Conference Paper

Classical and Classical-Quantum Sphere Packing Bounds: Rényi vs Kullback and Leibler

Author(s):
Dalai, Marco

Publication Date:
2016

Permanent Link:
https://doi.org/10.3929/ethz-a-010645622

Rights / License:
In Copyright - Non-Commercial Use Permitted

This page was generated automatically upon download from the ETH Zurich Research Collection. For more information please consult the Terms of use.
Classical and Classical-Quantum Sphere Packing
Bounds: Rényi vs Kullback and Leibler

Marco Dalai
Department of Information Engineering
University of Brescia
Via Branze 38, 25123, Brescia, Italy
Email: marco.dalai@unibs.it

Abstract—We review the use of binary hypothesis testing for the derivation of the sphere packing bound in channel coding, pointing out a key difference between the classical and the classical-quantum setting. In the first case, two ways of using the binary hypothesis testing are known, which lead to the same bound written in different analytical expressions. The first method (historically) compares output distributions by the codewords with an auxiliary fixed distribution, and naturally leads to an expression using the Rényi divergence. The second method compares the given channel with an auxiliary one and leads to an expression using the Kullback-Leibler divergence. In the classical-quantum case, due to a fundamental difference in the quantum binary hypothesis testing, these two approaches lead to two different bounds, the first being the “right” one. We discuss the details of this phenomenon, which suggests the question of whether auxiliary channels are used in the optimal decision region for message \( m \). Since \( Q \) is a distribution, for at least one \( m \) we have

\[
Q(Y_m) \leq \frac{1}{M} \tag{1}
\]

and

\[
\left( e^{-nR} \right). \tag{2}
\]

This second representation gives an intuitive interpretation of the bound. Roughly speaking, the probability of error for the optimal test is essentially due to those sequences in \( \mathcal{V}^n \) with empirical distribution close to \( P_s \), whose total probabilities under \( P_0 \) and \( P_1 \) vanish with exponents given by \( D(P_s\|P_0) \) and \( D(P_s\|P_1) \) respectively. One can notice that the problem of determining the trade-off of the error exponents in the test between \( P_0 \) and \( P_1 \) is essentially reduced to the problem of testing \( P_s \) against \( P_{i} \), \( i = 0, 1 \), in the Stein regime where \( P_{e_{i,s}} \) is bounded away from one.

I. CLASSICAL HYPOTHESIS TESTING

We start by recalling that in classical binary hypothesis testing between two distributions \( P_0 \) and \( P_1 \) on some set \( \mathcal{V} \), based on \( n \) independent extractions, the trade-off of the best possible exponents of error probabilities of the first and second kind can be expressed parametrically, for \( 0 < s < 1 \), as

\[
\frac{1}{n} \log P_{e_{0}} = -\mu(s) + s\mu'(s) + o(1) \tag{3}
\]

\[
\frac{1}{n} \log P_{e_{1}} = -\mu(s) - (1 - s)\mu'(s) + o(1), \tag{4}
\]

where

\[
\mu(s) = \log \sum_{v \in V} P_{0}(v)^{1-s}P_{1}(v)^{s}, \tag{5}
\]

is a scaled version of the Rényi divergence usually defined as

\[
D_{\alpha}(P||Q) = \frac{1}{\alpha - 1} \sum_{v \in V} P_{0}(v)Q^{1-\alpha}(v), \tag{6}
\]

so that \( \mu(s) = -sD_{1-s}(P_{0}\|P_{1}) \). An explicit computation - or just a different way of deriving the bound - shows that an equivalent expression is

\[
\frac{1}{n} \log P_{e_{0}} = D(P_s\|P_0) + o(1) \tag{7}
\]

\[
\frac{1}{n} \log P_{e_{1}} = D(P_s\|P_1) + o(1), \tag{8}
\]

where \( D(\cdot\|\cdot) \) is the Kullback-Leibler divergence

\[
D(P, Q) = \sum_{v} P(v) \log \frac{P(v)}{Q(v)} \tag{9}
\]

and \( P_s \) is the tilted mixture

\[
P_s(v) = \frac{P_{0}(v)^{1-s}P_{1}(v)^{s}}{\sum_{v'} P_{0}(v')^{1-s}P_{1}(v')^{s}}. \tag{10}
\]

This second representation gives an intuitive interpretation of the bound. Roughly speaking, the probability of error for the optimal test is essentially due to those sequences in \( \mathcal{V}^n \) with empirical distribution close to \( P_s \), whose total probabilities under \( P_0 \) and \( P_1 \) vanish with exponents given by \( D(P_s\|P_0) \) and \( D(P_s\|P_1) \) respectively. One can notice that the problem of determining the trade-off of the error exponents in the test between \( P_0 \) and \( P_1 \) is essentially reduced to the problem of testing \( P_s \) against \( P_{i} \), \( i = 0, 1 \), in the Stein regime where \( P_{e_{i,s}} \) is bounded away from one.

II. CLASSICAL SPHERE-PACKING

Given a discrete memoryless channel \( W : \mathcal{X} \to \mathcal{Y} \) with capacity \( C \), the sphere packing bound gives an exponential lower bound on the probability of error of codes at rate \( R < C \) in the form

\[
P_e \geq e^{-n(E_{w}(R)+o(n))}, \tag{11}
\]

where \( R \) is the coding rate, \( n \) the block length and \( E_{w}(R) \) is the so called sphere packing exponent. Two proofs are known for the classical version of the bound, which naturally lead to two equivalent yet different analytical expressions for the function \( E_{w}(R) \). A preliminary technical feature common to both procedures is that they both focus on some constant-composition sub-code which has virtually the same rate as the original code, but where all codewords have the same empirical composition \( P \). In both cases, then, the key ingredient is binary hypothesis testing (BHT).

A. The MIT proof

The first proof (see [1], [2]) is based on a binary hypothesis test between the output distributions \( X_m \) induced by the codewords \( x_1, \ldots, x_M \) and an auxiliary output product distribution \( Q = Q^{\otimes n} \) on \( \mathcal{Y}^n \). Let \( \mathcal{Y}_m \subseteq \mathcal{Y}^n \) be the decision region for message \( m \). Since \( Q \) is a distribution, for at least one \( m \) we have

\[
Q(\mathcal{Y}_m) \leq 1/M \tag{12}
\]

\[
= e^{-nR}. \tag{13}
\]

\[
198
\]
Considering a binary hypothesis test between $W_{xm}$ and $Q$, with $Y_m$ as decision region for $W_{xm}$, equation (1) gives an exponential upper bound on the probability of error under hypothesis $Q$ which implies a lower bound on the probability of error under hypothesis $W_{xm}$, which is $W_{xm}(Y_m)$, the probability of error for message $m$. Here the BHT is considered in the regime where both probabilities decrease exponentially. The standard procedure uses the first form of the bound mentioned in the previous section based on the Rényi divergence. The bound can be extended to the case of testing products of non-identical distributions; for the pair of distribution $W_{xm} = W_{xm,1} \otimes \cdots \otimes W_{xm,n}$ and $Q = Q \otimes \cdots \otimes Q$ it gives the performance of an optimal test in the form

$$-1/n \log P_{e|W_{xm}} = -\mu(s) + s\mu'(s) + o(1)$$

(3)

where

$$\mu(s) = \sum_x P(x) \left( \log \sum_{y \in \mathcal{Y}} W_x(y)^{1-s} Q(y)^s \right).$$

At this point the arguments in [1] and [2] diverge a bit; while the former is not rigorous, it has the advantage of giving the tight bound for the arbitrary codeword composition $P$. The latter is instead rigorous but only gives the tight bound for the optimal composition $P$. In [3] we proposed a variation which we believe to be rigorous and at the same time gives the tight bound for an arbitrary composition $P$. The need for this variation will be clear in the discussion of classical-quantum channels in the next section.

For the test based on the decoding region $Y_m$, the left hand side of (4) is lower bounded by $R$ due to (1). So, if we choose $s$ and $Q$ in such a way that the right hand side of (4) is roughly $R - \epsilon$, then $-(1/n) \log P_{e|W_{xm}}$ must be smaller than the right hand side of (3) computed for those same $s$ and $Q$ (for otherwise the decision region $Y_m$ would give a test strictly better than the optimal one).

This is obtained by choosing $Q$, as a function of $s$, as the minimizer of $-\mu(s)$ and then selecting $s$ which makes the right hand side of (4) (equal to $R - \epsilon$) accessible. Extracting $\mu'(s)$ from (4) in terms of $\mu(s)$ and $R$ and using it in (3), the probability of error for message $m$ is bounded in terms of $R$. After some tedious technicalities, cf. [3, Appendix A], we get

$$-1/n \log P_{e|W_{xm}} \leq \sup_{0 < s < 1} \left[ E_0(s, P) - \frac{s}{1-s} (R - \epsilon) \right] + o(1),$$

(5)

where

$$E_0(s, P) = \min_Q \left[ \frac{1}{s-1} \sum_x P(x) \log \sum_y W_x(y)^{1-s} Q(y)^s \right]$$

(6)

and

$$\min_Q \left[ \frac{s}{1-s} \sum_x P(x) D_{1-s}(W_x||Q) \right]$$

(7)

the minimum being over distributions $Q$ and $I_{s}(P, W)$ being the $\alpha$-mutual information as defined by Csiszár [4]. We thus find the bound, valid for codes with constant composition $P$

$$-1/n \log P_{e,max} \leq \sup_{0 < s < 1} \frac{1}{s} \left[ I_{1-s}(P, W) - R + \epsilon \right] + o(1).$$

(8)

It is worth pointing out that the chosen $Q$, which achieves the minimum in the definition of $E_0(s, P)$, satisfies the set of constraints (cf [1, eqs. (9.23), (9.24), (9.50)], [5, Cor. 3])

$$Q(y) = \sum_x P(x) V_x(y)$$

(9)

if we define $V_x(y)$ as

$$V_x(y) = \frac{W_x^{-s}(y) Q^s(y)}{\sum_{y'} W_x^{-s}(y') Q^s(y')}.$$  

(10)

So, the chosen $Q$ is such that its tilted mixtures with the distributions $W_x$ induce $Q$ itself on the output set $\mathcal{Y}$. Using the second representation of the error exponents in binary hypothesis testing mentioned in Section I (extended for independent extractions from non-identical distributions), we observe that the chosen $Q$ induces the construction of an auxiliary channel $V$ such that $I(P, V) = \sum_x P(x) D(V_x||Q) = R - \epsilon$. The second proof of the sphere packing bound, which is summarized in the next section, takes this line of reasoning as starting point.

### B. Haroutunian’s proof

In the second proof (see [6], [7]) one considers the performance of the given coding scheme for channel $W$ when used for an auxiliary channel $V$’ such that $I(P, V) < R$. Due to the strong converse for channel coding, when used with channel $V$ the coding scheme will incur an error probability $1 - o(1)$, which means that for at least one codeword $m$ we must have $V_{xm}(Y_m) = 1 - o(1)$. Applying the data processing inequality for the Kullback-Leibler divergence one thus finds that

$$V_{xm}(Y_m) \log \frac{V_{xm}(Y_m)}{W_{xm}(Y_m)} + V_{xm}(Y_m) \log \frac{V_{xm}(Y_m)}{W_{xm}(Y_m)} \leq n D(V||W|P),$$

from which

$$\log W_{xm}(Y_m) \geq -n D(V||W|P) + 1 + o(1).$$

So, the error exponent for channel $W$ is bounded as

$$-1/n \log P_{e|W_{xm}} \leq \min_{V : I(P, V) \leq R} D(V||W|P)(1 + o(1)).$$

Note that, thanks to the use of the strong converse, the data processing inequality is enough to get the desired result, but any converse for $V$ would work if followed by the more powerful Stein lemma.

The bound derived is precisely the same as in the previous section, and for the optimal choice of the channel $V$, if we
define the output distribution $Q = PV$ as in (9), then (10) is satisfied for some $s$. So, we notice that the two proofs really rely on a comparison between the original channel and equivalent auxiliary channels/distributions. In the first procedure we start with an auxiliary distribution $Q$, but we find that the optimal choice of $Q$ is such that the tilted mixtures with the $W_s$ distributions are the $V_s$ which give $PV = Q$. In the second procedure we start with the auxiliary channel $V$ but we find that the optimal $V$ induces an output distribution $Q$ whose tilted mixtures with the $W_s$ are the $V_s$ themselves. It is worth noticing that in this second procedure we use a converse for channel $V$; hidden in this step we are using the output distribution $Q$ induced by $V$.

These observations point out that while the MIT proof follows the first formulation of the binary hypothesis testing bound in terms of Rényi divergences, Haroutunian’s proof exploits the second formulation based on Kullback-Leibler divergences, but the compared quantities are equivalent. There seems to be no reason to prefer the first procedure given the simplicity of the second one.

III. QUANTUM HYPOTHESIS TESTING

In a binary hypothesis testing between two density operators $\sigma_0$ and $\sigma_1$, based on $n$ independent extractions (but with global measurement), the error exponents of the first and second kind can be expressed parametrically as (see [8])

\[
-\frac{1}{n} \log P_{e|\sigma_0} = -\mu(s) + s\mu'(s) + o(1),
\]

\[
-\frac{1}{n} \log P_{e|\sigma_1} = -\mu(s) - (1 - s)\mu'(s) + o(1),
\]

where, in complete analogy with the classical case,

\[
\mu(s) = \log \text{Tr} \sigma_0^{1-s}\sigma_1^s.
\]

Upon differentiation, one finds for example for (11)

\[
-\frac{1}{n} \log P_{e|\sigma_0} = -\log \text{Tr}(\sigma_0^{1-s}\sigma_1^s) + \text{Tr} \left[ \frac{\sigma_0^{1-s}\sigma_1^s}{\text{Tr} \sigma_0^{1-s}\sigma_1^s} \right] (\log \sigma_1^s - \log \sigma_0^s) + o(1).
\]

When $\sigma_0$ and $\sigma_1$ commute, that is, in the classical case, we can define the density operator

\[
\sigma_s = \frac{\sigma_0^{1-s}\sigma_1^s}{\text{Tr} \sigma_0^{1-s}\sigma_1^s}
\]

and use the property $\log \sigma_1^s - \log \sigma_0^s = \log \sigma_0^{1-s}\sigma_1^s - \log \sigma_0$ to obtain

\[
-\frac{1}{n} \log P_{e|\sigma_0} = \text{Tr} \sigma_s (\log \sigma_s - \log \sigma_0) + o(1) = D(\sigma_s||\sigma_0) + o(1).
\]

In a similar way we find

\[
-\frac{1}{n} \log P_{e|\sigma_1} = D(\sigma_s||\sigma_1) + o(1).
\]

This is indeed the second form of the bound as mentioned already in Section I. However, if $\sigma_0$ and $\sigma_1$ do not commute, the above simplification is not possible. Hence, the two error exponents cannot be expressed in terms of the Kullback-Leibler divergence. So, unlike in the classical binary hypothesis testing, the problem of determining the trade-off of the error exponents in the test between $\sigma_0$ and $\sigma_1$ cannot be reduced to the problem of testing some $\sigma_s$ against $\sigma_i$, $i = 0, 1$, in the Stein regime.

To verify that this is really a property of the quantum binary hypothesis testing and not an artificial effect of the used procedure, it is useful to consider the case of pure states, that is when operators $\sigma_0$ and $\sigma_1$ have rank 1, say $\sigma_0 = |\psi_0\rangle\langle\psi_0|$ and $\sigma_1 = |\psi_1\rangle\langle\psi_1|$, with non-orthogonal $\psi_0$ and $\psi_1$. In this case, $\sigma_0^{1-s} = \sigma_0$ and $\sigma_1^s = \sigma_1$, so that one simply has

\[
\mu(s) = \log \text{Tr} \sigma_0 \sigma_1 = \log |\langle\psi_0|\psi_1\rangle|^2
\]

and consequently the two error exponents both equal $-\log |\langle\psi_0|\psi_1\rangle|^2$. These quantity cannot be expressed as $D(\sigma_s||\sigma_i)$, $i = 0, 1$, for any $\sigma_s$ because

\[
D(\rho||\sigma_i) = \begin{cases} 0 & \rho = \sigma_i, \\ +\infty & \rho \neq \sigma_i, \quad i = 0, 1, \end{cases}
\]

since $\sigma_0$ and $\sigma_1$ are pure.

IV. CLASSICAL-QUANTUM SPHERE-PACKING

The different behavior of binary hypothesis testing in the quantum case with respect to the classical has a direct impact on the sphere packing bound for classical-quantum channels. Both the MIT and Haroutunian’s approaches can be extended to this setting, but the resulting bounds are different. In particular, since the binary hypothesis testing is correctly handled with the Rényi divergence formulation, the MIT form of the bound extends to what one expects as the right generalization (in particular, it matches known achievability bounds for pure-state channels), while Haroutunian’s form extends to a weaker bound. It was already observed in [9] that the latter gives a trivial bound for all pure state channels, which is a direct consequence of what already shown for the simple binary hypothesis testing in the previous section.

It is useful to investigate this weakness at a deeper level in order to see clearly where the problem really is. Let $W_x$, $x \in X$, be now density operators, let $W_x = W_{x_1} \otimes \cdots \otimes W_{x_n}$ be the state associated to a sequence $x$, and thus $W_{x_1}, \ldots, W_{x_M}$ the states associated to the $M$ messages, where $M = e^{nR}$. Let $\{\Pi_1, \Pi_2, \ldots, \Pi_M\}$ be the POVM used at the receiver for channel $W$, which means that the probability of decoding $m'$ when $m$ is sent is $\text{Tr} \Pi_m W_{x_m}$. Consider then an auxiliary classical-quantum channel with states $V_x$ and with capacity $C < R$. The strong converse still holds for channel $V$ which implies that for any decoding rule, at least one message the probability of error is $1 - o(1)$. In particular for the given POVM, for at least one $m$, $\text{Tr}(I - \Pi_m) V_{x_m} = 1 - o(1)$.

\footnote{More precisely, a correct formulation is that at least one of the two error exponents is not larger than $-\log |\langle\psi_0|\psi_1\rangle|^2$.}
Using again a data processing inequality for the quantum Kullback-Leibler divergence one then finds as in the classical case that
\[
\log \text{Tr}(I - \Pi_m)W_{\alpha m} \geq -\frac{nD(V||W|P) + 1}{1 + o(1)},
\]
and thus
\[
\frac{1}{n} \log P_e|W_{\alpha m}| \leq \min_{V : I(P,V) \leq R} D(V||W|P)(1 + o(1)).
\]
The problem is that if \( W \) is a pure state channel, at rates \( R < C \) any auxiliary channel \( V \not= W \) gives \( D(V||W|P) = \infty \), so that the bound is trivial for all pure state channels. It is important to observe that this is not due to a weakness in the used data processing inequality. In a binary hypothesis test between the the pure state \( W_{\alpha m} \) and a state \( V_{\alpha m} \) built from a different channel \( V \), one can notice that the POVM \( \{A, I - A\} \) with \( A = W_{\alpha m} \) satisfies
\[
\text{Tr}(I - A)V_{\alpha m} = 1 - o(1), \quad \text{Tr}(I - A)W_{\alpha m} = 0.
\]
So, it is really impossible to deduce a positive lower bound for \( \text{Tr}(I - \Pi_m)W_{\alpha m} \) using only the fact that \( \text{Tr}(I - \Pi_m)V_{\alpha m} = 1 - o(1) \).

It is also worth checking what happens with the MIT procedure. All the steps can be extended to the classical-quantum case (see [3] for details) leading to a bound which has the same form as (5) where \( E_0(s,P) \) is defined in analogy with (6) as
\[
E_0(s,P) = \min_Q \left[ \frac{1}{s - 1} \sum_x P(x) \log \text{Tr} W_1^{1-s}Q^x \right] = \min_Q \left[ \frac{1}{s - 2} \sum_x P(x) D_{1-s}(W_{\alpha m}||Q) \right],
\]
the minimum being over all density operators \( Q \), and \( D_{1-s}(\cdot||\cdot) \) being the quantum Rényi divergence. However, as far as we know, there is no analog of equations (9) and (10), and the optimizing \( Q \) does not induce an auxiliary \( V \) such that \( I(P,V) = R - \epsilon \).

V. Auxiliary Channels and Strong Converse

We have presented the two main approaches to sphere packing as different procedures which are equivalent in the classical case but not in the classical-quantum case. However, it is actually possible to consider the two approaches as particular instances of one general approach where the channel \( W \) is compared to an auxiliary channel \( V \), since the auxiliary distribution/state \( Q \) can be considered as a channel with constant \( V_x = Q \). This principle is very well described in [10], where it is shown that essentially all known converse bounds in classical channel coding can be cast in this framework.

According to this interpretation, the starting point in Haroutunian’s proof is general enough to include the MIT approach as a special case. So, the weakness of the method in the classical-quantum case must be hidden in one of the intermediate steps. It is not difficult to notice that the key point is how the strong converse is used in Haroutunian’s proof. The general auxiliary channel \( V \) is only assumed to have capacity \( C < R \), and the strong converse for \( V \) which is used is of the simple form \( P_e = 1 - o(1) \), which is good enough in the classical case. In the MIT proof, instead, the auxiliary channel is such that \( C = 0 \), so that the strong converse takes another simple form, \( P_e \geq 1 - e^{-nR} \). The critical point is that in the classical-quantum setting a converse of the form \( P_e = 1 - o(1) \) for \( V \) does not lead to a lower bound on \( P_e \) for \( W \) in general. What is needed is a sufficiently fast exponential convergence to 1 of \( P_e \) for channel \( V \), which essentially suggests that \( V \) should be chosen with capacity not too close to \( R \) and that the exact strong converse exponent for \( V \) should be used.

The natural question to ask at this point is what the optimal auxiliary channel is when the exact exponent of the strong converse is used. At high rates the question is not really meaningful for all those cases where the known versions of the sphere packing bound coincide with achievability results, that is, for classical channels and for pure state channels. However in the remaining cases, that is, in the low rate region for the mentioned channels or in the whole range of rates \( 0 < R < C \) for general non commuting mixed-state channels, the question is legitimate. In the classical case, since the choice of an (optimal) auxiliary channel with \( C = 0 \) or \( C = R^- \) leads to the same result, one might expect that any other intermediate choice would give the same result. However, to the best of our knowledge this has never been clarified in the literature.

For classical-quantum channels, the question is perhaps not trivial; it is worth pointing out that even the exact strong converse exponent has been determined only very recently [11]. What is very interesting is that while in the classical case the strong converse exponent for \( R > C \) is expressed in terms of Rényi divergence [12], [13], similarly as error exponents for \( R < C \), for classical-quantum channels the strong converse exponents are expressed in terms of the so called “sandwiched” Rényi divergence defined by
\[
\tilde{D}_\alpha(\rho, \sigma) = \frac{1}{\alpha - 1} \log \text{Tr} \left( \left( \frac{1 - \alpha}{\alpha - 1} \rho^{1/\alpha} \sigma^{1/\alpha} \right)^{\alpha} \right).
\]
The problem to consider would thus be more or less as follows. Consider an auxiliary channel \( V \) with capacity \( C < R \) and evaluate its strong converse exponent in terms of sandwiched Rényi divergences. Fix this exponent as the probability of error under hypothesis \( W_{\alpha m} \) in a test between \( W_{\alpha m} \) and \( V_{\alpha m} \), where \( \Pi_m \) is the operator in favor of \( W_{\alpha m} \) and \( I - \Pi_m \) is the one in favor of \( V_{\alpha m} \). Then deduce a lower bound for the probability of error under hypothesis \( W_{\alpha m} \) using the standard binary hypothesis testing bound in terms of Rényi divergences. It is not entirely clear, to this author, that the optimal auxiliary channel should necessarily always be the one such that \( C = 0 \) as used up to now. Since for non commuting mixed-state channels the current known form of sphere packing bound is not yet matched by any achievability result, one cannot exclude that it is not the tightest possible form.

\footnote{Here we mean optimal memoryless channel for bounding the error exponent in the asymptotic regime.}
REFERENCES


