# Algorithmic aspects of Optimal Channel Coding

Paul Fermé
Adviser: Omar Fawzi

*Laboratoire de l'Informatique du Parallélisme, École Normale Supérieure de Lyon*

## Contents

# 1 Channel Coding Function Properties

Instead of the limit quantity known as the capacity of a channel, introduced by Shannon in 1948 for classical channels [21], we rather study the best probability we can achieve with one channel and how to efficiently compute the code.

## 1.1 Definition and proven properties

We study classical-quantum channels, ie. a finite set $W$ of density operators $W_x$ indexed by $X$, which are trace 1 semidefinite positive complex matrices.

**Definition 1.1** (Channel Coding Function).

$$\begin{aligned} f_W(S) := \quad & \underset{\Lambda}{\text{maximize}} && \sum_{x \in S} \mathrm{Tr}(\Lambda_x W_x) \\ & \text{subject to} && \sum_{x \in S} \Lambda_x = \mathbb{1} \\ & && \Lambda_x \succcurlyeq 0, \forall x \in S \end{aligned} \tag{1}$$

**Property 1.2.** We can replace $\sum\limits_{x \in S} \Lambda_x = \mathbb{1}$ by $\sum\limits_{x \in S} \Lambda_x \preccurlyeq \mathbb{1}$

**Property 1.3** (Strong Duality, since there always exists strictly interior solutions).

$$\begin{aligned} f_W(S) = \quad & \underset{\rho}{\text{minimize}} && \mathrm{tr}(\rho) \\ & \text{subject to} && \rho \succcurlyeq W_x, \forall x \in S \end{aligned} \tag{2}$$

**Property 1.4** (Some properties by [3, 5]). Let $\{\Lambda_x\}_{x \in S}$ and $\rho$ be optimal solutions of the primal and the dual:

(1) $\rho$ is unique, although $\{\Lambda_x\}_{x \in S}$ is not.

(2) $\rho = \sum\limits_{x \in S} \Lambda_x W_x$

(3) $\forall i, j \in S, \Lambda_i (W_i - W_j) \Lambda_j = 0$ and in particular $(\rho - W_j)\Lambda_j = 0$

(4) If $d$ is the dimension of our system, then there is an optimal measurement of size smaller or equal to $d^2$.

(5) Let $\sigma_i$ a state and $r_i \geq 0$ such that $r_i \sigma_i = \rho - W_i \succcurlyeq 0$, then:

$$\forall i \in S, r_i = \|\rho - W_i\|_1 = r = \frac{\|W_j - W_k\|_1}{\|\sigma_j - \sigma_k\|_1}$$

which is the ratio of the congruent polytopes defined by points $\{W_i\}_{i \in S}$ and $\{r\sigma_i\}_{i \in S}$ and we have that:

$$f_W(S) = \mathrm{tr}(\rho) = 1 + \frac{1}{|S|} \sum_{i \in S} r_i = 1 + r$$

**Definition 1.5** (Channel Coding Optimization).

$$S(W, k) := \frac{1}{k} \max_{S \subseteq X, |S| \leq k} f_W(S)$$

$S(W,k)$ corresponds to the probability of success of the best strategy of decoding over the channel $W$ sending $k$ classical messages (so $\log_2(k)$ bits)

*Remark.*

$$S(W,k) = \frac{1}{k} \max_{S \subseteq X, |S|=k} f_W(S)$$

**Definition 1.6** (Marginal value)**.** Let $f$ a set function. The marginal value of $A$ having $B$ is defined by:

$$f(A|B) := f(A \cup B) - f(B)$$

**Definition 1.7** (Submodularity, Fractional Subadditivity (XOS))**.** Let $f : 2^X \to \mathbb{R}^+$ a nonnegative set function.

(1) $f$ is said to be *submodular* if

$$\forall S \subseteq T \subseteq X \text{ and } x \notin T, f(x|S) \geq f(x|T)$$

(2) $f$ is said to be *fractionally subadditive or XOS* if

$$f(A) \leq \sum_i \beta_i f(B_i) \text{ with } \beta_i \geq 0 \text{ and } \sum_{i:B_i \ni a} \beta_i \geq 1, \forall a \in A$$

*Remark.* Submodularity implies fractional subadditivity.

**Theorem 1.8** ([20])**.** If $f$ is nonnegative, nondecreasing and submodular, then the greedy algorithm described in page 8 outputs $S_{\text{greedy}}$ of size $k$ approximating $S_{\text{opt}}$ such that:

$$\begin{aligned} f(S_{\text{greedy}}) \geq & \left(1 - \left(1 - \frac{1}{k}\right)^k\right) f(S_{\text{opt}}) \\ \geq & \left(1 - \frac{1}{e}\right) f(S_{\text{opt}}) \end{aligned} \tag{3}$$

**Theorem 1.9** ([22, 1])**.** If $f$ is a nondecreasing, nonpositive and XOS set function we can access only with *value queries*, then for $\epsilon > 0$, there does not exist any polynomial $n^{\frac{1}{2}-\epsilon}$-approximation algorithm solving the maximum value under cardinality constraints problem.

*Proof.* □

**Property 1.10** (Easy properties)**.** (1) Positive: $f_W(S) \geq 0$

(2) Monotone: $f_W(S \cup \{x\}) \geq f_W(S)$

(3) Small Marginal Values: $0 \leq f_W(x|S) \leq 1$ (see [23] for applications)

(4) Upper Bound: $f_W(S) \leq \min(|S|, d)$

(5) Fractionally Subadditive (XOS): $f_W$ is XOS

*Proof.* (1) Taking the dual formulation of $f_W$, since $\rho \succcurlyeq W_x$ for some $x \in S$ and $\text{tr}(W_x) = 1$, then $\text{tr}(\rho) \geq \text{tr}(W_x) = 1 \geq 0$ and so does $f_W$.

(2) If $\Lambda$ is a solution of the maximization in $f$ for the set $S$, it is also for the set $S \cup \{x\}$ if we take $\Lambda_x = 0$, so $f_W(S \cup \{x\}) \geq f_W(S)$.

(3) $f_W(x|S) \geq 0$ is a restatement of the monotonicity of $f_W$.
On the other hand, we have that $0 \leq \text{Tr}(\Lambda_x W_x) \leq \text{Tr}(W_x) = 1$ since $0 \preccurlyeq \Lambda_x \preccurlyeq \mathbb{1}$ and $W_x \succcurlyeq 0$, and so $f_W(x|S) \leq 1$.

(4) We apply $|S|$ times the previous marginal bound to rebuild $S$ and get the bound on $|S|$. For the bound on $d$, we look at the dual on we have then that $\mathbb{1}$ is always a solution of this problem, and it is of trace $d$.

(5) Let $\{\Lambda_a\}_{a\in A}$ be on optimal measurement for $A$. Then $\{\Lambda_a\}_{a\in B_i\cap A}$ is an acceptable solution of $B_i$, so $\sum\limits_{a\in B_i\cap A} \mathrm{Tr}(\Lambda_a W_a) \leq f_W(B_i)$. Thus:

$$\begin{aligned}
\sum_i \beta_i f_W(B_i) &\geq \sum_i \beta_i \sum_{a\in B_i\cap A} \mathrm{Tr}(\Lambda_a W_a) \\
&= \sum_{a\in A}\Big(\sum_{i:B_i\ni a}\beta_i\Big)\mathrm{Tr}(\Lambda_a W_a) \\
&\geq \sum_{a\in A}\mathrm{Tr}(\Lambda_a W_a) = f_W(A)
\end{aligned} \tag{4}$$

□

**Property 1.11** (On submodularity)**.** (1) Submodularity: When $W$ is classical (ie. $W_x$ diagonals), then $f_W$ is submodular. Also, if $W_x$ are of dimension 2, real and pure (rank 1), then $f_W$ is submodular.

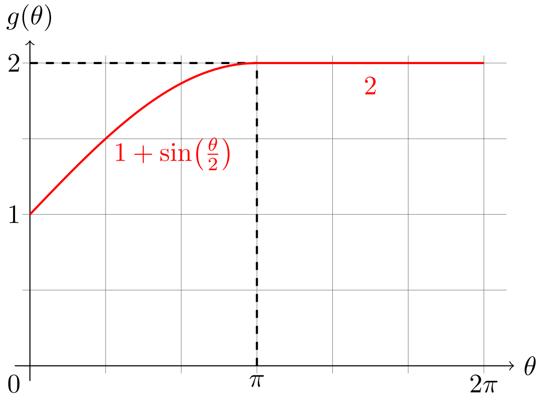
(2) No submodularity: As soon as you take $W_x$ to be complex and pure of dimension 2, then there exists $S \subseteq T$ and $x \notin T$ such that:

$$f_W(x|S) = 0 \text{ and } f_W(x|T) > 0$$

*Proof.* (1) For the classical case see [6].

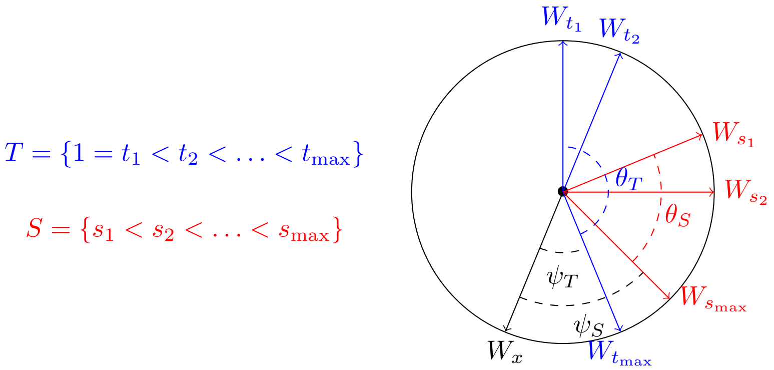
In the real, pure and dimension 2 case, we have the following property for a set of states $\{W_x = |\psi_x\rangle\langle\psi_x|\}_{x\in S}$. Let $\theta_S$ be the minimum angle such that all states of $S$ can be put in a slice of the Bloch circle of radius $\theta$. Then $f_W(S) = g(\theta_S)$ where:

$$g(\theta) = \begin{cases} 1+\sin\left(\frac{\theta}{2}\right) \text{ if } 0\leq\theta\leq\pi \\ 2 \text{ if } \pi\leq\theta\leq 2\pi \end{cases}$$



This is a consequence of results by [3]. We had done a direct proof of this, but it is not needed anymore since we have found a more general framework in section 3.5.2 where we can deduce this value. It will be proved in the corollary 3.28 page 23.

Since $g$ is concave, we have that $f_W$ is submodular:
Let $S \subseteq T$ and $x \notin T$, and let us order the $W_u$ in clockwise order and rename them with integers in $\{1, \dots, n\}$:



If $t_1 \leq x \leq t_{\max}$, then the angle is unchanged: $\theta_{T \cup \{x\}} = \theta_T$, and so $f(T \cup \{x\}) = f(T)$, and in particular we have the submodular property for those instances.

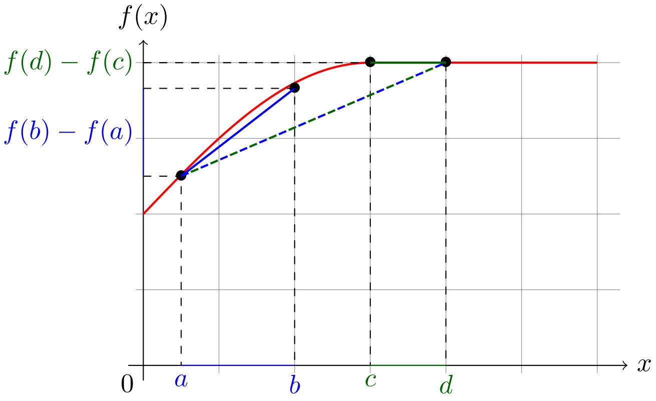
Otherwise, $x > t_{\max}$. We suppose that the angle $\psi_T$ between $W_x$ and $W_{t_{\max}}$ is smaller than the angle $\psi'_T$ between $W_x$ and $W_{t_1}$ (the analysis will be symmetric). Thus, we have that $\theta_{T \cup \{x\}} = \theta_T + \psi_T$.

If the angle $\psi_S$ between $W_x$ and $W_{s_{\max}}$ is smaller than the angle $\psi'_S$ between $W_x$ and $W_{s_1}$, then $\theta_{S \cup \{x\}} = \theta_S + \psi_S$. By definition:

$$\begin{aligned} f(x|T) &= g(\theta_{T \cup \{x\}}) - g(\theta_T) \\ f(x|S) &= g(\theta_{S \cup \{x\}}) - g(\theta_S) \end{aligned} \tag{5}$$

**Lemma 1.12.** If $f$ is concave, we have that for all $a, b, c, d$ s.t. $c \geq a$ and $b - a \geq d - c \geq 0$:

$$f(b) - f(a) \geq f(d) - f(c)$$



*Proof.* Indeed, we have that $a \leq b \leq d$ and $a \leq c \leq d$. When we fix one point, since $f$ is concave, its slope is decreasing, so:

$$\frac{f(b) - f(a)}{b - a} \geq \frac{f(d) - f(a)}{d - a} \text{ and } \frac{f(d) - f(a)}{d - a} \geq \frac{f(d) - f(c)}{d - c}$$

So

$$\frac{f(b)-f(a)}{b-a} \geq \frac{f(d)-f(c)}{d-c}$$

But $b-a \geq d-c$ and thus we get that $f(b)-f(a) \geq \frac{b-a}{d-c}(f(d)-f(c)) \geq f(d)-f(c)$. □

We have that $\theta_{S\cup\{x\}} - \theta_S = \psi_S \geq \psi_T = \theta_{T\cup\{x\}} - \theta_T$ and $\theta_T \geq \theta_S$, so applying the lemma 1.12:

$$g(\theta_{S\cup\{x\}}) - g(\theta_S) \geq g(\theta_{T\cup\{x\}}) - g(\theta_T)$$

ie. $f(x|S) \geq f(x|T)$: we have the submodular property.

Otherwise the angle $\psi'_S$ between $W_x$ and $W_{s_1}$ is strictly smaller than the angle $\psi_S$ between $W_x$ and $W_{s_{\max}}$. Then $\theta_{S\cup\{x\}} = \theta_S + \psi'_S$. But $\psi'_S \geq \psi'_T \geq \psi_T$ by hypothesis. So we can do the same analysis as before by replacing $\psi_S$ by $\psi'_S$. QED.

(2) Some of it was done in [14] (dimension 4 with real matrices).

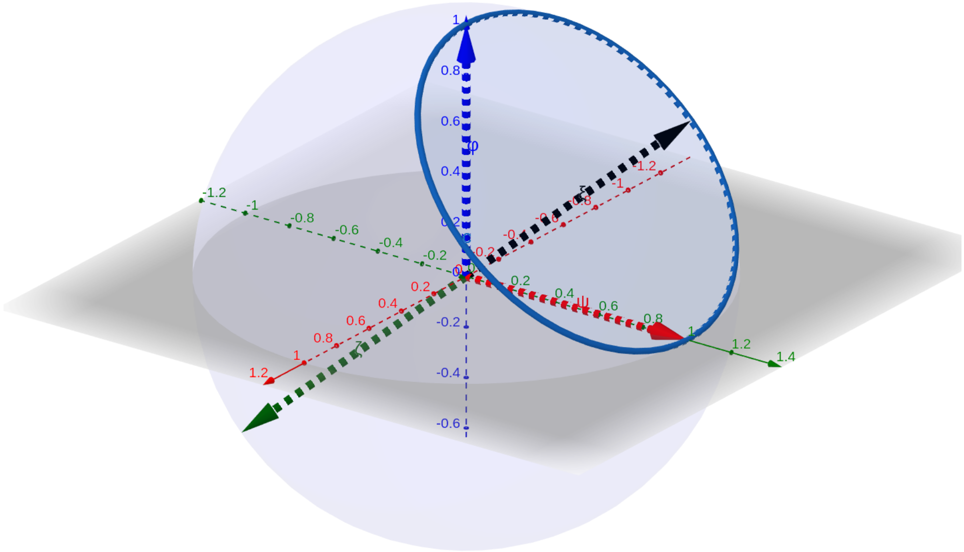


Figure 1: A counter-example of submodularity in the Bloch sphere for pure states

Let us look now at the dimension 2 complex case. We define the following vectors (and their coordinates in the Bloch sphere), which are represented in figure 1 page 6:

$$\begin{aligned} |\varphi\rangle = |0\rangle = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \qquad & |\psi\rangle = |+\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \\ |\xi\rangle = \frac{1}{2}\begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix} \qquad & |\zeta\rangle = \frac{1}{2}\begin{pmatrix} -1 \\ \sqrt{2} \\ -1 \end{pmatrix} \end{aligned} \tag{6}$$

Then we take $W = \{W_1, W_2, W_3, W_4\} = \{|\varphi\rangle\langle\varphi|, |\psi\rangle\langle\psi|, |\xi\rangle\langle\xi|, |\zeta\rangle\langle\zeta|\}$ and $S = \{1,2\}, T = S \cup \{3\}, x = 4$.

We have that $f_W(S) = 1 + \frac{\sqrt{2}}{2}$ using the value of $f_W$ for two vectors in the Bloch circle. Also, $f_W(T) < 2$ ($f_W(T) \simeq 1.976$) since vectors are strictly in one hemisphere (we will see this result in section 3.5.2, also this was shown in [16]). With the upcoming result, we will show that $f_W(S \cup \{x\}) = f_W(S) = 1 + \frac{\sqrt{2}}{2}$. Finally, $f_W(T \cup \{x\}) = 2$ since it contains two orthogonal vectors $|\xi\rangle$ and $|\zeta\rangle$. Thus:

$$f_W(x|S) = 0 < f_W(x|T) \simeq 0.024$$

and so we have found our counterexample.

## 1.2 Other properties investigated

**Definition 1.13** (Approximate Submodularity, Weak Submodularity)**.** Let $f : 2^X \to \mathbb{R}^+$ a nonnegative set function.

(1) $f$ is said to be *$\epsilon$-approximately submodular* [18, 17] if

$$\forall S \subseteq T \subseteq X \text{ and } x \notin T, f(x|S) \geq f(x|T) - \epsilon$$

(2) $f$ is said to be *$\gamma$-weakly submodular* [12, 8] if

$$\forall \Omega, S, \sum_{\omega \in \Omega - S} f(\omega|S) \geq \gamma f(\Omega|S)$$

We say that $\gamma_f$, the largest $\gamma$ such that $f$ is $\gamma$-weakly submodular, is the *submodular ratio* of $f$.

**Theorem 1.14** ([18])**.** If $f$ is nonnegative, nondecreasing and $\epsilon$-approximately submodular, then the greedy algorithm described in page 8 outputs $S_{\text{greedy}}$ of size $k$ approximating $S_{\text{opt}}$ such that:

$$\begin{aligned} f(S_{\text{greedy}}) \geq& \Big(1 - \Big(1 - \frac{1}{k}\Big)^k\Big) f(S_{\text{opt}}) - k\epsilon \\ \geq& \Big(1 - \frac{1}{e}\Big) f(S_{\text{opt}}) - k\epsilon \end{aligned} \tag{7}$$

**Property 1.15** ([8])**.** If $f$ is nondecreasing, then:

(1) $\gamma_f \in [0, 1]$

(2) $f$ is submodular $\Leftrightarrow \gamma_f = 1$

**Theorem 1.16** ([8])**.** If $f$ is nonnegative, nondecreasing and $\gamma$-weakly submodular, then the greedy algorithm described in page 8 outputs $S_{\text{greedy}}$ of size $k$ approximating $S_{\text{opt}}$ such that:

$$\begin{aligned} f(S_{\text{greedy}}) \geq& \Big(1 - \Big(1 - \frac{\gamma}{k}\Big)^k\Big) f(S_{\text{opt}}) \\ \geq& \Big(1 - \frac{1}{e^\gamma}\Big) f(S_{\text{opt}}) \end{aligned} \tag{8}$$

A notion of curvature is also described in [8], but useless for us as it seems that our function has always the worst case curvature.

Mixing both properties, with small $\epsilon$ we get $\gamma_f$ not so far from 1. It should be easy to prove that both bounds combine well. Also, using only one of those properties seems to fail.

# 2 Greedy algorithm

We study now properly the problem of computing $S(W,k)$, adn finding the code $S$ that corresponds to this. As shown in [6], this problem is NP-complete already in the classical case and is even $(1-e^{-1})$-hard to approximate, and this approximation is obtained by the following simple greedy algorithm:

```
Algorithm 1: Greedy algorithm
  Input: $k \in \{1, \dots, n\}$
  Output: An approximation of $S(W,k)$: A set $S \subseteq X$ of size at most $k$ that is
          *close* to maximize $f_W(S)$ over all of those sets
  Data: $W = \{W_1, \dots, W_n\}$
  $S = \emptyset$
  for $i \in \{1, \dots, k\}$ do
      $x^* = \operatorname*{argmax}_{x \in X - S} f_W(x|S)$
      $S = S \cup \{x^*\}$
  return $S$
```

## 2.1 Classical case

In fact, this algorithm in the classical case gives a $(1-(1-\frac{1}{k})^k)$-approximation for the problem of size $k$ since in that case $f_W$ is submodular [6]. We have found in particular a series of explicite examples where this bound is obtained, which shows the tightness of this approximation:

For $k=2$, this is obtained by the matrix $W^2$ (where line $i$ corresponds to the diagonal coefficients of $W_i^2$, since we are in the classical case):

$$W^2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & \\ & 1 \end{pmatrix}$$

The greedy algorithm gives the set $S_{\text{greedy}} = \{1,2\}$ whose value is $f_{W^2}(S_{\text{greedy}}) = 3/2$ as answer, whereas the optimal value is $f_{W^2}(S_{\text{opt}} = \{2,3\}) = 2$. This gives an approximation ratio of:

$$\frac{f_{W^2}(S_{\text{greedy}})}{f_{W^2}(S_{\text{opt}})} = \frac{3}{4} = 1 - \left(1 - \frac{1}{2}\right)^2$$

For $k=3$:

$$W^3 = \begin{pmatrix} \frac{1}{3} & & \frac{1}{3} & & \frac{1}{3} & \\ \frac{1}{3} & \frac{2}{9} & & \frac{2}{9} & & \frac{2}{9} \\ \frac{1}{2} & \frac{1}{2} & & & & \\ & & \frac{1}{2} & \frac{1}{2} & & \\ & & & & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

We get:

$$\frac{f_{W^3}(S_{\text{greedy}} = \{1,2,3\})}{f_{W^3}(S_{\text{opt}} = \{3,4,5\})} = \frac{19/9}{3} = \frac{19}{27} = 1 - \left(1 - \frac{1}{3}\right)^3$$

These first two cases were found by Daniel Szylagyi. We have generalized those examples for all $k$. We need the dimension $d = \frac{k(k-1)}{2}$ (number of columns) and the number of states $n = 2k-1$ (number of lines). Then we define the matrix $W^k = \begin{pmatrix} G \\ O \end{pmatrix}$ in the following way with:

$$G = \begin{pmatrix} U & V & \dots & V \end{pmatrix}$$

where V is repeated $(k-1)$ times and $U, V$ of size $(k-1) \times k$ defined by:

$$U = \begin{pmatrix} u_1^k & 0 & \dots & 0 \\ u_1^k & u_2^k & \ddots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u_1^k & u_2^k & \dots & u_{k-1}^k \end{pmatrix}$$

$$V = \begin{pmatrix} u_1^k & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & u_{k-1}^k \end{pmatrix}$$

and where the sequence $(u_j^k)_{j \in \mathbb{N}^*}$ is defined in the following way:

$$(u_j^k)_{j \in \mathbb{N}^*} : \begin{cases} u_1^k = \frac{1}{k} \\ u_{j+1}^k = u_j^k + \frac{1-u_j^k}{k} \end{cases}$$

and we define $O$:

$$O = \begin{pmatrix} \frac{1}{k-1} & \dots & \frac{1}{k-1} & 0 & & \dots & \dots & & 0 \\ 0 & \dots & 0 & \frac{1}{k-1} & \dots & \frac{1}{k-1} & 0 & \dots & 0 \\ \vdots & & \vdots & & \ddots & & \vdots & & \vdots \\ 0 & & \dots & \dots & & 0 & \frac{1}{k-1} & \dots & \frac{1}{k-1} \end{pmatrix}$$

where O has $k$ lines and each $\frac{1}{k-1}$ is repeated $k-1$ times.

We can show by induction on $j$ that:

$$u_j^k = \frac{\sum_{i=0}^{j-1} (-1)^{j-1-i} \binom{j}{i} k^i}{k^j}$$

and:

$$\frac{f_{W^k}(\{1, \dots, j\})}{k} = u_j^k$$

We deduce from this that:

$$\begin{aligned} \frac{f_{W^k}(S_{\text{greedy}} = \{1, \dots, k\})}{f_{W^k}(S_{\text{opt}} = \{k, \dots, 2k-1\})} &= \frac{f_{W^k}(\{1, \dots, k\})}{k} = u_k^k \\ &= \frac{\sum_{i=0}^{k-1} (-1)^{k-1-i} \binom{k}{i} k^i}{k^k} \\ &= \frac{k^k - (k-1)^k}{k^k} = 1 - \left(1 - \frac{1}{k}\right)^k \end{aligned} \tag{9}$$

For now, we do not write here precise proofs which are technical but not difficult.

## 2.2 Pure real quantum case

### 2.2.1 NP-hardness and inapproximability

We show in this section that the problem we are trying to solve with the greedy algorithm is NP-hard, even restricted to pure real states. However, the NP-hardness comes for different reasons than in the classical case.

**Theorem 2.1.** The following decision problem is NP-hard:

**Problem:** *PureRealChannelCoding*

**Instance:** A finite set of $W$ of pure real states $W_i$ (real, rank one states) of size $n$, $M \in \mathbb{R}$ and $k \in \{1, \dots, n\}$

**Question:** Does $S(W, k) \geq M$ ?

*Proof.* Inspired by a proof in [11], we give a reduction from *Exact cover by 3-sets* which is known to be NP-complete (see for instance [15]):

**Problem:** *Exact cover by 3-sets (X3C)*

**Instance:** A set $Q$ and a collection $C$ of 3-elements subsets of $Q$

**Question:** Does there exists an exact cover for $Q$, i.e. a sub-collection $S$ of indices of $C$ such that every element in $Q$ appears *exactly* once in $\{c_x\}_{x \in S}$ ?

We use the following reduction. Let $Q = \{q_1, \dots, q_m\}$ and $C = \{c_1, \dots, c_n\}$. We take $k = \lceil \frac{m}{3} \rceil$, $M = 1$, and $W = \{A^{(j)}(A^{(j)})^T, j \in \{1, \dots, n\}\}$, where $A^{(j)}$ is the $j$-th column of $A$ which is a $m \times n$ real matrix defined by:

$$A_{i,j} = \begin{cases} \frac{1}{\sqrt{3}} \text{ if } q_i \in c_j \\ 0 \text{ otherwise} \end{cases}$$

First, each $A^{(j)}$ is indeed a unit vector of $\mathbb{R}^m$ since for each $j$, $c_j$ contains exactly 3 elements of $Q$ since we have taken a valid instance of *X3C*. Thus we have defined a valid instance of *PureRealChannelCoding*.

Let us show that there exists an exact cover for $Q$ if and only if $S(W, k) \geq M$.

($\Rightarrow$) If there is an exact cover $S$ for $Q$, then $|S| = \frac{m}{3} \in \mathbb{N}$. We have then that $|S| = k$ and we will show that $f_W(S) \geq k$, which implies that $S(W, k) \geq \frac{f_W(S)}{k} = 1 = M$. Indeed, since $S$ is an exact cover, there is exactly one non-zero coefficient in each line of $\{A^{(j)}, j \in S\}$ seen as a matrix. Thus, we take as POVM $(\Lambda_j)_{j \in S}$ defined by:

$$\Lambda_j = \sum_{i : A^{(j)} \neq 0} |i\rangle\langle i| = \sum_{i : q_i \in c_j} |i\rangle\langle i|$$

which is in fact a PVM. Indeed, $\Lambda_j \succcurlyeq 0$ and:

$$\sum_{j \in S} \Lambda_j = \sum_{j \in S} \sum_{i : q_i \in c_j} |i\rangle\langle i| = \sum_{j \in Q} |i\rangle\langle i| = \mathbb{1}$$

since $S$ is an exact cover for $Q$.

Thus:

$$\begin{aligned} f_W(S) \geq & \sum_{j \in S} \mathrm{Tr}(\Lambda_j W_j) = \sum_{j \in S} \mathrm{Tr}\left( \sum_{i : A^{(j)} \neq 0} |i\rangle\langle i| \, A^{(j)}(A^{(j)})^T \right) \\ = & \sum_{j \in S} \sum_{i : A_i^{(j)} \neq 0} \mathrm{Tr}\Big( |i\rangle\langle i| \, A^{(j)}(A^{(j)})^T \Big) = \sum_{j \in S} \mathrm{Tr}\Big( A^{(j)}(A^{(j)})^T \Big) \qquad (10) \\ = & \sum_{j \in S} 1 = |S| = k \end{aligned}$$

($\Leftarrow$) Suppose that $S(W,k) \geq M$. We have in fact that $S(W,k) = M$. Indeed for all $S$ of size at most $k$, we have that $f_W(S) \leq k$ as seen in property 1.10, and so $S(W,k) \leq 1 = M$.

Let $S$ be a set such that $f_W(S) = k$ (and so $|S| = k$). To obtain such a value, necessarily for all $j$ in $S$, $\Lambda_j$ needs to be at least a projector on $W_j$, otherwise we have that $\mathrm{Tr}(\Lambda_j W_j) < \mathrm{Tr}(W_j) = 1$. Thus:

$$\Lambda_j \succcurlyeq \Pi_{\mathrm{vect}\{A^{(j)}\}} = \Pi_{\mathrm{vect}\{|i\rangle, i:q_i \in c_j\}} = \sum_{i:q_i \in c_j} |i\rangle\!\langle i|$$

And so:

$$\mathbb{1} = \sum_{j \in S} \Lambda_j \succcurlyeq \sum_{j \in S} \sum_{i:q_i \in c_j} |i\rangle\!\langle i|$$

but $S$ is of size $k = \lceil \frac{m}{3} \rceil$, and each $c_j$ is of size 3, so there are $3 * \lceil \frac{m}{3} \rceil$ in that previous sum. If $\frac{m}{3} \notin \mathbb{N}$, then $3 * \lceil \frac{m}{3} \rceil > m$, which is a contradiction with the previous inequality since $(|i\rangle)_{i \in [m]}$ is a basis of $\mathbb{R}^m$. Thus $\frac{m}{3} \in \mathbb{N}$, there are exactly $m$ terms in the previous sum and we have the equality $\Lambda_j = \sum_{i:q_i \in c_j} |i\rangle\!\langle i|$, which means that $S$ is an exact cover of $Q$ since there can't be any overlap.

□

A consequence of the precedent proof is the hardness of approximation of our problem. If there is an overlap of one coordinate of two vectors in the previous reduction, we get numerically that the value of $f_W(S)$ is decreased by $C_{\mathrm{hard}} \simeq 0.05719$. Thus we have:

**Corollary 2.2.** The optimization version of *PureRealChannelCoding*, which consists in computing $S(W,k)$, is NP-hard to approximate within a factor $1 - \frac{C_{\mathrm{hard}}}{k} + \epsilon$ for $\epsilon > 0$. So there does not exist a FPTAS that solves this problem if $P \neq NP$ (it fails if relative error is asked to be smaller than $\frac{C_{\mathrm{hard}}}{n}$ since $k \leq n$).

### 2.2.2 Best counter-example to the greedy algorithm

The greedy algorithm still seems to work: we have not been abled yet to find a counter-example to this ratio of approximation, although the precedent proof doesn't work since our function is not submodular in this broader case.

We will depict here the best counter-example we have been able to find so far.

We place ourselves in $\mathbb{R}^k$, where states are unit vectors. First we take a squeezed orthonormal basis:

$$|i'\rangle = \sqrt{1-\epsilon}\,|i\rangle + \sqrt{\frac{\epsilon}{k-1}} \sum_{j \neq i} |j\rangle$$

where $|1\rangle, \ldots, |k\rangle$ is the orthonormal basis and $\epsilon > 0$.

We construct a sequence $S_1, \ldots, S_k$ and $|g_1\rangle, \ldots, |g_k\rangle$ in the following way:

$$S_j = \{|j'\rangle, |(j+1)'\rangle, \ldots, |k'\rangle\}$$

And in particular:

$$S_1 = \{|1'\rangle, \ldots, |k'\rangle\}$$

Let O be the origin of $\mathbb{R}^k$. Take $|g_1\rangle$ the unique point at the same distance of all of points of $S_1$. Thus:

$$|g_1\rangle = \frac{1}{\sqrt{k}} \sum_i |i'\rangle = \frac{1}{\sqrt{k}} \sum_i |i\rangle$$

**Algorithm 2:** Counter-example to our greedy algorithm

```
$C = O$
$H = \mathbb{R}^k$
$|g_1\rangle = \frac{1}{\sqrt{k}}\sum_i |i'\rangle = \frac{1}{\sqrt{k}}\sum_i |i\rangle$
for $j \in \{1, \dots, k-1\}$ do
    $L = C + \text{vect}\{\overrightarrow{n}\}$ affine line where $\overrightarrow{n} = \frac{\overrightarrow{|g_j\rangle} - \overrightarrow{OC}}{\left\| \overrightarrow{|g_j\rangle} - \overrightarrow{OC} \right\|}$
    $H$ = only affine hyperplane containing all points of $S_j$ : $\overrightarrow{n}$ is a normal vector of $H$
    $P = H \cap L$ which is a point, we get it by the formula:
        $$P = C + ((\overrightarrow{|j'\rangle} - \overrightarrow{OC}) \cdot \overrightarrow{n})\overrightarrow{n}$$
    $C = P$ our new center
    $|g_{j+1}\rangle$ = central symmetry of $|j'\rangle$ around $C$ which is given by:
        $$\overrightarrow{|g_{j+1}\rangle} = 2 \times \overrightarrow{OC} - \overrightarrow{|j'\rangle}$$
return $\{|g_1\rangle, \dots, |g_{k-1}\rangle\}$
```

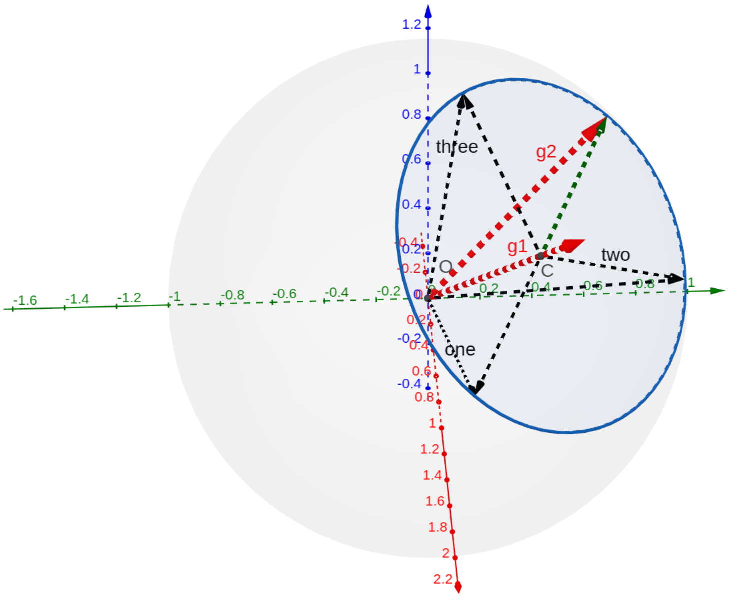


Figure 2: The construction of $g_1$ and $g_2$ (in red) for the case $k = 3$ (*one*, *two* and *three* correspond respectively to $|1'\rangle, |2'\rangle$ and $|3'\rangle$) where we have taken $\epsilon = 0.1$

In figure 2 page 12, we see how we construct $g_1$ and $g_2$. We see also that it is recursive: the way we construct $g_2$ for $k = 3$ is the same as the construction of $g_1$ for $k = 2$ when we stay inside the hyperplane $H$. In particular, we see also this in figure 3 page 13 where we show only the first projection and we see that getting $g_2$ here is the same as getting $g_1$ in dimension 3 and so on.

We take then $W = \{|g_1\rangle\langle g_1|, \dots, |g_{k-1}\rangle\langle g_{k-1}|, |1'\rangle\langle 1'|, \dots, |k'\rangle\langle k'|\}$. The greedy algo-

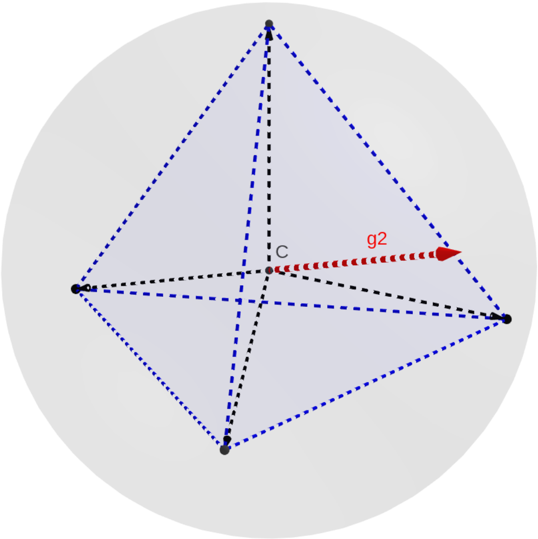


Figure 3: We have represented for $k = 4$ what happened after the first projection on $H$ of dimension 3 in order to find $g_2$ (in red), where the 4 black points are the elements of $S_1$. Note that the center $C$ here is different from $O$ which does not belong to $H$ (see figure 2 page 12 for instance)

rithm takes $S_{\text{greedy}} = \{|g_1\rangle\langle g_1|, \dots, |g_{k-1}\rangle\langle g_{k-1}|, |(k-1)'\rangle\langle(k-1)'|\}$ whereas the optimal solution is to take $S_{\text{opt}} = \{|1'\rangle\langle 1'|, \dots, |k'\rangle\langle k'|\}$.

Numerically, the best parameters in order to minimize the efficiency of our greedy algorithm are the following for small dimensions ($k \leq 10$): $k = 4$ and $\epsilon = 0.199$. Then:

$$\frac{f_W(S_{\text{greedy}})}{f_W(S_{\text{opt}})} \simeq 0.753 > 1 - \left(1 - \frac{1}{4}\right)^4 \simeq 0.684$$

□

# 3 Interpretation of our problem in terms of convex shape covering

## 3.1 Generalized Probabilistic Theories (GPTs)

Introduced in [7], GPTs are a generalization of classical and quantum theories into a more general framework where we consider any convex body as the set of states, and we look at the cone spanned by this space. One of the motives of these theories is to better understand the critical properties of quantum theory, such as entanglement.

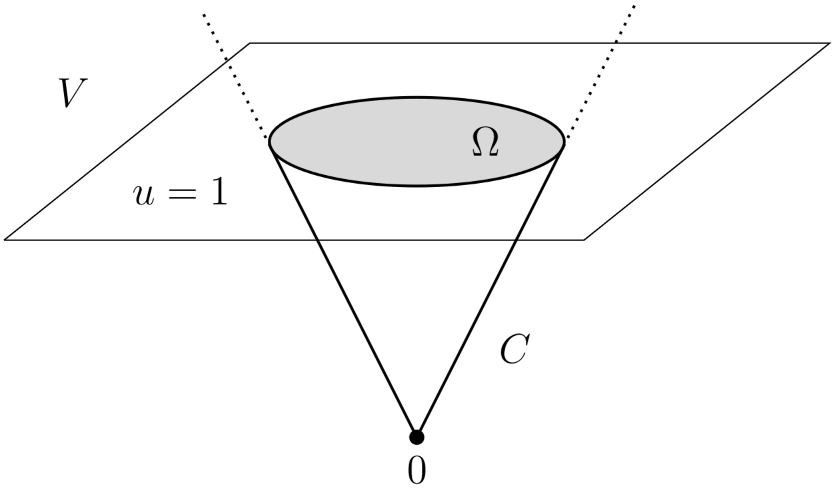


Figure 4: The basic ingredients of a GPT are a real finite-dimensional vector space $V$ and a cone $C$. The order unit functional $u$ defines a hyperplane $u^{-1}(1)$, whose intersection with $C$ identifies the state space $\Omega$.

- Formally, a *GPT* is a triple $(V, C, u)$, where $V$ is a finite-dimensional real vector space, $C \subset V$ is a proper cone, and the *order unit* $u \in \text{int}(C^*)$ is a strictly positive convex-linear functional on $C$: $\forall x \in C - \{0\}$, $u(x) > 0$.
- A subset $C \subseteq V$ is called a *cone* if $\lambda C = C$ for all $\lambda > 0$. A cone is said to be pointed if $C \cap (-C) = \{0\}$, generating if $C - C = \text{span}(C) = V$, and *proper* if it is convex, topologically closed, pointed, and generating. Mathematically, this gives $V$ the structure of an *ordered vector space*: the ordering is defined for $x, y \in V$ by saying that $x \leq_C y$ if $y - x \in C$.
- Positive functionals on $C$ in the dual space $V^*$ form the dual cone $C^*$. If $C$ is proper, then $C^{**} = C$ modulo the identification $V^{**} = V$. We can extend there the order in $C$ for elements of $V^*$: $f \leq_{C^*} g$ if $f(x) \leq_C g(x)$ for all $x \in C$.
- A *classical* cone $C$ is one that is generated by a basis of $V$ , ie. $C = \{\sum_i \lambda_i e_i : \forall i, \lambda_i \geq 0\}$ for some basis $\{e_i\}_i$.
- A natural way to construct a cone is through one of its sections. Namely, given a convex set $K \subseteq V$, let us define the cone $\mathcal{C}(K) = \{(tx, t) : x \in K, t \in \mathbb{R}^+\} \subseteq V \times \mathbb{R}$. One can verify that $\mathcal{C}(K)$ is a proper cone if and only if $K \subset V$ is a *convex body*

(compact convex set with non-empty interior), and that it is non-classical if and only if $K$ is not a simplex. Moreover, every proper cone in dimension $d$ is linearly isomorphic to $\mathcal{C}(K)$ for some $(d-1)$-dimensional convex body $K$

- The *state space* is defined as $\Omega := C \cap u^{-1}(1)$. It is the base of $C$, ie. $C = \{\lambda\omega : \lambda \geq 0, \omega \in \Omega\}$, such that $\forall \omega \in \Omega, u(\omega) = 1$.
- A *measurement* is a finite collection of *effects* $(e_i)_{i \in I}$ such that $e_i \in V^*$ satisfies $e_i \geq_{C^*} 0$ and $\sum_{i \in I} e_i = u$ (so in particular $0 \leq_{C^*} e \leq_{C^*} u$). The probability of obtaining the outcome $i$ when measuring the state $\omega$ is $e_i(\omega) \in [0,1]$.

For example, regular quantum theory can be restated in terms of a GPT in the following way:

- $V := \mathcal{H}_d$ is the real vector space.
- $C := \mathrm{PSD}_d = \{M \in \mathcal{H}_d : M \succcurlyeq 0\}$ is the cone.
- $u := \mathrm{Tr} = \mathrm{Tr}(\mathbb{1}\cdot) = \langle \mathbb{1}, \cdot \rangle$ is the order unit, and it can be identified to $\mathbb{1}$ through the scalar product $\langle A, B \rangle := \mathrm{Tr}\big(A^\dagger B\big)$ and the associated isomorphism between $V^*$ and $V$. Inequalites $\leq_{C^*}$ become $\leq_C$ through this isomorphism.
- In particular, $\Omega = C \cap u^{-1}(1) = \{M \in \mathcal{H}_d : M \succcurlyeq 0 \text{ and } \mathrm{Tr}(M) = 1\} = \mathcal{D}_d$ the usual space state. An effect $e \in V^*$ is such that $0 \leq_{C^*} e \leq_{C^*} u$. Through the previous isomorphism, it can be identified to $0 \preccurlyeq \Lambda_e \preccurlyeq \mathbb{1}$, and the probability of the measuring $e$ becomes $e(M) = \langle \Lambda_e, M \rangle = \mathrm{Tr}\big(\Lambda_e^\dagger M\big)$. A measurement is a finite set of such effects $(e_i)_{i \in I}$ such that $\sum_{i \in I} e_i = u$. Through this isomorphism, this translate to the usual definition of a POVM: a finite set of positive operators $(\Lambda_{e_i})_{i \in I}$ such that $\sum_{i \in I} \Lambda_{e_i} = \mathbb{1}$.

## 3.2 Generalization of the channel coding problem in a GPT

We study here our channel coding problem in a GPT $(V, C, u)$ (and $\Omega$ state space). We suppose we have $\{W_x\}_{x \in X} \subseteq \Omega$, and we define:

**Definition 3.1** (General Channel Coding Function)**.**

$$\begin{aligned} f_W(S) := \quad \underset{(e_x)_{x \in S}}{\text{maximize}} \quad & \sum_{x \in S} e_x(W_x) \\ \text{subject to} \quad & \sum_{x \in S} e_x = u \\ & e_x \geq_{C^*} 0, \forall x \in S \end{aligned} \tag{11}$$

**Property 3.2.** We can replace $\sum\limits_{x \in S} e_x = u$ by $\sum\limits_{x \in S} e_x \leq_{C^*} u$.

This is a conic program, we can compute its dual through the Lagrangian as it is done in [4]. We have also strong duality in this general case so:

**Property 3.3** (Strong Duality)**.**

$$\begin{aligned} f_W(S) = \quad \underset{\rho}{\text{minimize}} \quad & u(\rho) \\ \text{subject to} \quad & \rho \geq_C W_x, \forall x \in S \end{aligned} \tag{12}$$

*Remark.* If the input data is not uniform, but rather follows the probabilities $(q_x)_{x \in S}$, then the primal objective becomes $\sum_{x \in S} q_x |S| e_x(W_x)$, and dual constraints become $\rho \geq_C q_x |S| W_x$.

**Definition 3.4** (Channel Coding Optimization).

$$S(W,k) := \frac{1}{k} \max_{S \subseteq X, |S| \leq k} f_W(S)$$

*Remark.* 1. $\rho$ won't be always unique: it will depend on the shape of the convex body $\Omega$.

2. If we can compute efficiently one instance of this problem, we can use the same greedy algorithm as defined in algorithm 1 to approximate in polynomial time $S(W,k)$. One of the fundamental questions we want to solve is to understand for which shapes this algorithm gives a good approximation.

3. We can also define a relaxation of the dual problem as in [14], and the method to transform one of its solutions to the original dual problem. However we can't analyse its efficiency with the same arguments as in the quantum case. Another question we want to solve is whether this is efficient and for which shapes.

## 3.3 Equivalence with the minimum enclosing convex shape problem (MECP)

**Definition 3.5** (MinEffect). For a GPT $(V,C,u)$, it is the following conic problem:

$$\begin{aligned} f_W(S) = \quad & \underset{\rho}{\text{minimize}} && u(\rho) \\ & \text{subject to} && \rho \geq_C W_x, \forall x \in S \end{aligned} \tag{13}$$

**Definition 3.6** (Minimum Enclosing Convex Problem (MECP) [9]). $\Gamma \subseteq \mathbb{R}^n$ convex body that contains 0, $P \subseteq -\Gamma$ finite set of points, find the least dilatation factor $r \geq 0$, such that a translate of $r\Gamma$ contains $P$:

$$\begin{aligned} R(P,\Gamma) := \quad & \text{minimize} && r \\ & \text{subject to} && P \subseteq \mathcal{D}_\Gamma(\gamma, r) := \gamma + r\Gamma \\ & && \gamma \in \mathbb{R}^n, r \geq 0 \end{aligned} \tag{14}$$

*Remark.* Since the problem is translation invariant, and a dilatation of $\Gamma$ can be found directly on the coefficient $r$, we could have stated MECP with $\Gamma \not\ni 0$ and $P \subseteq \mathbb{R}^n$.

**Definition 3.7** (Lower Cone). For a GPT $(V,C,u)$, the lower cone $\mathcal{C}_{\leq_C}(x)$ of an element $x \in V$ is defined by:

$$\mathcal{C}_{\leq_C}(x) := \{y \in V \text{ s.t. } y \leq_C x\}$$

*Remark.* $\mathcal{C}_{\leq_C}(x)$ is an (affine) proper cone of vertex $x$

For some fixed $c \in \Omega$, we define $\Omega_c := \Omega - c$ the translation of $\Omega$ in the direction $c$. In particular, $\Omega_c \subseteq u^{-1}(0)$ by linearity of $u$.

**Property 3.8.** For $x \in V$ such that $r_x := u(x) > 0, \mathcal{D}(x) := \mathcal{D}_{-\Omega_c}(x - r_x c, r_x) = x - r_x\Omega$ is a base of $\mathcal{C}_{\leq_C}(x)$ included in $u^{-1}(0)$.

*Proof.* $\mathcal{D}(x) \subseteq u^{-1}(0)$ by linearity of $u$. Let us show that $\mathcal{D}(x)$ is a base of $\mathcal{C}_{\leq_C}(x)$, ie. $\mathcal{C}_{\leq_C}(x) = x + \mathbb{R}^+(\mathcal{D}(x) - x)$. Let $y \in \mathcal{C}_{\leq_C}(x)$, ie. $y \leq_C x$. Suppose that $y \neq x$ ($x = x + 0$ otherwise), ie. $y <_C x$. Then $\frac{1}{u(x-y)}(x-y)$ is a state so:

$$y_0 = x - \frac{r_x}{u(x-y)}(x-y) \in \mathcal{D}(x)$$

Then $y = x + \frac{u(x-y)}{r_x}(y_0 - x)$ so $\mathcal{C}_{\leq_C}(x) \subseteq x + \mathbb{R}^+(\mathcal{D}(x) - x)$.

On the other hand, let us take $y \in x + \mathbb{R}^+(\mathcal{D}(x) - x)$: there exist $k \in \mathbb{R}^+$ and $z \in \Omega$ such that $y = x + k[(x - r_x z) - x] = x - kr_x z$. But $kr_x z \geq_C 0$ since $k, r_x \geq 0$ and $z \in \Omega \subseteq C$, so $y = x - kr_x z \leq_C x$, ie. $x + \mathbb{R}^+(\mathcal{D}(x) - x) \subseteq \mathcal{C}_{\leq_C}(x)$. □

We show in the following sections that both problems MinEffect and MECP are equivalent.

### 3.3.1 MinEffect → MECP

We consider a MinEffect instance, ie. a GPT $(V, C, u)$, $\Omega = C \cap u^{-1}(1)$ the state space and a finite set of states $\{W_x\}_{x \in S} \subseteq \Omega$. Let $c \in \Omega$ a reference state we call the center state. We define the following MECP instance:

- $\Gamma := -\Omega_c \subseteq u^{-1}(0) = \mathbb{R}^n$ modulo a choice of a basis, since $V$ is a real finite dimension vector space. $0 = -(c - c) \in \Gamma$ since $c \in \Omega$.
- $P := \{W_x - c\}_{x \in S} \subseteq \Omega_c = -\Gamma$ with the same basis choice.

We define:

**Property 3.9.**

$$\begin{array}{rrcl} \varphi : & \mathbb{R}^n \times \mathbb{R}^+ & \to & V \\ & (\gamma, r) & \mapsto & \gamma + (r+1)c \end{array} \tag{15}$$

is a bijection with $\varphi^{-1}(x) = (x - (u(x) - 1)c, u(x) - 1)$

**Theorem 3.10.** $(\gamma, r)$ is an optimal solution of MECP if and only if $\varphi(\gamma, r)$ is an optimal solution of MinEffect for the instances we have described.

*Proof.* Let $(\gamma, r) \in \mathbb{R}^n \times \mathbb{R}^+$. If $r > 0$, $\mathcal{D}_{-\Omega_c}(\gamma, r)$ is a basis of $\mathcal{C}_{\leq_C}(\varphi(\gamma, r) - c)$ included in $u^{-1}(0)$ by property 3.8. In particular, we have that for $y \in u^{-1}(0)$, $y \in \mathcal{D}_{-\Omega_c}(\gamma, r)$ if and only if $y \leq_C \varphi(\gamma, r) - c$, which is trivially true when $r = 0$. Thus in particular, $W_x - c \in \mathcal{D}_{-\Omega_c}(\gamma, r) \iff W_x \leq_C \varphi(\gamma, r)$

Thus, $P \subseteq \mathcal{D}_{-\Omega_c}(\gamma, r) \iff W_x \leq_C \varphi(\gamma, r), \forall x \in S$, and $u(\varphi(\gamma, r)) = r + 1$ and $r_{\varphi^{-1}(x)} = u(x) - 1$, so we minimize the same parameter up to a constant 1. Furthermore $\varphi$ is a bijection between the variable spaces: with the previous equivalence it means that it maps MECP feasible solution to MinEffect feasible solution and reciprocally. Since we minimize the same parameter through $\varphi$, it maps optimal solutions of MECP to optimal solutions of MinEffect and reciprocally. □

### 3.3.2 MECP → MinEffect

We consider a MECP instance $(\Gamma, \{P_x\}_{x \in S}) \subseteq \mathbb{R}^n \times -\Gamma$ and $0 \in \Gamma$. We define a MinEffect instance composed of a GPT $(V, C, u)$ and a set of states $W \subseteq \Omega = C \cap u^{-1}(1)$ in the following way:

- $V := \mathbb{R}^{n+1}$
- $C := \{(tx, t) : x \in -\Gamma, t \in \mathbb{R}^+\}$
- $\begin{array}{rrcl} u : & \mathbb{R}^n \times \mathbb{R} & \to & \mathbb{R} \\ & (x, t) & \mapsto & t \end{array}$
- So, in particular, $\Omega = C \cap u^{-1}(1) = \{(x, 1) : x \in -\Gamma\}$
- $W := \{(P_x, 1)\}_{x \in S} \subseteq \Omega$

- We define $c := (0, \dots, 0, 1) \in \Omega$ (since $0 \in \Gamma$) and we get then that $\Gamma = -\Omega_c$ if we see $\Gamma$ living in $\mathbb{R}^{n+1}$. Thus $P \subseteq \Omega_c = -\Gamma$.

**Theorem 3.11.** $(\gamma, r)$ is an optimal solution of MECP if and only if $(\gamma, r+1)$ is an optimal solution of MinEffect for the instances we have described.

*Proof.* It is exactly the same proof as in theorem 3.10. □

### 3.3.3 Non-uniform inputs case

In this section we do not suppose anymore that we have uniform input. Thus the dual problem to compute our Channel Coding function becomes:

**Definition 3.12** (MinEffect*)**.** For a GPT $(V, C, u)$, it is the following conic problem:

$$\begin{aligned} f_W(S) = \quad & \underset{\rho}{\text{minimize}} && u(\rho) \\ & \text{subject to} && \rho \geq_C W_x, \forall x \in S \end{aligned} \tag{16}$$

where we require only $W_x \geq_C 0$: $u(W_x)$ is only known to be nonnegative.

We also define a variant of MECP:

**Definition 3.13** (Generalized Minimum Enclosing Convex Problem (MECP*))**.** $\Gamma \subseteq \mathbb{R}^n$ convex body that contains 0, $P \subseteq \bigsqcup_{\lambda \geq 0} \{\alpha + \lambda\Gamma, \alpha \in \mathbb{R}^n\}$ finite set of dilatations of $\Gamma$, find the least dilatation factor $r \geq 0$, such that a translate of $r\Gamma$ contains $P$:

$$\begin{aligned} R(P, \Gamma) := \quad & \text{minimize} && r \\ & \text{subject to} && p \subseteq \mathcal{D}_\Gamma(\gamma, r) := \gamma + r\Gamma, \forall p \in P \\ & && \gamma \in \mathbb{R}^n, r \geq 0 \end{aligned} \tag{17}$$

Instead of having only points, we have also dilations of $\Gamma$ to cover with $\Gamma$.

Let us consider the points $W_x$. Each of them generates a cone $\mathcal{C}_{\leq_C}(W_x)$ with a base in $u^{-1}(0)$ given by $\mathcal{D}(W_x) = \mathcal{D}_{-\Omega_c}(W_x - u(W_x)c, u(W_x)) = W_x - u(W_x)\Omega$ by property 3.8. We transform an instance of MinEffect* to MECP* by taking $P = \{\mathcal{D}(W_x)\}_{x \in S}$ and reciprocally we transform an instance of MECP* to MinEffect* by taking $W = \{\alpha_x + \lambda_x c\}_{x \in S}$. Otherwise the transformations are as before. Then we have:

**Theorem 3.14.**

1. If $\rho$ is a feasible solution of MinEffect*, then $(\rho - u(\rho)c, u(\rho))$ is a feasible solution of MECP*
2. If $(r, \gamma)$ is a feasible solution of MECP*, then $\rho := \gamma + rc \geq_C W_x$ for all $x \in S$, ie. $\rho$ is a feasible solution of MinEffect*

*Remark.* Here, the parameter we minimize is *exactly* the same in both problems through this bijection.

In order to prove this theorem, we will uses some properties on *gauges* of convex bodies containing 0:

**Definition 3.15.** Let $\Gamma \ni 0$ a convex body in $\mathbb{R}^n$. For $x \in \mathbb{R}^n$, we define $\|x\|_\Gamma$, the *gauge* of $x$, in the following way:

$$\|x\|_\Gamma := \inf\{\lambda > 0 : x \in \lambda\Gamma\}$$

**Property 3.16.** We have some properties on gauges of convex bodies:

1. $\|\cdot\|_\Gamma$ is positive and sublinear, ie. subadditive and positively homogeneous.
2. $\|x\|_\Gamma = 0 \iff x = 0$.
3. $\|x\|_{r\Gamma} = \frac{1}{r}\|x\|_\Gamma$ for $r > 0$.
4. The point $y$ at the border of $\mathcal{D}_\Gamma(x,r) = x + r\Gamma$ (which is closed) going from $x$ in the direction $\vec{u}$ is given by:
$$y = x + r\frac{\vec{u}}{\|\vec{u}\|_\Gamma}$$

*Proof.* 1. Classic result, true for any convex set that contains the origin.

2. Classic result, true for any finite dimensonal bounded closed convex set that contains the origin.
3. $\frac{1}{r}x \in \Gamma \iff x \in r\Gamma$ and we conclude by positive homogeneity.
4. By definition of the gauge, a point $z$ is at the border of $\Gamma$ if and only if $\|z\|_\Gamma = 1$ since $\Gamma$ is compact. If we ask that $z$ is positively proportional to $\vec{u}$, we get by positive homogeneity that $z = \frac{\vec{u}}{\|\vec{u}\|_\Gamma}$. If we want $y$ to be obtained from $x$ as the border point of $\mathcal{D}_\Gamma(x,r) = x + r\Gamma$ in the direction $\vec{u}$, this means that $y - x$ is a border point of $r\Gamma$ in the direction $\vec{u}$, ie:
$$y - x = \frac{\vec{u}}{\|\vec{u}\|_{r\Gamma}} = r\frac{\vec{u}}{\|\vec{u}\|_\Gamma}$$

□

*Proof of Theorem 3.14.* 1. Since $\rho \geq_C W_x$, then in particular $\rho \geq_C \mathcal{D}(W_x)$ since it is a base of $\mathcal{C}_{\leq_C}(W_x)$ and then $\mathcal{D}(W_x) \subseteq \mathcal{C}_{\leq_C}(\rho)$ if $W_x \neq 0$, and otherwise it is trivially true. But the base of $\mathcal{C}_{\leq_C}(\rho)$ in $u^{-1}(0)$ is $\mathcal{D}(\rho) = \mathcal{D}_{-\Omega_c}(\rho - u(\rho)c, u(\rho)) = \mathcal{D}_\Gamma(\rho - u(\rho)c, u(\rho))$. Thus $\mathcal{D}(W_x) \subseteq \mathcal{D}_\Gamma(\rho - u(\rho)c, u(\rho))$. So if $\rho \geq_C W_x$ for all $x \in S$, ie. $\rho$ a feasible solution of MinEffect*, then $(\rho - u(\rho)c, u(\rho))$ is a feasible solution of MECP*.

2. Call $r_x = u(w_x)$ and $w_x = W_x - r_x c$. We have then $\mathcal{D}(W_x) = \mathcal{D}_\Gamma(w_x, r_x) \subseteq \mathcal{D}_\Gamma(\gamma, r)$ since $(\gamma, r)$ is a feasible solution of MECP*. If $r = r_x$, then $\mathcal{D}(W_x) = \mathcal{D}(\rho)$ so $W_x = \rho$ and in particular $W_x \leq_C \rho$. Suppose now $r > r_x$. Let us show that $W_x \in \rho + \mathbb{R}^+(\mathcal{D}_\Gamma(\gamma, r) - \rho) = \mathcal{C}_{\leq_C}(\rho)$ by property 3.8.

   Let $\tilde{w_x} = w_x + \frac{r_x}{r - r_x}(w_x - \gamma) = \gamma + \frac{r}{r-r_x}(w_x - \gamma)$. We show that $W_x = \rho + \frac{r-r_x}{r}(\tilde{w_x} - \rho)$ and $\tilde{w_x} \in \mathcal{D}_\Gamma(\gamma, r)$, which means that $W_x \in \mathcal{C}_{\leq_C}(\rho)$ since $r_x \leq r$ by definition of $r$. First:

$$\begin{aligned} \rho + \frac{r - r_x}{r}(\tilde{w_x} - \rho) &= \gamma + rc + \frac{r - r_x}{r}(\gamma + \frac{r}{r - r_x}(w_x - \gamma) - (\gamma + rc)) \\ &= \gamma + (w_x - \gamma) + (r - (r - r_x))c = w_x + r_x c = W_x \end{aligned} \tag{18}$$

   Let us show now that $\tilde{w_x} \in \mathcal{D}_\Gamma(\gamma, r)$. Let us call the following unit gauge direction $\vec{u} := \frac{w_x - \gamma}{\|w_x - \gamma\|_\Gamma}$. Let $w'_x$ be the point at the border of $\mathcal{D}_\Gamma(w_x, r_x)$ in the direction $\vec{u}$ going from $w_x$: $w'_x = w_x + r_x\vec{u} \in \mathcal{D}_\Gamma(w_x, r_x)$. For all $p \in [0,1]$ we have by convexity of $\mathcal{D}_\Gamma(w_x, r_x)$ that $w_p = pw'_x + (1-p)w_x = w_x + pr_x\vec{u} \in \mathcal{D}_\Gamma(w_x, r_x) \subseteq \mathcal{D}_\Gamma(\gamma, r)$. But:

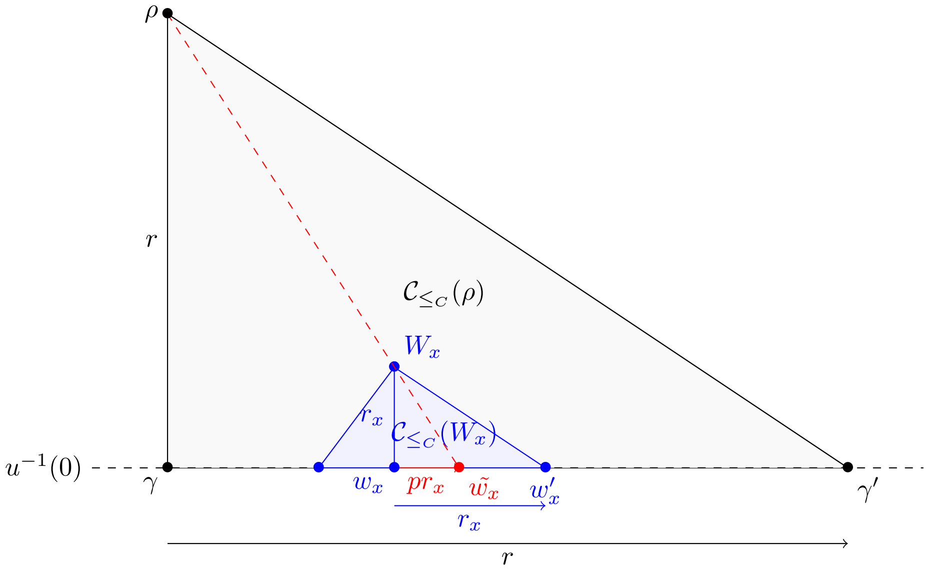


Figure 5: Geometric visualization of points and gauge values defined in the proof of theorem 3.14

$$\tilde{w_x} = w_x + \frac{r_x}{r - r_x}(w_x - \gamma) = w_x + \frac{\|w_x - \gamma\|_\Gamma}{(r - r_x)} r_x \vec{u} = w_{\frac{\|w_x - \gamma\|_\Gamma}{(r - r_x)}}$$

So we have only to prove that $\|w_x - \gamma\|_\Gamma \leq r - r_x$ to get that $\tilde{w_x} \in \mathcal{D}_\Gamma(w_x, r_x) \subseteq \mathcal{D}_\Gamma(\gamma, r)$.

Let $\gamma'$ be the point at the border of $\mathcal{D}_\Gamma(\gamma, r)$ in the direction $\vec{u}$ going from $\gamma$. Then $\gamma' = \gamma + r\vec{u}$. Also, $w'_x = w_x + r_x\vec{u} = \gamma + (r_x + \|w_x - \gamma\|_\Gamma)\vec{u}$ is in the direction $\vec{u}$ going from $\gamma$. Since $w'_x \in \mathcal{D}_\Gamma(w_x, r_x) \subseteq \mathcal{D}_\Gamma(\gamma, r)$ and $\gamma'$ is at the border of $\mathcal{D}_\Gamma(\gamma, r)$ in the same direction $\vec{u}$ from the same point $\gamma$, we have that $\|w'_x - \gamma\|_\Gamma \leq \|\gamma' - \gamma\|_\Gamma$ so $r_x + \|w_x - \gamma\|_\Gamma \leq r$ ie. $\|w_x - \gamma\|_\Gamma \leq r - r_x$.

Thus $W_x \leq_C \rho$, and this is true for all $x \in S$, so $\rho$ is a feasible solution of MinEffect$^*$. □

**Corollary 3.17.** MECP$^*$ and MinEffect$^*$ are equivalent problems.

*Proof.* Both problems have the same feasible solutions thanks to theorem 3.14, and the parameter we minimize is the same in both case, so optimal solutions from one problem map to optimal solutions of the other problem. □

## 3.4 Useful properties of MECP from [9, 10]

Here we assume here that $P$ in MECP is only compact and not finite. We have the following results:

**Theorem 3.18** (Optimality conditions). *$P$ is optimally contained in $\Gamma$ iff:*

1. *$P \subseteq \Gamma$*

2. *For some $2 \leq k \leq n + 1$, there exist $p_1, \ldots, p_k \in P$ and hyperplanes $H(a_i, 1)$ supporting $P$ and $\Gamma$ in $p_i$ s.t. $0 \in \text{conv}\{a_1, \ldots, a_k\}$.*

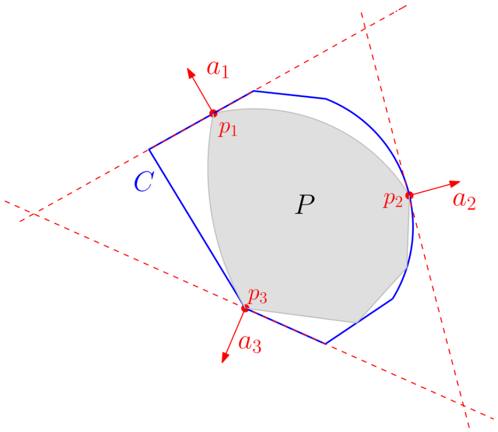


Figure 6: Illustration of optimality conditions (Theorem 3.18)

**Definition 3.19** (Core-radii)**.** The *k-th core-radius* of $P$ with respect to $\Gamma$ is defined by:

$$R_k(P,\Gamma) := \max_{S\subseteq P, |S|\leq k+1} R(S,\Gamma)$$

In particular: $S(W,k) = \frac{1}{k}(1 + R_{k-1}(W-c, -\Omega_c))$

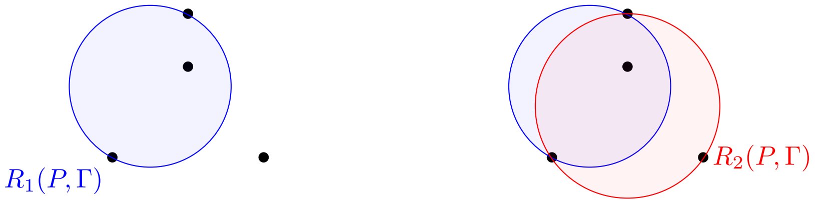


Figure 7: Illustration of several core-radii

**Theorem 3.20** (Helly's theorem)**.** If $\dim(P) \leq k$, then $R_k(P,\Gamma) = R(P,\Gamma)$. For us: $\dim(V)+1$ measurement outputs used at most.

**Theorem 3.21** ($R_k/k$ decreases)**.**

$$\left(\frac{R_k(P,\Gamma)}{k}\right)_{k\in[n]} \text{ decreases. } \underline{\text{For us:}} \left(\frac{S(W,k)-\frac{1}{k}}{1-\frac{1}{k}}\right)_{k\in[n]} \text{ decreases.}$$

**Property 3.22.** Let $\varphi(\gamma) := \min\{r : P \subseteq r\Gamma\}$ defined on $\mathbb{R}^n$. Then $\varphi$ is convex, and in particular it has subgradients.

*Proof.* Let $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$. Let us show that:

$$\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$$

For this we will use the following lemma on convex sets:

**Lemma 3.23.** If $C$ is a convex set and $a, b > 0$, then $aC + bC = (a+b)C$.

*Proof.* $aC + bC \supseteq (a+b)C$ is immediate and always true. If $x \in aC + bC$, there exists $y, z \in C$ such that $x = ay + bz = (a+b)w$ where:

$$w = \frac{a}{a+b}y + \frac{b}{a+b}z \in C$$

since $C$ is convex and $\frac{a}{a+b} + \frac{b}{a+b} = 1$. □

Then by definition of $\varphi$, we have that $P \subseteq C_x := x + \varphi(x)\Gamma$ and $P \subseteq C_y := y + \varphi(y)\Gamma$, so $P \subseteq \lambda C_x + (1-\lambda)C_y$. Since $C_x$ and $C_y$ are convex, by lemma 3.23, we have that:

$$\begin{aligned} P \subseteq \lambda C_x + (1-\lambda)C_y = &\quad \lambda x + (1-\lambda)y + \lambda\varphi(x)C + (1-\lambda)\varphi(y)C \\ = &\quad \lambda x + (1-\lambda)y + (\lambda\varphi(x) + (1-\lambda)\varphi(y))C \end{aligned} \tag{19}$$

But since $\varphi(\lambda x + (1-\lambda)y)$ is the minimum dilation coefficient of such an inclusion, we have that $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$. □

**Property 3.24.** If $a$ is a subgradient of $\varphi$ at $c$ then $-a$ is an outer normal of a hyperplane supporting $c + \varphi(c)\Gamma \supseteq P$ at $p_i \in P$.

**Definition 3.25** (Core-set)**.** $S \subseteq P$ is a *$\epsilon$-core-set* of P if $R(S,\Gamma) \leq R(P,\Gamma) \leq (1+\epsilon)R(S,\Gamma)$

**Theorem 3.26.** For all $P, \Gamma \subseteq \mathbb{R}^n, \epsilon \geq 0$, there exists an $\epsilon$-core-set of size at most $\lceil \frac{n}{1+\epsilon} \rceil + 1$.
Moreover, for any $\epsilon < 1$ there exist $P \subseteq \mathbb{R}^n$ and a 0-symmetric convex body $\Gamma$ (ie. $\Gamma = -\Gamma$) such that no smaller subset of $P$ suffices.
Also, this bound is optimal for the $n$-regular simplex.

*Remark.* The MEC of some set of points is not necessarely unique, for instance it isn't for squares. However, in the shapes we will study, we will look at the cases where it is unique.

## 3.5 Geometric interpretation

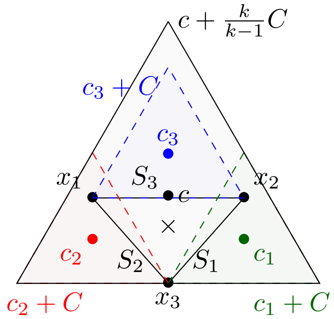
### 3.5.1 Simplex: classical case

When the shape $\Gamma$ is the opposite of the probability simplex $\Omega_c = T^n := \{x \in {\mathbb{R}_+}^{n+1} : u(x) := \sum_{i=1}^{n+1} x_i = 1\}$, it corresponds to the classical problem of channel coding. Up to an affine transformation, we get in this way all classical cones. For them, the function $f_W$ is submodular (since $f_W$ is submodular for the probability simplex and MECP is stable by affine transformations). In particular we have some basic interpretation of submodularity in the shapes:

**Property 3.27.** If $f_W$ is submodular, then $\text{MEC}(S) \subseteq \text{MEC}(T)$ for $S \subseteq T$ if MEC is unique.

*Proof.* By contraposition. Let $S \subset T$ such that $\text{MEC}(S) \not\subseteq \text{MEC}(T)$. There exists thus $x$ such that $x \in \text{MEC}(S)$ and $x \notin \text{MEC}(T)$, so in particular $x \notin T$. This implies that $\text{MEC}(S \cup \{x\}) = \text{MEC}(S)$, so $f_w(S \cup \{x\}) = f_W(S)$, but $\text{MEC}(T) \not\subseteq \text{MEC}(T \cup \{x\})$, so by unicity of MEC and monocity of $f_w$ that $f_w(T \cup \{x\}) > f_W(T)$, so $f_w$ is not submodular. □

We conjecture that this property is in fact an equivalence, and is fulfilled only for simplices, where MEC ar unique as a consequence of the uniqueness of the supremum $\rho$ in that cone.



### 3.5.2 $\mathcal{D}_d$: quantum case

Let us call $\mathcal{D}_d := \{M \in \mathcal{H}_d : M \succcurlyeq 0 \text{ and } \operatorname{Tr}(M) = 1\}$ the set of quantum states. In the quantum context, we take for convenience $c = \mathbb{1}_d := \frac{1}{d}\mathbb{1}$, and we call $Q_d := \Omega_c = \mathcal{D}_d - \mathbb{1}_d$. The equivalent MEC problem is given by $P = W - \mathbb{1}_d \subseteq Q_d$ and $\Gamma := -\Omega_c = -Q_d$

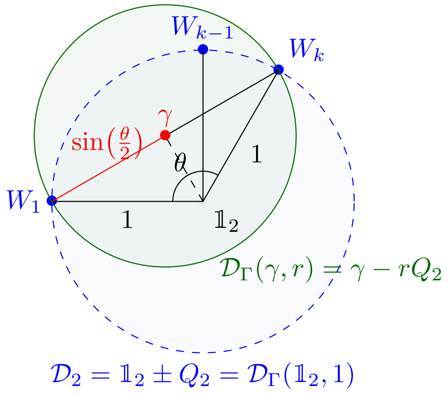
**Qubit case ($d = 2$):** Here we have that $Q_2$ is isometric to the Bloch sphere $\mathbb{B}_3$ and $P$ are the coordinates of $W$ in that sphere. We have in particular that $Q_2$ is symmetric, ie. $Q_2 = -Q_2 = \Gamma$. Thus we are reduced to find the Minimum Enclosing Ball of points (which will be discussed more generally in section 3.5.3) to compute one value of MECP. This interpretation was found by [13] in the dimension 2 case, with some interpretation of non-uniform inputs which we can also interpret with our problem MECP* as a Minimum Enclosing Ball of Balls problem, first studied by [19]. In both cases, these can be computed efficiently without using SDPs.

In particular, for pure qubit states, we can now prove the following corollary:

**Corollary 3.28.** In the real, pure and dimension 2 case, we have the following property for a set of states $\{W_x = |\psi_x\rangle\langle\psi_x|\}_{x\in S}$. Let $\theta_S$ be the minimum angle such that all states of $S$ can be put in a slice of the Bloch circle of radius $\theta$. Then $f_W(S) = g(\theta_S)$ where:

$$g(\theta) = \begin{cases} 1 + \sin\left(\frac{\theta}{2}\right) \text{ if } 0 \leq \theta \leq \pi \\ 2 \text{ if } \pi \leq \theta \leq 2\pi \end{cases}$$

*Proof.* If all states are in one hemisphere of the Bloch circle, the minium enclosing circles of all states is the minimum enclosing circle of extremal ones (depicted as $W_1$ and $W_k$ in the drawing below).


$W_{k-1}$
$W_k$
$\gamma$
$\sin\left(\frac{\theta}{2}\right)$
$\theta$
1
$W_1$
1
$\mathbb{1}_2$
$\mathcal{D}_\Gamma(\gamma, r) = \gamma - rQ_2$
$\mathcal{D}_2 = \mathbb{1}_2 \pm Q_2 = \mathcal{D}_\Gamma(\mathbb{1}_2, 1)$

Then, by a simple use of Pythagore, we get that the radius of that circle is given by $r = \sin\left(\frac{\theta}{2}\right)$, so $f_W(S) = 1 + \sin\left(\frac{\theta_S}{2}\right) = g(\theta_S)$ for $0 \leq \theta_S \leq \pi$.

Otherwise, if we can't put allo states in one hemisphere, by optimality conditions of MECP, we get that $\mathcal{D}_\Gamma(\mathbb{1}_2, 1)$ is optimal, so that $r = 1$, ie. $f_W(S) = 2 = g(\theta_S)$ for $\theta_S \geq \pi$. □

**General case:**

- We look at $Q_d = \mathcal{D}_d - \mathbb{1}_d \subseteq \mathbb{R}^n$ with $n = d^2 - 1$ which is not symmetric as soon as $d > 2$.
- MEQ of a set of points is unique since the dual quantum channel coding problem has a unique solution.
- In $\Gamma$: 0 is equidistant from all extremal points (pure states).
- $\mathcal{B}\left(0, \frac{1}{\sqrt{d(d-1)}}\right) \subseteq \Gamma \subseteq \mathcal{B}\left(0, \sqrt{1 - \frac{1}{d}}\right)$ and tight ratio $d - 1$.
- Like $T^{d-1}$, no $\epsilon$-core-sets of size $< \lceil \frac{d-1}{1+\epsilon} \rceil + 1$: if $k < \lceil \frac{d-1}{1+\epsilon} \rceil$, then for $|S| \leq k + 1$: $R(\mathcal{D}_d, -Q_d) = \frac{d-1}{k} R_k(\mathcal{D}_d, -Q_d) > (1+\epsilon) R_k(\mathcal{D}_d, -Q_d) \geq (1+\epsilon) R(S, -Q_d)$

### 3.5.3 Ball $\mathbb{B}_n$

We study here the case of MECP where we have $\Gamma := \mathbb{B}_n$ is a 2−norm ball in dimension $d$, which is a symmetric convex body. It is a well-studied problem, and a specific instance of MEBP can be computed efficiently (via an FPTAS), thanks to an algorithm by [2] for instance. We can speak of the MEB of a set of points since it is unique for that shape.

**Efficient approximation** In the case of balls, optimality conditions can be rewritten in the following form:

**Lemma 3.29** (Half-space lemma)**.** MEB$(P)$ is the minimum enclosing ball of $P \subseteq \mathbb{R}^n$ iff any closed half-space that contains the center $c_{\mathcal{B}(P)}$ also contains a point of $P$ on the boundary of MEB$(P)$.

We have not yet been able to prove the efficiency of the generic greedy algorithm, but we hav another algorithm wich will efficiently approximate $S(W, k)$ via the existence of $\epsilon$-core-sets of size independent to the dimension in the case of balls:

**Theorem 3.30** ([2])**.** An $\epsilon$-core-set $S_\epsilon$ of $P \subseteq \mathbb{R}^n$ of size $\lceil \frac{2}{\epsilon} \rceil$ can be efficiently computed. In particular, this gives a $(1 + \frac{2}{k})$-approximation of $R_k(P, C)$.

*Proof of Corollary.* Let $\epsilon$ such that $k+1 = \lceil \frac{2}{\epsilon} \rceil = |S_\epsilon|$. Then $\epsilon \leq \frac{2}{k}$, so $R_k(P,\Gamma) \leq R(P,\Gamma) \leq (1+\epsilon)R(S_\epsilon,\Gamma) \leq (1+\frac{2}{k})R(S_\epsilon,\Gamma)$. □

*Remark.* This decreasing approximation ratio gives in particular the existence of a PTAS computing $R_k(P,\Gamma)$ (and so $S(W,k)$):

Let $\epsilon > 0$ fixed.

- If $\epsilon \geq \frac{2}{k}$, then we run the previous algorithm which gives a $(1+\frac{2}{k})$-approximation, so in particular a $(1+\epsilon)$-approximation.
- If $\epsilon < \frac{2}{k}$, then $k < \frac{2}{\epsilon}$ is bounded by a constant, so we do an exhaustive search of all sets of size $k$, which will be done in time $\mathcal{O}(|P|^{\frac{2}{\epsilon}})$ which is polynomial.

To get such a core-set we wil use the following algorithm.

**Algorithm 3:** Ball algorithm [2]

```
Input: $k \in \{1,\dots,|P|\}$
Output: An $\frac{2}{k}$-core-set $S_k$ of $P \subseteq \mathbb{R}^n$ of size $k+1$
$S_0 = \{p\}$ with $p$ any point of $P$
$c_0 = p, r_0 = 0$                    // Center and radius of MEB of $S_0$
for $i \in \{1,\dots,k\}$ do
    $q_i$ = furthest point from $c_{i-1}$ in $P$
    $S_i = S_{i-1} \cup \{q_i\}$
    $c_i, r_i$ such that MEB$(S_i) = \mathcal{B}(c_i, r_i)$
return $S_k$
```

We show now some example where the greedy choice is different from the generic greedy algorithm:

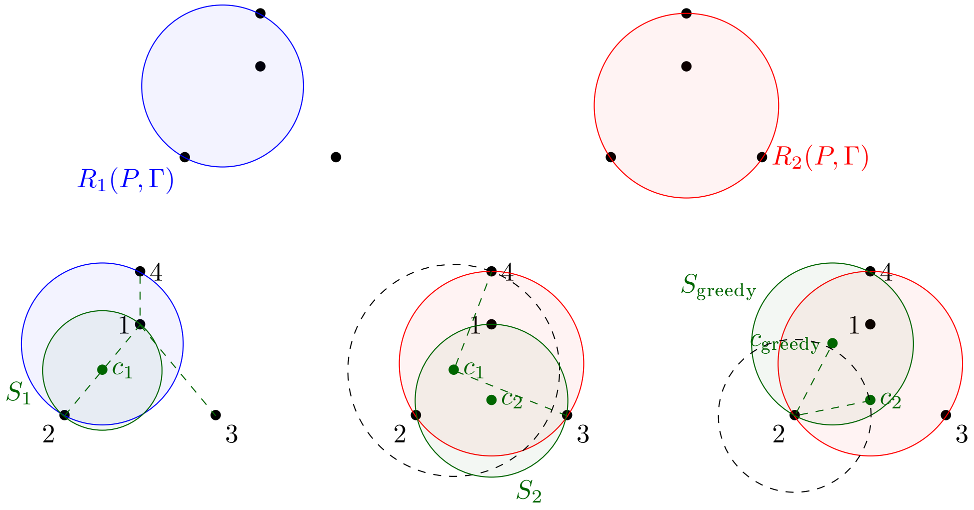


Figure 8: Differences between algorithms 1 and 3

*Proof sketch of Theorem 3.30, done in [2].* If at some point $c_{i+1} = c_i$, then we have an MEB of all $P$, so it will be in particular a $\epsilon$-core-set. We assume this never happens. At step $i$, take $H \ni c_i$ with $H \perp \overrightarrow{c_i c_{i+1}}$, and $p \in \partial\mathcal{B}(S_i) \cap H^+ \cap S_i$ thanks to the half-space lemma. By geometric arguments, we get two inequalities:

1. $r_{i+1} \geq \|c_{i+1} - p\|_2 \geq \sqrt{r_i^2 + \|c_{i+1} - c_i\|_2^2}$ by Pythagore.
2. $r_{i+1} \geq \|c_{i+1} - q_{i+1}\|_2 \geq R - \|c_{i+1} - c_i\|_2$ since $q_{i+1}$ furthest point away from $c_i$.

This will imply by studying the worst case in this inequality, $\sqrt{r_i^2 + \|c_{i+1} - c_i\|_2^2} = R - \|c_{i+1} - c_i\|_2$, that after $k = \lceil \frac{2}{\epsilon} \rceil$ steps, we get $R \leq (1 + \epsilon) r_k$. □

### NP-hardness and inapproximability result

**Problem:** *BallChannelCoding*

**Instance:** A finite set $P \subseteq \mathbb{R}^d$ of points of size $n$, $R \in \mathbb{R}$ and $k \in \{1, \ldots, n\}$

**Question:** Does $R_{k-1}(P, \Gamma) (= kS(W, k) - 1) \geq R$, where $\Gamma = \mathbb{B}^d$ ?

**Theorem 3.31.** *BallChannelCoding* is NP-hard. Furthermore, its optimization version is even NP-hard to approximate within a factor $1 - \frac{1}{4k^2} + \epsilon$ for $\epsilon > 0$. In particular, there does not exist a FPTAS that solves this problem if $P \neq NP$.

In order to prove this theorem, we will need some lemmas on the MEB of points:

**Lemma 3.32.** 1. The center $c$ of the MEB of $P$ lies in the convex hull of $P$.

2. If a point $c$ is equidistant from all points in $P$ and $c$ lies in the convex hull of $P$, then $c$ is the center of the MEB of $P$ and its radius is given by $\|p - c\|_2$ for any $p \in P$.

*Proof.* 1. This is a consequence of the half-space lemma: if $c$ is outside the convex hull of $P$, there exist a hyperplan $H \ni c$ such that $P$ is included in the one of the open half-spaces induced by $H$, which is in contradiction with the half-space lemma.

2. We use the fact that $a \in \mathrm{CH}(P) \iff \forall u : \|u\|_2 = 1, \min_{p \in P} (p - a)^T u \leq 0$.

We first remark that $\mathbb{B}(c, d)$, where $d = \|c - p\|_2$ for $p \in P$ is an enclosing ball of $P$ since $c$ is equidistant from $P$. Let us assume by contradiction that it is not the MEB of $P$, but that $\mathrm{MEB}(P) = \mathbb{B}(c', d')$. We have $d' < d$ since otherwise $\mathbb{B}(c, d)$ is the MEB of $P$. Thus this implies that $c' \neq c$ since $p \notin \mathbb{B}(c, d')$ for $p \in P$ since $\|c - p\|_2 = d > d'$.

Let us define $u = \frac{c' - c}{\|c' - c\|_2}$. Since $c \in \mathrm{CH}(P)$, we have that $\min_{p \in P} (p - c)^T u \leq 0$, so there exists some $p \in P$ such that $(p - c)^T u \leq 0$. Thus:

$$\begin{aligned} \|p - c'\|_2 = &\quad \|p - c + c - c'\|_2 = \sqrt{\|p - c\|_2^2 + \|c - c'\|_2^2 + 2(p - c)^T (c - c')} \\ = &\quad \sqrt{d^2 + \|c - c'\|_2^2 - 2\|c - c'\|_2 (p - c)^T u} \geq \sqrt{d^2 + \|c - c'\|_2^2} > d \end{aligned} \tag{20}$$

since $(p - c)^T u \leq 0$, which leads to a contradiction since $\|p - c'\|_2 \leq d' < d$. □

*Proof of Theorem 3.31.* Like for *PureRealChannelCoding*, we will show that *BallChannelCoding* is NP-complete with a reduction from *X3C* define in section 2.2.1. We

We use the following reduction. Let $Q = \{q_1, \ldots, q_m\}$ and $C = \{c_1, \ldots, c_n\}$. We take $k = \lfloor \frac{m}{3} \rfloor$, $R = \sqrt{1 - \frac{1}{k}}$, and $P = \{p^j\}_{j \in [n]} \subseteq \mathbb{R}^m$, where

$$p_i^j = \begin{cases} \frac{1}{\sqrt{3}} \text{ if } q_i \in c_j \\ 0 \text{ otherwise} \end{cases}$$

This is a valid instance of *BallChannelCoding*.

Let us show that there exists an exact cover for $Q$ if and only if $R_{k-1}(P, \Gamma) \geq R$.

($\Rightarrow$) If there is and exact cover $S$ of $Q$, then $|S| = \frac{m}{3} \in \mathbb{N}$. We have then $|S| = k$. We show that the MEB of $P_S := \{p^j\}_{j \in S}$ has a radius equals to $R = \sqrt{1 - \frac{1}{k}}$, which is enough to show that $R_{k-1}(P, \Gamma) \geq R$.

Since $S$ is an exact cover of $Q$, this means that in $P_s$, each coordinate of all points is non-zero exactly once for some point where it is equal to $\frac{1}{\sqrt{3}}$. We have then that their barycenter $c_{\text{opt}} := \frac{1}{k}\begin{pmatrix} \frac{1}{\sqrt{3}} \\ \vdots \\ \frac{1}{\sqrt{3}} \end{pmatrix}$ is the center of the MEB, since it lies in their convex hull and it is equidistant from all of them, thanks to lemma 3.32. In that case, all points of $P_S$ are equidistant to the center $c_{\text{opt}}$, so the radius of the MEB is given by:

$$\begin{aligned} r_{P_S} = \; & \|p^{s_1} - c_{\text{opt}}\|_2 = \sqrt{3 \times \Big(\frac{1}{\sqrt{3}}\Big(1 - \frac{1}{k}\Big)\Big)^2 + (m-3) \times \Big(\frac{1}{k\sqrt{3}}\Big)^2} \\ = \; & \sqrt{\Big(1 - \frac{1}{k}\Big)^2 + \frac{m}{3k^2} - \frac{1}{k^2}} = \sqrt{\Big(1 - \frac{2}{k} + \frac{1}{k^2}\Big) + \frac{1}{k} - \frac{1}{k^2}} \\ = \; & \sqrt{1 - \frac{1}{k}} = R \end{aligned} \tag{21}$$

($\Leftarrow$) We first remark that in the previous situation, the MEB of $P_S$ which we call $\mathbb{B}(c_{\text{opt}}, R)$ is in fact enclosing all the points possible coming from a reduction from *X3C*. Thus $R$ is the maximal value that can be achieved.

We show this by contradiction. We assume that there does not exist any exact cover for $Q$. Thus for any set $T$ of size $k$, there is a least one coordinate which is not covered by any points of $P_T$ since $3k \leq m$, so all coordinates are covered if and only if we have an exact cover. Thus their MEB is different from the previous one $\mathbb{B}(c_{\text{opt}}, R)$, since its center lies in a subspace which does not contain $c_{\text{opt}}$ thanks to lemma 3.32. By unicity of MEBs, this implies that the radius of the MEB of $P_T$ is strictly smaller than the radius of $\mathbb{B}(c_{\text{opt}}, R)$, since it is also an enclosing ball of our current instance. Thus we get that $r_{P_S} < R$, ie. $R_{k-1}(P, \Gamma) < R$ since it is true for any set $T$ of size $k$.

In order to get the inapproximability result, we need to quantify the minimum gap in the previous proof between the optimal MEB of any exact cover $\mathbb{B}(c_{\text{opt}}, R)$, and any MEB of a set of points which is not an exact cover, so when there is at least one zero-coordinate. We look at cases where $k = \frac{m}{3} \in \mathbb{N}$. We discuss differents cases depending on the number $N$ of zero-coordinates, ie. that are not covered by any point of $P_S$:

$\underline{N \geq 3}$: First, as soon as there at least 3 zero-coordinates, we have that the MEB of $P_S$ is covered by an enclosing ball of radius $r_{\geq 3} = \sqrt{1 - \frac{1}{k-1}}$, simply by seeing this as the previous problem with a parameter $k-1$. Thus we get in this situation that:

$$\frac{r_{\geq 3}}{R} \leq \sqrt{\frac{1 - \frac{1}{k-1}}{1 - \frac{1}{k}}} = \sqrt{\frac{\frac{k-2}{k-1}}{\frac{k-1}{k}}} = \sqrt{\frac{k^2 - 2k}{k^2 - 2k + 1}} = \frac{1}{\sqrt{1 + \frac{1}{k^2 - 2k}}} \leq 1 - \frac{1}{2k^2} \Big(\leq 1 - \frac{1}{4k^2}\Big)$$

Thus we have only to look at all the cases where there is one or two zero-coordinates. We will show that the worst case is obtained in the cas when there is only one zero-coordinate (where there is only one case to study) and that we get the bound

$\leq 1 - \frac{1}{4k^2}$ here. When there are two zero-coordinates, then there will be three situations to study, and all of them will lead to gaps bigger than the one with one zero-coordinate.

$\underline{N=1:}$ We suppose that we have one zero-coordinate. In order to simplify, we assume this is the last coordinate, and that the only overlap is obtained in the last non-zero coordinate of $p^{k-1}$ and the first non-zero coordinate of $p^k$, where we assume all coordinates are sorted with relation to our set $P = \{p^1, \dots, p^k\}$ with:

$$p^1 := \frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\\0\\\vdots\\0\end{pmatrix};p^2 := \frac{1}{\sqrt{3}}\begin{pmatrix}0\\0\\0\\1\\1\\1\\0\\\vdots\\0\end{pmatrix};\dots;p^{k-1} := \frac{1}{\sqrt{3}}\begin{pmatrix}0\\\vdots\\0\\1\\1\\1\\0\\0\\0\end{pmatrix};p^k := \frac{1}{\sqrt{3}}\begin{pmatrix}0\\\vdots\\0\\0\\0\\1\\1\\1\\0\end{pmatrix}$$

In that case, the point $c_1$ defined below is equidistant from all of thes points and in their convex hull, so thanks to lemma 3.32, it will be the center of their MEB which will be of radius $r_1 = \sqrt{1 - \frac{1}{k-\frac{1}{2}}}$:

$$c_1 := \frac{1}{\sqrt{3}\left(k - \frac{1}{2}\right)}\begin{pmatrix}1\\\vdots\\1\\3/4\\3/4\\3/2\\3/4\\3/4\\0\end{pmatrix}$$

First, $c_1 = \left(\sum_{i=1}^{k-2} \frac{1}{k-\frac{1}{2}} p^i\right) + \frac{3}{4}\frac{1}{k-\frac{1}{2}}(p^{k-1} + p^k)$, so $c_1$ is in the convex hull of $P$ since $\left(\sum_{i=1}^{k-2} \frac{1}{k-\frac{1}{2}}\right) + 2 \times \frac{3}{4}\frac{1}{k-\frac{1}{2}} = 1$. Let us show that it is equidistant from all $p^i$:

$$\begin{aligned}
&\text{If } i \leq k-2, \left\|p^i - c_1\right\|_2 \\
&= \sqrt{\left(1 - \frac{1}{k-\frac{1}{2}}\right)^2 + (k-3)\left(\frac{1}{k-\frac{1}{2}}\right)^2 + \frac{1}{3}\left(4\left(\frac{3}{4}\left(\frac{1}{k-\frac{1}{2}}\right)\right)^2 + \left(\frac{3}{2}\left(\frac{1}{k-\frac{1}{2}}\right)\right)^2\right)} \\
&= \sqrt{1 - \frac{2}{k-\frac{1}{2}} + \frac{1}{(k-\frac{1}{2})^2} + (k-3)\frac{1}{(k-\frac{1}{2})^2} + \frac{3}{2}\frac{1}{(k-\frac{1}{2})^2}} \\
&= \sqrt{1 - \frac{2}{k-\frac{1}{2}} + \frac{k-\frac{1}{2}}{(k-\frac{1}{2})^2}} = \sqrt{1 - \frac{1}{k-\frac{1}{2}}}
\end{aligned} \tag{22}$$

If $i = k-1$ or $k$, $\left\|p^i - c_1\right\|_2$

$$= \sqrt{(k-2)\Big(\frac{1}{k-\frac{1}{2}}\Big)^2 + \frac{1}{3}\Big(2\Big(\frac{3}{4}\Big(\frac{1}{k-\frac{1}{2}}\Big)\Big)^2 + 2\Big(1-\frac{3}{4}\Big(\frac{1}{k-\frac{1}{2}}\Big)\Big)^2 + \Big(1-\frac{3}{2}\Big(\frac{1}{k-\frac{1}{2}}\Big)\Big)^2\Big)}$$
$$= \ldots = \sqrt{1 - \frac{2}{k-\frac{1}{2}} + \frac{k-\frac{1}{2}}{(k-\frac{1}{2})^2}} = \sqrt{1 - \frac{1}{k-\frac{1}{2}}} \tag{23}$$

Thus $c_1$ is equidistant form all $p \in P$, so $\mathrm{MEB}(P) = \mathbb{B}\Big(c_1, r_1 := \sqrt{1-\frac{1}{k-\frac{1}{2}}}\Big)$. We get then the bound on the ratio exepected:

$$\frac{r_1}{R} \le \sqrt{\frac{1-\frac{1}{k-\frac{1}{2}}}{1-\frac{1}{k}}} = \sqrt{\frac{\frac{k-\frac{3}{2}}{k-\frac{1}{2}}}{\frac{k-1}{k}}} = \sqrt{\frac{k^2-\frac{3}{2}k}{k^2-\frac{3}{2}k+\frac{1}{2}}} = \frac{1}{\sqrt{1+\frac{1}{2k^2-3k}}} \le 1 - \frac{1}{4k^2}$$

$\underline{N=2}$: There are three kinds of overlap of coordinates corresponding to two zero-coordinates in $P$:

1. $p^{k-1} := \frac{1}{\sqrt{3}}\begin{pmatrix}0\\ \vdots\\ 0\\ 1\\ 1\\ 1\\ 0\\ 0\\ 0\end{pmatrix}; p^k := \frac{1}{\sqrt{3}}\begin{pmatrix}0\\ \vdots\\ 0\\ 0\\ 1\\ 1\\ 1\\ 0\\ 0\end{pmatrix}$ : 2 overlaps on 2 points

2. $p^{k-2} := \frac{1}{\sqrt{3}}\begin{pmatrix}0\\ \vdots\\ 0\\ 1\\ 1\\ 0\\ 0\\ 1\\ 0\\ 0\\ 0\\ 0\end{pmatrix}; p^{k-1} := \frac{1}{\sqrt{3}}\begin{pmatrix}0\\ \vdots\\ 0\\ 0\\ 0\\ 1\\ 1\\ 0\\ 1\\ 0\\ 0\\ 0\end{pmatrix}; p^k := \frac{1}{\sqrt{3}}\begin{pmatrix}0\\ \vdots\\ 0\\ 0\\ 0\\ 0\\ 0\\ 1\\ 1\\ 1\\ 0\\ 0\end{pmatrix}$ : 2 overlaps on 2 different coordinates of 3 points

3. $p^{k-2} := \frac{1}{\sqrt{3}}\begin{pmatrix}0\\ \vdots\\ 0\\ 1\\ 1\\ 0\\ 0\\ 1\\ 0\\ 0\\ 0\\ 0\end{pmatrix}; p^{k-1} := \frac{1}{\sqrt{3}}\begin{pmatrix}0\\ \vdots\\ 0\\ 0\\ 0\\ 1\\ 1\\ 1\\ 0\\ 0\\ 0\\ 0\end{pmatrix}; p^{k} := \frac{1}{\sqrt{3}}\begin{pmatrix}0\\ \vdots\\ 0\\ 0\\ 0\\ 0\\ 0\\ 1\\ 1\\ 1\\ 0\\ 0\end{pmatrix}$ : 1 overlap on 1 coordinate of 3 points

We get respectively by using the same calculation as in the $N = 1$ case:

1.
$$c_2^1 = \frac{1}{\sqrt{3}\left(k - \frac{4}{5}\right)}\begin{pmatrix}1\\ \vdots\\ 1\\ 3/5\\ 6/5\\ 6/5\\ 3/5\\ 0\\ 0\end{pmatrix} \text{ and } r_2^1 = \sqrt{1 - \frac{1}{k - \frac{4}{5}}}$$

2.
$$c_2^2 := \frac{1}{\sqrt{3}\left(k - \frac{6}{7}\right)}\begin{pmatrix}1\\ \vdots\\ 1\\ 6/7\\ 6/7\\ 6/7\\ 6/7\\ 9/7\\ 9/7\\ 3/7\\ 0\\ 0\end{pmatrix} \text{ and } r_2^2 = \sqrt{1 - \frac{1}{k - \frac{6}{7}}}$$

3.

$$c_2^3 := \frac{1}{\sqrt{3}\left(k - \frac{6}{5}\right)} \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 3/5 \\ 3/5 \\ 3/5 \\ 3/5 \\ 9/5 \\ 3/5 \\ 3/5 \\ 0 \\ 0 \end{pmatrix} \text{ and } r_2^3 = \sqrt{1 - \frac{1}{k - \frac{6}{5}}}$$

Since they give all smaller coefficients $r$, we have proved that the worst case was indeed the $N = 1$ case, where we have certified that there is always a gap $\frac{r}{R} \leq 1 - \frac{1}{4k^2}$ in our reduction. Thus if we could approximate $R$ within a gap $1 - \frac{1}{4k^2} + \epsilon$ for $\epsilon > 0$, then because of the gap, it would mean that in that reduction we were able to find in fact the optimal solution, ie. solve $X3C$, which is known to be NP-hard. We get this way the inapproximability result announced.

□

**Theorem 3.33** (Channel Coding for Balls)**.** In the case of balls, there exists a PTAS to compute $S(W, k)$ but there does not exist any FPTAS if $P \neq NP$.

# References

[1] Ashwinkumar Badanidiyuru, Shahar Dobzinski, and Sigal Oren. Optimization with demand oracles. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 110–127, 2012.

[2] Mihai Badoiu and Kenneth L Clarkson. Smaller core-sets for balls. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 801–802. Society for Industrial and Applied Mathematics, 2003.

[3] Joonwoo Bae. Structure of minimum-error quantum state discrimination. *New Journal of Physics*, 15(7):073037, 2013.

[4] Joonwoo Bae, Dai-Gyoung Kim, and Leong-Chuan Kwek. Structure of optimal state discrimination in generalized probabilistic theories. *Entropy*, 18(2):39, 2016.

[5] Joonwoo Bae and Leong-Chuan Kwek. Quantum state discrimination and its applications. *Journal of Physics A: Mathematical and Theoretical*, 48(8):083001, 2015.

[6] Siddharth Barman and Omar Fawzi. Algorithmic aspects of optimal channel coding. *IEEE Transactions on Information Theory*, 64(2):1038–1045, 2017.

[7] Jonathan Barrett. Information processing in generalized probabilistic theories. *Physical Review A*, 75(3):032304, 2007.

[8] Andrew An Bian, Joachim M Buhmann, Andreas Krause, and Sebastian Tschiatschek. Guarantees for greedy maximization of non-submodular functions with applications. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 498–507. JMLR. org, 2017.

[9] René Brandenberg and Stefan König. No dimension-independent core-sets for containment under homothetics. *Discrete & Computational Geometry*, 49(1):3–21, 2013.

[10] René Brandenberg and Lucia Roth. Minimal containment under homothetics: a simple cutting plane approach. *Computational Optimization and Applications*, 48(2):325–340, 2011.

[11] Ali Çivril and Malik Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410(47-49):4801–4811, 2009.

[12] Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

[13] Matthieu E Deconinck and Barbara M Terhal. Qubit state discrimination. *Physical Review A*, 81(6):062304, 2010.

[14] Omar Fawzi, Johanna Seif, and Dániel Szilágyi. Approximation algorithms for classical-quantum channel coding. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2569–2573. IEEE, 2019.

[15] Michael R Garey and David S Johnson. *Computers and intractability*, volume 174. freeman San Francisco, 1979.

[16] Kieran Hunter. Results in optimal discrimination. In *AIP Conference Proceedings*, volume 734, pages 83–86. American Institute of Physics, 2004.

[17] Andreas Krause and Volkan Cevher. Submodular dictionary selection for sparse representation. In *International Conference on Machine Learning (ICML)*, number CONF, 2010.

[18] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(Feb):235–284, 2008.

[19] Nimrod Megiddo. On the ball spanned by balls. *Discrete & Computational Geometry*, 4(6):605–610, 1989.

[20] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical programming*, 14:265–294, 1978.

[21] C Shannon. A mathematical theory of communication, bell system technical journal 27: 379-423 and 623–659. *Mathematical Reviews (MathSciNet): MR10, 133e*, 1948.

[22] Yaron Singer. Budget feasible mechanisms. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 765–774. IEEE, 2010.

[23] Jan Vondrák. A note on concentration of submodular functions. *arXiv preprint arXiv:1005.2791*, 2010.