjecture 5.2** **(Brown, Susskind [BS18]).** Most local (random) circuits of size $T$ have a complexity that scales linearly in $T$ for an exponentially long time.

the complexity of random circuits is conjectured to grow linearly with size

This conjecture is visualized in Figure 5.1 Today, we will prove a related statement regarding the growth of state complexity under local random circuits on $n$ qui$d$its ($D = d^n$). To achieve such a goal, we will work with the following standard notion of state complexity.

**Definition 5.3** **(state complexity).** Let $\mathsf{V}_R$ be the set of all size-$R$ circuits on $n$ qui$d$its with gates from a universal gate set $\mathsf{G}$, let $|\psi_0\rangle$ be a fixed 'simple' state. For $\delta \in (0,1)$, we say that $|\psi\rangle \in \mathbb{C}^D$ has $\delta$-complexity at most $R$ if

formal definition of state complexity

$$\min_{V \in \mathsf{V}_R} \| |\psi\rangle\langle\psi| - U|\psi_0\rangle\langle\psi_0|U^\dagger \|_1 \leq \delta,$$

or, equivalently

$$\max_{V \in \mathsf{V}_R} |\langle\psi|V|\psi_0\rangle|^2 \geq 1 - \delta^2.$$

if this is the case, we write $C_\delta(|\psi\rangle) \leq R$.

Based on this formal definition, we will prove the following rigorous lower bound on typical state complexity generated by local random circuits.

**Theorem 5.4** **('polynomial' growth in state complexity, informal).** 'Very many' local random circuits of size $T$ produce states with complexity (at least) $\Omega\left(T^{1/(5+o(1))}\right)$. This growth persists up to circuit sizes $T \approx \sqrt{D} = d^{n/2}$.

'polynomial' state complexity growth for exponentially long time

Similar statements are true for stronger notions of state and circuit complexity as well. These are inspired by state and channel discrimination tasks and we will briefly discuss them in Section 5.5.

## 5.2 Proof part 1: Almost all states have high complexity

Let us start by analyzing the complexity of generic (i.e. Haar-random) $n$-qu$d$it states. These are states sampled uniformly from the complex unit sphere in $D = d^n$ dimensions. There, concentration of measure together with a simple counting argument yield exponentially strong lower bounds on the state complexity.

**Lemma 5.5 (exponential concentration for Haar random states).** Fix $|v\rangle \in \mathbb{C}^D$ ($D = d^n$) and let $|h\rangle \overset{Haar}{\sim} \mathbb{C}^D$ be a Haar random state. Then,

exponential concentration for Haar-random states

$$\Pr_{|h\rangle}\left[|\langle v,h\rangle|^2 \geq \tau\right] \leq 2\mathrm{e}^{-D\tau/2} \quad \text{for any } \tau \geq 0.$$

This is a poor man's variant of a beautiful measure concentration phenomenon called Levy's lemma. It applies much more generally and is best proved via isoperometric inequalities. The argument presented here, however, does provide valuable guidance on how to deal with ensembles that are not quite Haar random. Our proof will be based on the following fact that is somewhat folklore in the quantum information community, see e.g. [Kue19, Lecture 05].

**Fact 5.6 (Haar integration; folklore).** Let $|h\rangle \overset{Haar}{\sim} \mathbb{C}^D$ ($D = d^n$) be a Haar random state. Then, for all $k \in \mathbb{N}_+$

$$\mathbb{E}_{|h\rangle}\left[(|h\rangle\langle h|)^{\otimes k}\right] = \int_{\text{Haar}} (|h\rangle\langle h|)^{\otimes k}\,\mathrm{d}\mu(h) = \binom{D+k-1}{k}^{-1} P_{\vee^k}, \tag{5.1}$$

where $P_{\vee^k}$ is the projector onto the totally symmetric subspace of $\left(\mathbb{C}^D\right)^{\otimes k}$. ■

*Proof of Lemma 5.5.* Let us start by computing the moments of the random variable $|\langle v,h\rangle|^2$. For any $k \in \mathbb{N}_+$ the Haar integration formula (Fact 5.6) yields

$$\begin{aligned}\mathbb{E}_{|h\rangle}\left[|\langle v,h\rangle|^{2k}\right] =& \mathrm{tr}\left((|v\rangle\langle v|)^{\otimes k}\,\mathbb{E}_{|h\rangle}\left[(|h\rangle\langle h|)^{\otimes k}\right]\right)\\ =& \mathrm{tr}\left((|v\rangle\langle v|)^{\otimes k}\binom{D+k-1}{k}^{-1} P_{\vee^k}\right) \qquad (5.2)\\ =& \binom{D+k-1}{k} \leq \frac{k!}{D^k}. \qquad (5.3)\end{aligned}$$

This moment behavior indicates sub-exponential moment growth. We can now use some elementary tricks from probability theory to turn these moment bounds into an exponential concentration bound:

$$\begin{aligned}
\Pr_{|h\rangle}\left[|\langle v,h\rangle|^2 \geq \tau\right] =& \Pr_{|h\rangle}\left[D|\langle v,h\rangle|^2/2 \geq D\tau/2\right] \\
=& \Pr_{|h\rangle}\left[\exp\left(D|\langle v,h\rangle|^2/2\right) \geq \exp\left(D\tau/2\right)\right] \\
\leq& \mathrm{e}^{-D\tau/2}\mathbb{E}\left[\exp\left(D|\langle v,h\rangle|^2/2\right)\right] \\
=& \mathrm{e}^{-D\tau/2}\sum_{k=0}^{\infty}\frac{1}{k!}\frac{D^k}{2^k}\mathbb{E}_{|h\rangle}\left[|\langle v,h\rangle|^{2k}\right] \\
\leq& \mathrm{e}^{-D\tau/2}\sum_{k=0}^{\infty}\frac{1}{2^k} = 2\mathrm{e}^{-D\tau/2}.
\end{aligned}$$

The key step is Markov's inequality ($\Pr\left[S \geq \alpha\right] \leq \mathbb{E}\left[S\right]/\alpha$ for any nonnegative random variable $S$) in line three. ■

Exponential concentration for Haar-random states implies the following strong claim about the complexity minimal complexity.

**Proposition 5.7 (Haar random states have almost maximal complexity).** A Haar-random state $|h\rangle \in \mathbb{C}^D$ obeys

almost all states have almost maximal state complexity

$$\Pr_{|h\rangle}\left[C_\delta(|h\rangle) \leq R\right] \leq n^{2R}|\mathsf{G}|^R\mathrm{e}^{-D(1-\delta^2)/2} \quad \text{for any } R \in \mathbb{N}_+.$$

This probability remains tiny until

$$R \approx \frac{D}{2\log(n)} = \frac{d^n}{2\log(n)}.$$

Note that the Haar measure is fair in the sense that it assigns the same infinitesimal weight to each $D$-dimensional quantum state. Laplace's definition of probability therefore allows us to interpret Eq. (5.7) as a bound on the relative volume of complexity-$R$ states. This volume remains tiny until $R$ approaches the overall Hilbert space dimension $D = 2^n$. In other words: *almost all states have almost maximal state complexity*.

*Proof of Proposition 5.7.* Insert the definition of state complexity, see Definition 5.3, and apply a union bound (Boole's inequality) to obtain

$$\begin{aligned}
\Pr_{|h\rangle}\left[C_\delta(|h\rangle) \leq R\right] =& \Pr_{|h\rangle}\left[\max_{V\in\mathsf{V}_R}|\langle V\psi_0,h\rangle|^2 \geq 1-\delta^2\right] \\
\leq& \sum_{V\in\mathsf{V}_R}\Pr\left[|\langle V\psi_0,h\rangle|^2 \geq 1-\delta^2\right].
\end{aligned}$$

We can now apply Lemma 5.5 to each term in the sum. This produces

$$\Pr_{|h\rangle}\left[C_\delta(|h\rangle) \leq R\right] \leq |\mathsf{V}_R|\,\mathrm{e}^{-D(1-\delta^2)/2}$$

and the claim follows from counting the number of different circuits in $\mathsf{V}_R$. ■

## 5.3 Proof part 2: complexity by design

Our study of the complexity of Haar random states is a promising starting point. But it doesn't allow us to address less generic state ensembles. One solution is to apply a *partial derandomization* based on the Haar integration formula (5.1). Proposition 5.7 is contingent on the assumption that this formula is true for *all* tensor powers $k \in \mathbb{N}_+$ This allowed us to control all $k$ moments of $|\langle v, h\rangle|^2$ and arrive at an exponentially strong concentration formula (Lemma 5.5). We can relax these assumptions by assuming that the Haar integration formula is only approximately true for the first $K$ tensor powers. Ensembles with this property are called $\epsilon$-approximate $K$-designs.

(approximate) $K$-design

**Definition 5.8 (approximate $K$-design).** Fix $\epsilon \in (0,1)$ and a threshold $K \in \mathbb{N}_+$. We say that a state ensemble $\{p_i, |h_i\rangle\} \subset \mathbb{C}^D$ forms an $\epsilon$-approximate $K$-design if

$$\left\| \mathbb{E}_{|h\rangle}\left[ (|h\rangle\langle h|)^{\otimes k} \right] - \binom{D+k-1}{k}^{-1} P_{\vee^k} \right\|_1 \leq \epsilon \quad \text{for all } k = 1, \ldots, K.$$

Note that for $K = 1$, this requirement is met by any state ensemble whose average is close to the maximally mixed state. Sampling computational basis states uniformly at random is one concrete example. Haar random states are another extreme case that occurs when we let $K$ tend to infinity. Adjusting the design order $K$ allows us to interpolate between those extremes. And, remarkably, the typical state complexity associated with such ensembles varies accordingly.

most $K$-design states have complexity linear in $K$

**Proposition 5.9 (complexity by $K$-design).** Suppose that $|h\rangle \in \mathbb{C}^D$ is sampled from an $\epsilon$-approximate $K$-design. Then,

$$\Pr_{|h\rangle}\left[ C_\delta(|h\rangle) \right] \leq n^R |\mathsf{G}|^R \left( \frac{K}{(1-\delta^2)D} \right)^K \quad \text{for all } R \in \mathbb{N}_+.$$

This probability remains tiny until

$$R \approx \frac{K\,(n - \log(K))}{\log(n)}.$$

> **Warning 5.10** Eq. (5.9) becomes vacuous once the design order $K$ approaches the total system size $D = d^n$. This puts an upper limit on the amount of progress we can make by letting $K$ become larger and larger. ▪

*Proof of Proposition 5.9.* We will do the proof for the extreme case $\epsilon = 0$. An extension to $\epsilon > 0$ is relatively straightforward. The key ingredient is to replace the Haar concentration formula with a weaker concentration bound that only uses the first $K$ moments. Fix $|v\rangle \in \mathbb{C}^D$ arbitrary and $\tau \geq 0$. Markov's inequality then implies

polynomial concentration for $K$-designs

$$\begin{aligned}\Pr_{|h\rangle}\left[|\langle v,h\rangle|^2 \geq \tau\right] =& \Pr_{|h\rangle}\left[|\langle v,h\rangle|^{2K} \geq \tau^K\right] \\ \leq& \tau^{-K}\mathbb{E}\left[|\langle v,h\rangle|^{2K}\right] \\ =& \tau^{-k}\binom{D+K-1}{K}^{-1} \leq \left(\frac{K}{D\tau}\right)^K .\end{aligned}$$

Here, we have made use of the assumption that the $K$-design approximation is perfect ($\epsilon = 0$). This allows us to directly recycle the exact Haar integration from Eq. (5.3). The claim then readily follows from retracing the steps of the previous proof, but with this weaker polynomial concentration formula. ■

## 5.4 Proof part 3: connection to local random circuits

Proposition 5.9 highlights that the complexity of a randomly selected $K$-design state increases linearly with $K$ until a certain threshold is met ($K \approx D = d^n$). But, so far, $K$-designs have ben a rather abstract concept. The following deep result allows us to relate $K$-designs to random circuits of increasing size.

**Fact 5.11 (local random circuits generate $k$-designs [Haf22]).** Local random circuits of size $T = O\left(nK^{4+o(1)}\left(nK + \log(1/\epsilon)\right)\right)$ produce state ensembles $U|\psi_0\rangle$ that form $\epsilon$-approximate $K$-designs. ■

local random circuits form $K$-designs

This is a very recent and substantial improvement of a seminal result by Brandão, Horodecki and Harrow [BHH16]. Our main result is now an immediate consequence of Proposition 5.9 and Fact 5.11. The detailed conversion is a bit cumbersome, but here is the main gist.

**Theorem 5.12 ('polynomial' growth in state complexity, formal).** Fix $|\psi_0\rangle \in \mathbb{C}^D$ and let $U$ be a local random circuit of size $T$. Then, with high probability

$$C_\delta\left(U|\psi_0\rangle\right) = \Omega\left(T^{1/(4+o(1))}\right),$$

where we have suppressed the dependence on $n$ (nr. of qu$d$its) and $\delta$. This growth persists up to exponential circuit sizes $T \approx \sqrt{D} = d^{n/2}$.

'polynomial' state complexity growth (formal)

It is also possible to turn this probabilistic statement into a quantitative bound on the minimal number of high-complexity states that have this property. It must grow exponentially in circuit depth as $\exp\left(\Omega(T^{1/(5+o(1))}\right)$. The trick is to exploit the fact that the weight distribution of a $K$-design cannot be too spiky. This then implies a direct relation between the probability of producing a high complexity state and the minimal number of high-complexity states within the entire ensemble.

## 5.5 Generalizations and follow-up work

The proof technique introduced above is very versatile and can be adjusted to cover stronger notions of complexity as well. Circuit complexity is one such example. We say that a unitary $U \in \mathsf{U}(D)$ has $\delta$-complexity at most $R$ if

formal definition of circuit complexity

$$\min_{V \in \mathsf{V}_R} \left\| U \cdot U^\dagger - V \cdot V^\dagger \right\|_\diamond \leq \delta. \tag{5.4}$$

Here, $\| \cdot \|_\diamond$ denotes the diamond distance of the two unitary channels involved. If Eq. (5.4) holds, we write $C_\delta(U) \leq R$. It should be not surprising at this point that circuit complexity also grows with circuit size.

**Theorem 5.13 ('polynomial' growth in state complexity, informal).** 'Most' local random circuits of size $T$ produce unitaries with complexity (at least) $\Omega\left(T^{1/(5+o(1))}\right)$. This growth persists up to circuit sizes $T \approx \sqrt{D} = d^{n/2}$.

'polynomial' circuit complexity growth

To prove this claim, it is helpful to first relate the diamond distance to another property that is easier to control:

$$\|U \cdot U^\dagger - V \cdot V^\dagger\|_\diamond \leq \delta \quad \Rightarrow \quad \left|\mathrm{tr}\left(V^\dagger U\right)\right|^2 \geq D^2(1-\delta^2).$$

This necessary condition is much easier to control. In particular, we can use more general Haar integration techniques to bound moments and deduce polynomial concentration bounds. The rest of the proof is then almost identical to the state complexity case.

Next, we want to point out that it is possible to introduce *stronger/operational complexity notions* for both states and unitary circuits. These are based on the operational task of distinguishing the state/unitary in question from the most useless state/channel conceivable. For states this is the maximally mixed state $\tau = \mathbb{I}/D$, while for channels this is the completely depolarizing channel $\mathcal{D}(\rho) = \mathrm{tr}(\rho)\tau$. In both cases, the optimal single-shot distinguishability protocols are known. They give rise to the trace distance $\||\psi\rangle\langle\psi| - \tau\|_1$ and the diamond distance $\|U \cdot U^\dagger - \mathcal{D}\|_\diamond$, respectively. But achieving these optimal values requires measurement procedures whose complexity mimics that of the state/unitary in question. This allows us to indirectly capture complexity by limiting the circuit size allowed for executing distinguishing measurements. A formal definition would go beyond the scope of this talk. Instead, we refer to Figure 5.2 (state complexity) and Figure 5.3 (circuit complexity) for visual illustrations. These stronger/operational notions of complexity imply the ones used so far, but the converse is not necessarily true, as the following example shows.

stronger/operational definitions of complexity

**Example 5.14** Let $|h\rangle$ be a Haar-random state on $(n-1)$ qu$d$its and define the $n$-qu$d$it state $|\psi\rangle = |h\rangle \otimes |0\rangle$. Then, this state has exponential state complexity according to Definition 5.3. But it is actually very easy to distinguish $|\psi\rangle$ from $\tau = \mathbb{I}/D$. A computational basis measurement on the last qubit (and ignoring everything else) does the job with reasonable probability – especially if the local dimension $q$ is large. ■

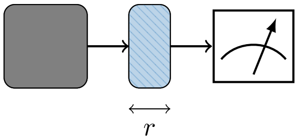

**Figure 5.2** *Pictographic illustration of strong state complexity.* A black-box either outputs a (known) pure state $\rho = |\psi\rangle\langle\psi|$, or the maximally mixed state $\rho_0 = \frac{1}{d}\mathbb{I}$. The task is to correctly guess which one it produced by applying a pre-processing circuit $V$ (blue line pattern) of limited size $r$ and performing a simple measurement (right). We say that $|\psi\rangle$ has *strong/operational state complexity* at most $r$ if the probability of correctly distinguishing both possibilities is close to optimal.

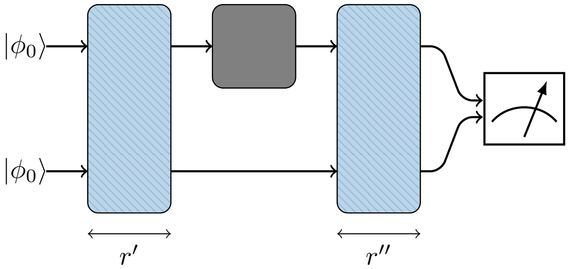

**Figure 5.3** *Pictographic illustration of strong circuit complexity.* A black box (center) takes quantum states as inputs and applies either a unitary channel $\mathcal{U}(\rho) = U\rho U^\dagger$, or the depolarizing channel $\mathcal{D}(\rho) = \tau = \mathbb{I}/D$. The task is to correctly guess which evolution occurred. The rules of the game allow short pre- and post-processing circuits (blue line patterns) that may involve a quantum memory. The final guess must be based on a simple measurement (right). We say that $U$ has *strong/operational circuit complexity* at most $r = r' + r''$ if the probability of correctly distinguishing both options is close to optimal.

This feature of strong/operational complexity delays the onset of complexity growth up to circuit sizes that cover all qu$d$its involved. Such a behavior accurately addresses physical effects like operator growth and the switchback effect in holography.

We conclude with a beautiful result by Haferkamp *et al.* that proves Conjecture 5.2 for certain random circuit families acting on $n$ qubits ($d = 2$).

## Bibliography

[BHH16] F. G. S. L. Brandão, A. W. Harrow, and M. Horodecki. "Local Random Quantum Circuits are Approximate Polynomial-Designs". In: *Communications in Mathematical Physics* 346.2 (2016), pages 397–434. DOI: 10.1007/s00220-016-2706-8. URL: https://doi.org/10.1007/s00220-016-2706-8.

[Bra+21] F. G. Brandão et al. "Models of Quantum Complexity Growth". In: *PRX Quantum* 2 (3 2021), page 030316. DOI: 10.1103/PRXQuantum.2.030316. URL: https://link.aps.org/doi/10.1103/PRXQuantum.2.030316.

[BS18] A. R. Brown and L. Susskind. "Second law of quantum complexity". In: *Phys. Rev.* D97 (2018), page 086015. DOI: 10.1103/PhysRevD.97.086015. arXiv: 1701.01107 [hep-th].

[Haf22] J. Haferkamp. *Random quantum circuits are approximate unitary $t$-designs in depth $O\left(nt^{5+o(1)}\right)$*. 2022. DOI: 10.48550/ARXIV.2203.16571. URL: https://arxiv.org/abs/2203.16571.

[Haf+22] J. Haferkamp et al. "Linear growth of quantum circuit complexity". In: *Nature Physics* 18.5 (2022), pages 528–532. DOI: 10.1038/s41567-022-01539-6. URL: https://doi.org/10.1038/s41567-022-01539-6.

[Kue19] R. Kueng. *Quantum and Classical Information Processes with Tensors (lecture notes)*. Caltech course notes: https://iqim.caltech.edu/classes. 2019.

# Bibliography

[BHH16] F. G. S. L. Brandão, A. W. Harrow, and M. Horodecki. “Local Random Quantum Circuits are Approximate Polynomial-Designs”. In: *Communications in Mathematical Physics* 346.2 (2016), pages 397–434. DOI: 10.1007/s00220-016-2706-8. URL: https://doi.org/10.1007/s00220-016-2706-8.

[Bra+21] F. G. Brandão et al. “Models of Quantum Complexity Growth”. In: *PRX Quantum* 2 (3 2021), page 030316. DOI: 10.1103/PRXQuantum.2.030316. URL: https://link.aps.org/doi/10.1103/PRXQuantum.2.030316.

[BS18] A. R. Brown and L. Susskind. “Second law of quantum complexity”. In: *Phys. Rev.* D97 (2018), page 086015. DOI: 10.1103/PhysRevD.97.086015. arXiv: 1701.01107 [hep-th].

[Haf22] J. Haferkamp. *Random quantum circuits are approximate unitary $t$-designs in depth $O\left(nt^{5+o(1)}\right)$*. 2022. DOI: 10.48550/ARXIV.2203.16571. URL: https://arxiv.org/abs/2203.16571.

[Haf+22] J. Haferkamp et al. “Linear growth of quantum circuit complexity”. In: *Nature Physics* 18.5 (2022), pages 528–532. DOI: 10.1038/s41567-022-01539-6. URL: https://doi.org/10.1038/s41567-022-01539-6.

[Hel69] C. W. Helstrom. “Quantum detection and estimation theory”. In: *J. Statist. Phys.* 1 (1969), pages 231–252. ISSN: 0022-4715. DOI: 10.1007/BF01007479. URL: https://doi.org/10.1007/BF01007479.

[Hol73] A. S. Holevo. “Optimal quantum measurements”. In: *Teoret. Mat. Fiz.* 17 (1973), pages 319–326. ISSN: 0564-6162.

[Kli+16] M. Kliesch et al. “Improving Compressed Sensing With the Diamond Norm”. In: *IEEE Trans. Inform. Theory* 62.12 (2016), pages 7445–7463. DOI: 10.1109/TIT.2016.2606500.

[Kue19] R. Kueng. *Quantum and Classical Information Processes with Tensors (lecture notes)*. Caltech course notes: https://iqim.caltech.edu/classes. 2019.

[Mic+18] U. Michel et al. "Comments on "Improving Compressed Sensing With the Diamond Norm"–Saturation of the Norm Inequalities Between Diamond and Nuclear Norm". In: *IEEE Trans. Inform. Theory* 64.11 (2018), pages 7443–7445. DOI: 10.1109/TIT.2018.2861887.

[Wat11] J. Watrous. *Theory of Quantum Information (lecture notes)*. 2011.