MIT Open Access Articles

Saddle Point in the Minimax Converse for Channel Coding

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation	Polyanskiy, Yury. Saddle Point in the Minimax Converse for Channel Coding. IEEE Transactions on Information Theory 59, no. 5 (May 2013): 2576-2595.
As Published	http://dx.doi.org/10.1109/TIT.2012.2236382
Publisher	Institute of Electrical and Electronics Engineers
Version	Author's final manuscript
Citable link	http://hdl.handle.net/1721.1/79372
Terms of Use	Creative Commons Attribution-Noncommercial-Share Alike 3.0
Detailed Terms	http://creativecommons.org/licenses/by-nc-sa/3.0/

Massachusetts Institute of Technology

DSpace@MIT

Saddle point in the minimax converse for channel coding

Yury Polyanskiy

***Abstract*—A minimax meta-converse has recently been proposed as a simultaneous generalization of a number of classical results and a tool for the non-asymptotic analysis. In this paper it is shown that the order of optimizing the input and output distributions can be interchanged without affecting the bound. In the course of the proof, a number of auxiliary results of separate interest are obtained. In particular, it is shown that the optimization problem is convex and can be solved in many cases by the symmetry considerations. As a consequence it is demonstrated that in the latter cases the (multi-letter) input distribution in information-spectrum (Verdú-Han) converse bound can be taken to be a (memoryless) product of single-letter ones. A tight converse for the binary erasure channel is re-derived by computing the optimal (non-product) output distribution. For discrete memoryless channels, a conjecture of Poor and Verdú regarding the tightness of the information spectrum bound on the error-exponents is resolved in the negative. Concept of the channel symmetry group is established and relations with the definitions of symmetry by Gallager and Dobrushin are investigated.**

I. Introduction

The meta-converse method proposed in [1, Sections III.E-III.G] has been successfully applied to prove impossibility results in problems of point-to-point channel coding [1], communication with feedback [2], energy-efficient transmission [3], generalized to lossy source compression [4], multiple-access communication [5], quantum-assisted coding [6] and several other problems [7]–[9]. Most of these applications employed a particular variation of the general method – *a minimax converse*. The focus of the present paper is to provide general results on and techniques for *exact* evaluation of the minimax converse bound.

Exact evaluation is important from several viewpoints. First, in the domain of finite blocklength analysis it is preferable to isolate provably optimal bounds, so that time-consuming numerical evaluations are carried out only for them. Since the minimax converse dominates a number of other results [10, Section 2.7.3], its evaluation becomes crucial. Second, theoretically it is required to understand what (multi-letter) input distribution optimizes the converse bound. This problem is emphasized by information-spectrum converse bounds, such as the one by Verdú and Han [11], in which it is not clear whether even for a memoryless channel one may restrict optimization to memoryless input distributions. In this paper this is positively resolved for symmetric channels. Satisfyingly, we find out that the optimal (multi-letter) input distribution coincides with (power of) the capacity achieving one. Next, for the characterization of the third (logarithmic) term in the expansion of the maximum achievable rate, see [10, Section 3.4.5] and [7], a common technique of reduction to constant-composition subcodes results in loose estimates of the third term. Thus, for this question knowledge of the optimal input distribution in the minimax converse is also crucial.

Consider an abstract channel coding problem, that is a random transformation defined by a pair of measurable spaces of inputs A and outputs B and a conditional probability measure $P_{Y|X} : \mathsf{A} \mapsto \mathsf{B}$. Let M be a positive integer and

$$0 \le \epsilon \le 1 - \frac{1}{M}\,. \tag{1}$$

An (M, ϵ) *code* the random transformation $(\mathsf{A}, \mathsf{B}, P_{Y|X})$ is a pair of (possibly randomized) maps $\mathsf{f} : \{1, \dots, M\} \to \mathsf{A}$ (the encoder) and $\mathsf{g} : \mathsf{B} \to \{1, \dots, M\}$ (the decoder), satisfying

$$\frac{1}{M}\sum_{m=1}^{M} P[g(Y) \neq m | X = f(m)] \le \epsilon\,. \tag{2}$$

In practical applications, we take A and B to be n-fold Cartesian products of alphabets $\mathcal{A}$ and $\mathcal{B}$, and a channel to be a sequence of random transformations $\{P_{Y^n|X^n} : \mathcal{A}^n \to \mathcal{B}^n\}$ [11]. In this paper, however, it is preferable not to assume that A and B have any structure such as a Cartesian product.

Given a pair of distributions P and Q on common measurable space W, a randomized test between those two distributions is defined by a random transformation $P_{Z|W} : \mathsf{W} \mapsto \{0, 1\}$ where 0 indicates that the test chooses Q. In the Neyman-Pearson (non-Bayesian) formulation, to a pair of P and Q we associate the fundamental region of the unit square defined as

$$\mathcal{R}(P, Q) \stackrel{\triangle}{=} \{(\alpha, \beta) \colon \exists P_{Z|W} \colon \alpha = P[Z = 1], \beta = Q[Z = 1]\}\,. \tag{3}$$

Clearly, $\mathcal{R}(P, Q)$ is closed convex, contains the diagonal and is fixed by the symmetry $(\alpha, \beta) \mapsto (1 - \alpha, 1 - \beta)$, see [12, Section 3.2 and Fig. 3.1]. The lower boundary of $\mathcal{R}(P, Q)$ is denoted by

$$\beta_\alpha(P, Q) \stackrel{\triangle}{=} \min\{\beta : (\alpha, \beta) \in \mathcal{R}(P, Q)\} \tag{4}$$

$$= \min \int P_{Z|W}(1|w) Q(dw)\,, \tag{5}$$

where the minimum is over all probability distributions $P_{Z|W}$ satisfying

$$P_{Z|W} : \int P_{Z|W}(1|w) P(dw) \ge \alpha\,. \tag{6}$$

The author is with the Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 02139 USA. e-mail: yp@mit.edu. This work supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

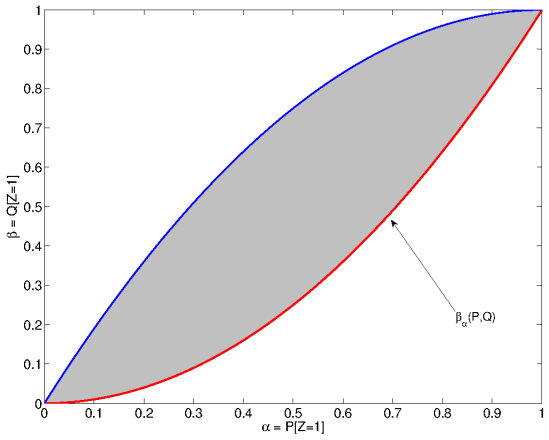

Fig. 1: Relation between the hypothesis testing region $\mathcal{R}(P,Q)$ and the Neyman-Pearson function $\beta_\alpha(P,Q)$ (schematic).

The minimum in (5) is guaranteed to be achieved by the Neyman-Pearson lemma. In other words, $\beta_\alpha(P,Q)$ gives the minimum probability of error under hypothesis Q if the probability of error under hypothesis P is not larger than $1-\alpha$. Because of the mentioned symmetry and closedness, knowledge of $\beta_\alpha(P,Q)$ is enough to reconstruct the entire $\mathcal{R}(P,Q)$ and, moreover, $\alpha \mapsto \beta_\alpha(P,Q)$ is a convex continuous function on $[0,1]$. This is illustrated on Fig. 1.[1]

In [1] it was shown that a number of classical converse bounds, including Fano's inequality, Shannon-Gallager-Berlekamp, Wolfowitz strong converse and Verdú-Han information spectrum converse, can be obtained in a unified manner as a consequence of the meta-converse theorem [1, Theorem 26]. One of such consequences is the following minimax converse [1]:

Theorem 1 (minimax converse): Every (M,ϵ) code for the random transformation $P_{Y|X}$ satisfies

$$\left(1-\epsilon, \frac{1}{M}\right) \in \bigcup_{P_X} \bigcap_{Q_Y} \mathcal{R}(P_{XY}, P_X \times Q_Y)\,.$$

In particular,

$$\frac{1}{M} \geq \inf_{P_X} \sup_{Q_Y} \beta_{1-\epsilon}(P_{XY}, P_X \times Q_Y)\,, \tag{7}$$

where P_X ranges over all input distributions on A, Q_Y ranges over all output distributions on B and $P_{XY} = P_X P_{Y|X}$ denotes the joint distribution on $\mathsf{A}\times\mathsf{B}$:

$$P_{XY}(dx,dy) = P_X(dx)P_{Y|X}(dy|x)\,. \tag{8}$$

In this paper we discuss the problem of exact computation of the minimax problem in (7). This is unlike the majority of applications of Theorem 1 (for example, those discussed above) , in which one selects a convenient Q_Y and then proves a lower bound on $\beta_\alpha(P_{XY}, P_X Q_Y)$ independent of P_X. In essence, such an argument invokes a looser bound (10) evaluated at only one Q_Y:

$$\frac{1}{M} \geq \inf_{P_X} \sup_{Q_Y} \beta_{1-\epsilon}(P_{XY}, P_X \times Q_Y) \tag{9}$$

$$\geq \sup_{Q_Y} \inf_{P_X} \beta_{1-\epsilon}(P_{XY}, P_X \times Q_Y)\,. \tag{10}$$

Our primary goal is to develop tools to evaluate the optimizing P_X, Q_Y and the values in (9)-(10), instead of relying on a "lucky guess" of a good Q_Y. The paper is structured as follows:

1) Section II-A shows that the inner optimization in (9) is equivalent to solving a composite hypothesis testing problem. This is a simple consequence of the Wald-LeCam theory of completeness of Bayes rules in minimax decision problems [13], [14].
2) Optimal composite tests correspond exactly to non-signalling assisted (NSA) codes, thereby explaining the mysterious result of W. Matthews [6] that NSA codes achieve the minimax meta-converse bound (9) with equality (Section II-B).
3) Next we proceed to studying general properties of the function
$$P_X \mapsto \beta_\alpha(P_{XY}, P_X \times Q_Y)\,.$$
It is shown that this function is convex (Section III-A), continuous in the topology of total variation (Section III-B) and under regularity assumptions weakly continuous (Section III-C). It is also shown that functions of ϵ appearing in the right-hand sides of (9) and (10) are convex.
4) The bound (10) is simplified by replacing the domain of the inner optimization with the elements of A (instead of measures on A) and taking the convex hull (Section IV).
5) For compact (in particular, finite) A a simple consequence of the convexity-continuity results in Section III and Fan's minimax theorem [15] is the *saddle point* property for β_α:
$$\min_{P_X} \max_{Q_Y} \beta_{1-\epsilon}(P_{XY}, P_X \times Q_Y) = \max_{Q_Y} \min_{P_X} \beta_{1-\epsilon}(P_{XY}, P_X \times Q_Y)\,. \tag{11}$$
In Section V the result is extended to non-compact A. Thus, under regularity conditions the bounds (9) and (10) are equal.
6) Next, we discuss how the general concept of channel symmetry can be defined and how it simplifies calculation of the optimal P_X and Q_Y (Section VI-A).
7) Classes of symmetric channels and their inter-relations are discussed in Section VI-B.
8) The saddle point is computed for the binary symmetric channel (BSC) in Section VI-C, for the additive white Gaussian noise (AWGN) channel in Section VI-F and for the binary erasure channel (BEC) in Section VI-D. Interestingly, for the latter we discover that the optimal Q_Y is not a product distribution despite the channel being memoryless.
9) For discrete memoryless channels (DMC) the bound (9) exponentially coincides with the sphere-packing bound

[1]Note that some authors prefer α to carry meaning of the probability of error, while β denotes the probability of success. The resulting region, however, is the same: e.g., compare Fig. 1 with [12, Fig. 3.1].

of Shannon, Gallager and Berlekamp [16]. This resolves the conjecture of Poor-Verdú [17] regarding the tightness of their bound on the error-exponents (Section VI-E).

10) Discussion and general remarks conclude the paper (Section VII).

As suggested by the title, our exposition focuses on deriving the saddle point result (11) in Section V. One reason we emphasize this result among others is that we see it as a non-asymptotic analog of the classical characterization of channel capacity:

$$C = \max_{P_X}\min_{Q_Y} D(P_{Y|X}||Q_Y|P_X) \quad (12)$$

$$= \min_{Q_Y}\max_{P_X} D(P_{Y|X}||Q_Y|P_X)\,. \quad (13)$$

In fact, this analogy is to be expected as for memoryless channels Stein's lemma shows that

$$\beta_\alpha(P_{XY}^n, P_X^n Q_Y^n) = \exp\{-nD(P_{Y|X}||Q_Y|P_X) + o(n)\}\,.$$

Notation and assumptions: Throughout this paper we assume that there exists a σ-finite measure μ such that the kernel $P_{Y|X}$ is given by

$$P_{Y|X}[E|x] \stackrel{\triangle}{=} \int_E \rho(y|x)\mu(dy) \quad (14)$$

for some measurable function $\rho : \mathsf{A}\times\mathsf{B}\to\mathbb{R}$ and that all singletons $\{x\}$ and $\{y\}$ are measurable subsets of A and B. Criteria for satisfying condition (14) are discussed in [12, Section A.4]. We also denote by $\mathcal{M}(\mathsf{A})$ the set of all finite signed (countably-additive) measures on A, $\mathcal{M}_+(\mathsf{A})$ the subset of positive measures, and by $\mathcal{M}_1(\mathsf{A})$ the set of all probability measures. Absolute continuity of measure μ with respect to ν is denoted as $\mu\ll\nu$, and we write $\mu\sim\nu$ for the case when $\mu\ll\nu$ and $\nu\ll\mu$. We specify distributions of random variables as $X\sim P_X$, e.g. $W\sim\mathcal{N}(0,1)$ defines W to be standard Gaussian.

II. Composite hypothesis testing problem

Fix a distribution P_X and a random transformation $P_{Y|X}$ and consider a (simple vs. composite) hypothesis testing problem:

$$H_0 : (X,Y)\sim P_{XY} \quad (15)$$

$$H_1 : X\sim P_X \text{ and independent of } Y\,, \quad (16)$$

that is under H_1 the pair (X,Y) can be distributed according to $P_X\times Q_Y$ with an arbitrary Q_Y. Following the minimax formulation to each randomized test $P_{Z|XY}:\mathsf{A}\times\mathsf{B}\to\{0,1\}$ we associate a pair of numbers

$$\alpha = P_{XY}[Z=1]\,, \quad (17)$$

$$\beta = \sup_{Q_Y} P_X Q_Y[Z=1]\,, \quad (18)$$

where we adopted an intuitive notation

$$P_{XY}[Z=1] = \int_{\mathsf{A}}\int_{\mathsf{B}} P_{Z|XY}(1|x,y)P_X(dx)P_{Y|X}(dy|x) \quad (19)$$

$$P_X Q_Y[Z=1] = \int_{\mathsf{A}}\int_{\mathsf{B}} P_{Z|XY}(1|x,y)P_X(dx)Q_Y(dy)\,. \quad (20)$$

Analogous to (3) we define the fundamental region associated to this hypothesis testing problem as

$$\tilde{\mathcal{R}}(P_X, P_{Y|X}) = \{(\alpha,\beta) : \exists P_{Z|XY} \text{ s.t. (17)-(18) hold}\}$$

and its lower boundary

$$\tilde{\beta}_\alpha(P_X, P_{Y|X}) \stackrel{\triangle}{=} \inf\{\beta : (\alpha,\beta)\in\tilde{\mathcal{R}}(P_X,P_{Y|X})\} \quad (21)$$

To describe region $\tilde{\mathcal{R}}(P_X,P_{Y|X})$, first notice that it clearly contains the diagonal $\{(\alpha_0,\alpha_0),\alpha_0\in[0,1]\}$, which corresponds to trivial tests $P'_{Z|XY}(1|x,y)=\alpha_0$. Next, for an arbitrary test $P_{Z|XY}$ we may consider

$$P'_{Z|XY} = (1-\lambda)P_{Z|XY} + \lambda\alpha_0\,, \qquad \lambda\in[0,1]\,,$$

which demonstrates that $\mathcal{R}$ contains a line segment connecting any of its points to a point (α_0,α_0). Hence, $\mathcal{R}$ does not have "holes" (formally, has diagonal as its strong deformation retract) and it suffices to describe its upper and lower boundary.

In this paper we will only be concerned with the lower boundary of $\tilde{\mathcal{R}}$, described by (21). For completeness, though, we briefly inspect the upper boundary, whose height at α corresponds to finding

$$\sup_{P_{Z|XY}}\sup_{Q_Y} P_X Q_Y[Z=1] = \sup_{Q_Y}\sup_{P_{Z|XY}} P_X Q_Y[Z=1]$$

taken over all tests with $P_{XY}[Z=1]\ge\alpha$. It is possible to show that this supremum for $\alpha>0$ is given by

$$\alpha\mapsto\sup_{y_0\in\mathsf{B}}\min\left(\frac{\alpha}{P_Y(y_0)},1\right)\,.$$

Thus, when $\min_{y_0}P_Y(y_0)$ exists the upper boundary consists of two linear segments $(0,0)\to(\min P_Y(y_0),1)\to(1,1)$ and is contained inside $\tilde{\mathcal{R}}$. If $\inf_{y_0}P_Y(y_0)=0$ not achievable at any y_0 then, the boundary is $(0,0)\to(0,1)\to(1,1)$ but the vertical segment (except for the origin) does not belong to $\mathcal{R}$[2]. Thus, the portion of $\tilde{\mathcal{R}}$ above the diagonal is convex but maybe non-closed. Also, we note that unlike $\mathcal{R}$ the region $\tilde{\mathcal{R}}$ does not have the symmetry $(\alpha,\beta)\mapsto(1-\alpha,1-\beta)$. This fact is especially clear if one considers an example with $|\mathsf{A}|=1$.

The lower boundary, parametrized by $\alpha\mapsto\tilde{\beta}_\alpha$, is not as elementary. It is also convex, since for any two points (α_j,β_j), $j=0,1$ and corresponding tests $P_{Z_j|XY}$ we may consider

$$P_{Z|XY} = \lambda P_{Z_1|XY} + (1-\lambda)P_{Z_0|XY}$$

which according to (17)-(18) achieves $\alpha=\lambda\alpha_1+(1-\lambda)\alpha_0$ and

$$\beta\le\lambda\beta_1+(1-\lambda)\beta_0\,.$$

Thus, function $\tilde{\beta}_\alpha$ is convex on $[0,1]$ and thus continuous on $[0,1)$. In fact, we show next it is also continuous at $\alpha=1$ and the lower boundary $(\alpha,\tilde{\beta}_\alpha)$ is contained in $\mathcal{R}$.

To that end consider the following result:

Proposition 2: For any test $P_{Z|XY}$ we have

$$\sup_{Q_Y} P_X Q_Y[Z=1] = \sup_{y\in\mathsf{B}}\int_{\mathsf{A}} P_{Z|XY}(1|x,y)P_X(dx)\,. \quad (22)$$

[2]For example, $|\mathsf{A}|=1$ and P_Y is geometric distribution on positive integers.

Furthermore, any test $P_{Z|XY}$ can be modified to $P_{Z'|XY}$ such that

$$P_{XY}[Z'=1] = P_{XY}[Z=1] \quad (23)$$
$$\sup_{Q_Y} P_X Q_Y[Z'=1] \le \sup_{Q_Y} P_X Q_Y[Z=1] \quad (24)$$

and $P_{Z'|XY}$ is regular in the sense that

$$\sup_{Q_Y} P_X Q_Y[Z'=1] = \operatorname*{esssup}_{y\in\mathsf{B}} \int_{\mathsf{A}} P_{Z|XY}(1|x,y)P_X(dx) \quad (25)$$
$$= \sup_{Q_Y \ll \mu} P_X Q_Y[Z=1] \quad (26)$$

where essential supremum esssup is taken with respect to μ.

Proof: (22) follows by linearity of $Q_Y \mapsto P_X Q_Y[Z=1]$ and assumption of measurability of singletons $\{y\}$. Denote

$$h(y) \stackrel{\triangle}{=} \int_{\mathsf{A}} P_{Z|XY}(1|x,y)P_X(dx)\,.$$

Since h is measurable, e.g. [18, Proposition 1.6.9] , its essential supremum is well defined. We set

$$P_{Z'|XY}(1|x,y) \stackrel{\triangle}{=} P_{Z|XY}(1|x,y)1\{h(y) \le \operatorname{esssup} h\}\,.$$

Since $P_{Z'|XY}(1|x,y) \le P_{Z|XY}(1|x,y)$ everywhere and equality holds $P_X \times \mu$-almost everywhere, the (23)-(24) are satisfied. Then (25) follows from (22) applied to $P_{Z'|XY}$.

From (25) we obtain (26) by noticing that $L_\infty(\mathsf{B},\mu)$ is the dual of $L_1(\mathsf{B},\mu)$ and thus for any $f \in L_\infty(\mathsf{B},\mu)$ we have

$$\operatorname{esssup}|f| \stackrel{\triangle}{=} ||f||_\infty = \sup_{g:||g||_1=1} \int_f g(y)f(y)\mu(dy)\,,$$

while every $g \in L_1$ is naturally identified with $Q_Y \ll \mu$. ∎

By Proposition (2) we conclude that for the purpose of evaluating $\tilde\beta_\alpha(P_X, P_{Y|X})$ we may replace each $P_{Z|XY}$ with its regularization $P_{Z'|XY}$ and restrict supremization in (21) to $Q_Y \ll \mu$. Thus, the set of regularized tests $P_{Z|XY}$ is naturally identified with a closed convex subset of $L_\infty(\mathsf{A}\times\mathsf{B}, P_X\times\mu)$, while the set of $P_X \times Q_Y$ (with $Q_Y \ll \mu$) is identified with a closed convex subset of $L_1(\mathsf{A}\times\mathsf{B}, P_X\times Q_Y)$. Considering the standard dual pairing between these two spaces and a standard weak-$*$ compactness result of Banach and Alaoglu we conclude that the set of all tests is convex and compact in the topology induced by L_1, cf. [12, Theorem A.5.1]. Thus, the infimum in (21) is attained and we obtain a simplified characterization:

$$\tilde\beta_\alpha(P_X, P_{Y|X}) \stackrel{\triangle}{=} \min\{\beta : (\alpha,\beta) \in \tilde{\mathcal{R}}(P_X, P_{Y|X})\} \quad (27)$$
$$= \min_{P_{Z|XY}} \sup_{Q_Y \ll \mu} P_X Q_Y[Z=1]\,, \quad (28)$$

where the minimum is over all non-negative L_∞ functions $(x,y) \mapsto P_{Z|XY}(1|x,y)$ satisfying

$$\int_{\mathsf{A}}\int_{\mathsf{B}} P_{Z|XY}(1|x,y)P_X(dx)P_{Y|X}(dy|x) \ge \alpha\,. \quad (29)$$

Correspondingly, $\tilde{\mathcal{R}}$ contains its lower boundary, and function $\alpha \mapsto \tilde\beta_\alpha$ is convex and continuous on $[0,1]$.

A. Relation to minimax converse

Composite hypothesis testing region can be used to bound performance of error-correcting codes as follows:

Theorem 3: Every (M,ϵ) code for the random transformation $P_{Y|X}$ satisfies

$$\frac{1}{M} \ge \inf_{P_X} \tilde\beta_{1-\epsilon}(P_X, P_{Y|X})\,, \quad (30)$$

where P_X ranges over all input distributions on A.

Proof: Let P_X be the distribution induced by the encoder f with message equiprobable on $\{1,\dots,M\}$. Derivation of Theorem 1 in [1] consisted of noticing that any code (f,g) defines a hypothesis test

$$P_{Z|XY}(1|x,y) = 1\{\mathsf{f}(x) = \mathsf{g}(y)\}$$

for P_{XY} vs. $P_X Q_Y$ with parameters $\alpha \ge 1-\epsilon$ and $\beta = \frac{1}{M}$. Clearly this test has the same parameters for the composite hypothesis test (15)-(16). Thus, $(1-\epsilon, \frac{1}{M})$ belongs to $\tilde{\mathcal{R}}(P_X, P_{Y|X})$ and must satisfy (30). ∎

Immediately from the definition we notice that

$$\tilde\beta_\alpha(P_X, P_{Y|X}) \ge \sup_{Q_Y} \beta_\alpha(P_{XY}, P_X Q_Y)\,.$$

Thus Theorem 3 is at least as strong as Theorem 1. It turns out the two are equivalent:

Theorem 4: For any P_X and $P_{Y|X}$ the lower boundaries of $\tilde{\mathcal{R}}(P_X, P_{Y|X})$ and $\bigcap_{Q_Y} \mathcal{R}(P_{XY}, P_X Q_Y)$ coincide:

$$\tilde\beta_\alpha(P_X, P_{Y|X}) = \sup_{Q_Y} \beta_\alpha(P_{XY}, P_X Q_Y)\,. \quad (31)$$

Proof: In fact, (31) simply expresses the classical fact in statistical decision theory that an optimal minimax rule can be arbitrarily well approximated in the class of Bayes decision rules. Indeed according to (28), among the decision rules $P_{Z|XY}$ constrained by (29) one seeks the one minimizing the worst case risk. Notice, however by linearity of the risk function in Q_Y taking a prior on the set of Q_Y's is equivalent to choosing a prior concentrated at a single point (the average). Hence the left-hand side of (31) is just the worst-case Bayes risk.

When the space B of values of Y is finite then the set of distributions Q_Y is compact. Thus, computation of the minimax tradeoff $\tilde\beta_\alpha$ is facilitated by the existence of the least favorable prior Q_Y, cf. [12, Section 3.8]. To satisfy the regularity conditions in the general case, we first show that just like in (26) it is sufficient to restrict attention to

$$Q_Y \ll \mu$$

in the right-hand side of (31). Indeed, for any $Q_Y \not\ll \mu$ by the Lebesgue decomposition there exist probability measures Q_1, Q_2, a number $0 < \lambda \le 1$ and a pair of disjoint measurable sets S_1, S_2 such that $Q_1 \ll \mu$, $Q_2 \perp \mu$ and

$$Q_Y = \lambda Q_1 + (1-\lambda)Q_2$$

and $Q_1[S_1] = Q_2[S_2] = 1$. Thus, any test $P_{Z|XY}$ can be improved by restricting to S_1:

$$P_{Z|XY}(1|x,y) \quad \to \quad P_{Z|XY}(1|x,y)1\{y \in S_1\}$$

which does not change $P_{XY}[Z=1]$ (since $P_Y \ll \mu$) but possibly reduces $P_X Q_Y[Z=1]$. But then

$$P_X Q_Y[Z=1] = \lambda P_X Q_1[Z=1] < P_X Q_1[Z=1]\,,$$

and the measure Q_1 achieves a strictly larger β compared to Q_Y. Thus, to any $Q_Y \not\ll \mu$ there is a $Q_1 \ll \mu$ which is less favorable.

Next the space of measures $P_X \times Q_Y \ll P_X \times \mu$ can be identified with a convex subset of a complete metric space $L_1(\mathsf{A}\times\mathsf{B}, P_X\times\mu)$, while the set of $P_{Z|XY}$ with a convex subset of $L_\infty(\mathsf{A}\times\mathsf{B}, P_X\times\mu)$, corresponding to functions taking values in $[0,1]$. By σ-finiteness of $P_X\times\mu$ and a theorem of Banach-Alaoglu the set of $P_{Z|XY}$ is thus weak-$*$ compact; see also [12, Theorem A.5.1]. The result then follows from the completeness of the (closure of the) family of Bayes decision functions [13, Chapter 3] and [14, Section 5]. Indeed, as explained above $\beta_\alpha(P_{XY}, P_X Q_Y)$ corresponds to the Bayes test for a prior Q_Y, while by completeness such tests approach $\tilde\beta_\alpha(P_X, P_{Y|X})$ arbitrarily close. Alternatively, by weak compactness of $\{P_{Z|XY}\}$, (31) follows directly by the Fan's minimax theorem [15] applied to (28). ∎

B. Relation to non-signalling assisted codes

Since for any test $P_{Z|XY}$ we have (22) it makes sense to consider the following:

Definition 1: A randomized test $P_{Z|XY}: \mathsf{A}\times\mathsf{B}\to\{0,1\}$ is said to be P_X-balanced if the function

$$y \mapsto \int_{\mathsf{A}} P_{Z|XY}(1|x,y) P_X(dx)$$

is constant.

Remark: For deterministic tests $Z = 1\{(x,y)\in E\}$, P_X-balancedness means that the slices of the critical region $\{x : (x,y_0)\in E\}$ have equal P_X measure.

It can be seen that because of (22) every non-balanced test can be modified (by increasing some of the $P_{Z|XY}(1|x,y)$) to a P_X-balanced one without changing the

$$\sup_{Q_Y} P_X Q_Y[Z=1] \tag{32}$$

and without decreasing $P_{XY}[Z=1]$. This proves:

Theorem 5: In the computation of $\tilde\beta_\alpha(P_X, P_{Y|X})$ one may restrict optimization to P_X-balanced tests only:

$$\tilde{\mathcal{R}}(P_X, P_{Y|X}) = \{(\alpha,\beta)\colon \exists P_X\text{-balanced } P_{Z|XY} \text{ s.t. (17)-(18) hold}\} \tag{33}$$

As explained in [6, Section III] (see equation (36) in particular), every P_X and a P_X-balanced $P_{Z|XY}$ can be converted into a so-called non-signalling assisted (NSA) code for the channel $P_{Y|X}$ with number of codewords equal to the reciprocal of (32) and the probability of successful decoding equal to $P_{XY}[Z=1]$. Thus, we see that the maximal number of codewords $M^*(\epsilon)$ in an NSA code decodable with (average) probability of error ϵ satisfies

$$M^*(\epsilon) \ge \left\lfloor \frac{1}{\tilde\beta_{1-\epsilon}(P_X, P_{Y|X})} \right\rfloor . \tag{34}$$

On the other hand, it is easy to show that the minimax converse (7) also applies to the NSA codes. Overall, taking supremum over all P_X in (34) and applying Theorem 4 we get

$$M^*(\epsilon) = \left\lfloor \frac{1}{\inf_{P_X}\sup_{Q_Y}\beta_{1-\epsilon}(P_{XY}, P_X\times Q_Y))} \right\rfloor .$$

For the case of finite A and B, this result was shown in [6] by indirect arguments relying on the duality in linear programming. Here, however, we see that NSA codes are simply equivalent to P_X-balanced composite tests, which by virtue of Theorem 4 are in turn equivalent to solving the original minimax converse (7).

III. Convexity and continuity properties of β_α

A. Convexity in P_X

Each of the regions $\mathcal{R}(P_{XY}, P_X Q_Y)$ is convex. However, the union of such regions need not be convex, unless there is a special relationship between the sets. In this section we show that the latter is indeed the case. The following is a key new ingredient of this paper:

Theorem 6: For every P_X let P_{XY} and Q_{XY} denote the joint distributions on $\mathsf{A}\times\mathsf{B}$ defined as:

$$P_{XY}(dx\,dy) \stackrel{\triangle}{=} P_X(dx) P_{Y|X}(dy|x) \tag{35}$$

$$Q_{XY}(dx\,dy) \stackrel{\triangle}{=} P_X(dx) Q_{Y|X}(dy|x)\,. \tag{36}$$

Then the function

$$(\alpha, P_X) \to \beta_\alpha(P_{XY}, Q_{XY}) \tag{37}$$

is convex.

Proof: Take a finite convex combination of points in the domain of the function:

$$(\alpha, P_X) = \sum_j \lambda_j \cdot (\alpha_j, P_j)\,,$$

with $\sum_j \lambda_j = 1$ and $\lambda_j > 0$. Let $P_{Z_j|XY}$ be the tests achieving the optimal value β_j for each j. Note that $P_j \ll P_X$ and thus there exist Radon-Nikodym derivatives $\frac{dP_j}{dP_X}$. Define a new test

$$P_{Z|XY}(1|x,y) = \sum_j P_{Z_j|XY}(1|x,y)\lambda_j \frac{dP_j}{dP_X}(x)\,. \tag{38}$$

Since

$$\sum_j \lambda_j \frac{dP_j}{dP_X}(x) = 1$$

for P_X-almost all x the value in the right-hand side of (38) is between 0 and 1 and hence the test $P_{Z|XY}$ is well defined. Notice that by the definition of $\frac{dP_j}{dP_X}$ we have in the notation (19)

$$\begin{aligned} P_{XY}[Z=1] &= \int_{\mathsf{A}}\int_{\mathsf{B}} P_{Z|XY}(1|x,y) P_{Y|X}(dy|x) P_X(dx) && (39)\\ &= \sum_j \lambda_j \int_{\mathsf{A}}\int_{\mathsf{B}} P_{Z_j|XY}(1|x,y) P_{Y|X}(dy|x) P_j(dx) && (40)\\ &= \sum_j \lambda_j \alpha_j && (41)\\ &= \alpha && (42) \end{aligned}$$

Similarly, replacing $P_{Y|X}$ with $Q_{Y|X}$ we obtain

$$Q_{XY}[Z=1] = \sum_j \lambda_j \beta_j \,.$$

Thus, we have shown

$$\beta_\alpha(P_{XY}, Q_{XY}) \le \sum_j \lambda_j \beta_j \,,$$

which establishes convexity of (37). ∎

From the general properties of convex functions we obtain the following:

Corollary 7: Let $\mathcal{Q}$ be a family of random transformations $Q_{Y|X} : \mathsf{A} \to \mathsf{B}$ and Π be a convex set of probability measures on A. Then

$$(\alpha, P_X) \mapsto \sup_{Q_{Y|X}\in\mathcal{Q}} \beta_\alpha(P_{XY}, Q_{XY}) \tag{43}$$

$$\alpha \mapsto \inf_{P_X\in\Pi} \sup_{Q_{Y|X}\in\mathcal{Q}} \beta_\alpha(P_{XY}, Q_{XY}) \tag{44}$$

are convex.

We can restate the results in terms of the unions and intersections of the regions $\mathcal{R}(P_{XY}, Q_{XY})$ as follows:

Theorem 8: Let Π be a convex set of probability measures on A. Let $Q_{Y|X} : \mathsf{A} \to \mathsf{B}$ be a random transformation. For every P_X let P_{XY} and Q_{XY} denote the joint distributions (35)-(36) on $\mathsf{A}\times\mathsf{B}$. Then the set

$$\bigcup_{P_X\in\Pi} \mathcal{R}(P_{XY}, Q_{XY}) \tag{45}$$

is convex. Moreover, for any family $\mathcal{Q}$ of random transformations $Q_{Y|X} : \mathsf{A} \to \mathsf{B}$ the set

$$\bigcup_{P_X\in\Pi} \bigcap_{Q_{Y|X}\in\mathcal{Q}} \mathcal{R}(P_{XY}, Q_{XY}) \tag{46}$$

is convex.

Proof: By the symmetry $(\alpha,\beta) \leftrightarrow (1-\alpha, 1-\beta)$ it is sufficient to prove convexity of the union of the upper-extended regions:

$$\mathcal{R}'(P_{XY}, Q_{XY}) \stackrel{\triangle}{=} \bigcup_{(\alpha,\beta)\in\mathcal{R}(P_{XY},Q_{XY})} \{(\alpha,\beta') : \beta' \ge \beta\} \,,$$

which is precisely the epigraph of $\alpha \mapsto \beta_\alpha(P_{XY}, Q_{XY})$, see [19]. Next notice that the set

$$\bigcup_{P_X\in\Pi} \mathcal{R}'(P_{XY}, Q_{XY})$$

is in fact a projection of the epigraph of the convex function (Theorem 6)

$$(\alpha, P_X) \mapsto \beta_\alpha(P_{XY}, Q_{XY})$$

defined on $[0,1]\times\Pi\times[0,1]$ onto the first and third coordinate. The projection being linear must preserve the convexity.

Convexity of (46) follows from the fact that the convexifying test (38) did not in fact depend on the kernel $Q_{Y|X}$. ∎

For the purpose of this paper the following application of Theorems 6 and 8 is important:

Theorem 9: The set

$$\bigcup_{P_X} \bigcap_{Q_Y} \mathcal{R}(P_{XY}, P_X Q_Y) \tag{47}$$

is convex. Consequently,

$$\alpha \mapsto \inf_{P_X} \tilde\beta_\alpha(P_X, P_{Y|X}) \tag{48}$$

is a convex function on $[0,1]$.

Proof: Convexity of (47) is established by (46). By Theorem 4 the function in (48) is a lower boundary of the closure of (47), which must be convex. ∎

B. Continuity in P_X: general input space

We next consider the continuity properties of $\beta_\alpha(P_{XY}, P_{XY})$ as a function of P_X.

Theorem 10: For any Q_Y and $\alpha \in [0,1]$ the functions

$$P_X \mapsto \beta_\alpha(P_{XY}, P_X Q_Y) \tag{49}$$

$$P_X \mapsto \tilde\beta_\alpha(P_X, P_{Y|X}) \tag{50}$$

are continuous in the topology of total variation.

Proof: If $Q_Y \not\ll \mu$ then we can replace Q_Y with the absolutely continuous part of the latter without affecting the $\beta_\alpha(P_{XY}, P_X Q_Y)$, thus turning Q_Y into a sub-probability measure. So we assume Q_Y is given by

$$Q_Y[E] = \int_E q(y)\mu(dy) \,, \qquad \forall E \subset \mathsf{B} \,,$$

for some $q \ge 0$ with $||q||_1 \le 1$ in $L_1(\mathsf{B}, \mu)$.

First we consider the case $\alpha = 1$. No matter what Q_Y is the optimal test $P_{Z|XY}$ achieving $\beta_1(P_{XY}, P_X Q_Y)$ is

$$P_{Z|XY}(1|x,y) = 1\{\rho(y|x) > 0\} \,.$$

Indeed, consider reducing the value of $P_{Z|XY}(1|x,y)$ on any $E \subset \mathsf{A}\times\mathsf{B}$ with $P_X \times \mu[E] > 0$. Then for some $\epsilon > 0$ we must have

$$(P_X\mu)[E \cap \{\rho(Y|X) > \epsilon\}] > 0 \,,$$

which in turn implies, cf. (14), that

$$P_{XY}[E\cap\{\rho(Y|X) > \epsilon\}] > \epsilon\cdot(P_X\mu)[E\cap\{\rho(Y|X) > \epsilon\}] > 0$$

and thus

$$P_{XY}[\{\rho(Y|X) > 0\} \setminus E] < 1 \,.$$

Thus, we have

$$\beta_1(P_{XY}, P_X Q_Y) = \mathbb{E}\left[g(X)\right] , \tag{51}$$

where

$$g(x) = \int_{\mathsf{B}} 1\{\rho(y|x) > 0\} q(y)\mu(dy) \,.$$

Since $0 \le g \le 1$, from (51) we obtain for

$$|\beta_1(P_{XY}, P_X Q_Y) - \beta_1(P'_{XY}, P'_X Q_Y)| \le ||P_X - P'_X|| \,, \tag{52}$$

where $||\cdot||$ denotes the total variation distance and P'_{XY} – the joint probability distribution on $\mathsf{A}\times\mathsf{B}$ defined as in (8) with $P_X(dx)$ replaced by $P'_X(dx)$. Thus continuity of β_1 follows from (52) and continuity of $\tilde\beta_1$ follows from Theorem 4 and the fact that (52) holds uniformly for all Q_Y.

Now fix $\alpha \in (0,1)$. Note that if $\alpha \in (\epsilon, 1-\epsilon)$ for some $\epsilon > 0$ then from the definition (5) it follows that

$$\begin{aligned} \beta_{\alpha-\epsilon}(P,Q) - \epsilon &\le \beta_\alpha(P',Q') && (53)\\ &\le \beta_{\alpha+\epsilon}(P,Q) + \epsilon && (54)\end{aligned}$$

for every P, P', Q, Q' with

$$||P - P'|| \le \epsilon, \qquad ||Q - Q'|| \le \epsilon\,.$$

Now, since

$$\begin{aligned} ||P_{XY} - P'_{XY}|| &= ||P_X - P'_X|| && (55)\\ ||P_X Q_Y - P'_X Q_Y|| &= ||P_X - P'_X|| && (56)\end{aligned}$$

we have from (53)-(54) and continuity of $\alpha \mapsto \beta_\alpha(P_{XY}, P_X Q_Y)$ that

$$\beta_\alpha(P'_{XY}, P'_X Q_Y) \to \beta_\alpha(P_{XY}, P_X Q_Y)$$

as $P'_X \to P_X$.

To prove continuity of $P_X \mapsto \tilde\beta_\alpha(P_X, P_{Y|X})$ we consider P_X and P'_X with

$$||P_X - P'_X|| \le \epsilon\,.$$

Then by taking supremum over Q_Y in the obvious inequality

$$P_X Q_Y[Z=1] - \epsilon \le P'_X Q_Y[Z=1] \le P_X Q_Y[Z=1] + \epsilon$$

we prove the analog of (53)-(54) for $\tilde\beta_\alpha$:

$$\begin{aligned} \tilde\beta_{\alpha-\epsilon}(P_X, P_{Y|X}) - \epsilon &\le \beta_\alpha(P'_X, P_{Y|X}) && (57)\\ &\le \tilde\beta_{\alpha+\epsilon}(P_X, P_{Y|X}) + \epsilon\,. && (58)\end{aligned}$$

The statement follows by the continuity of $\alpha \mapsto \tilde\beta_\alpha(P_X, P_{Y|X})$. ∎

Note that on a finite-dimensional simplex there is only one topology that is compatible with the linear structure. Thus no matter how weak we choose the topology on the space of probability measures, we have:

Corollary 11: On every finite-dimensional simplex of probability distributions on A the functions (49) and (50) are continuous (in the trace of any topology compatible with the linear structure).

Remark: Or, equivalently, the map

$$(\lambda, \mu) \mapsto \beta_\alpha(\lambda P + (1-\lambda)P', \mu Q + (1-\mu)Q')$$

is continuous on $[0,1] \times [0,1]$.

Note that every convex and locally upper-bounded function is continuous on the interior of its domain. Thus, since $0 \le \beta_\alpha(P_{XY}, Q_{XY}) \le 1$ one may naturally wonder whether it is possible to show continuity of β_α from the convexity. It turns out this approach will not work for the subtle reason that the interior of $\mathcal{M}_1(\mathsf{A})$ is empty whenever A is infinite. In fact, in the vector space $\mathcal{M}(\mathsf{A})$ even the algebraic interior of a larger $\mathcal{M}_+(\mathsf{A})$ is empty. To see this, consider any measure ν. If ν is purely atomic with finitely many atoms, then since $|\mathsf{A}| = \infty$ there is a singleton $\{x_0\}$ and a δ-measure μ on it such that $\nu - \lambda\mu \notin \mathcal{M}_+$ for any $\lambda > 0$. Otherwise, in the space $L_1(\mathsf{A}, \nu)$ there exists an unbounded integrable function f, e.g. [18, Theorem 2.3.19], and hence setting

$$d\mu = f \cdot d\nu$$

we again conclude $\nu - \lambda\mu \notin \mathcal{M}_+$ for any $\lambda > 0$. Thus unlike the finite-dimensional case, every positive (in particular, probability) measure is a boundary point in any topology on the space of measures. That is why it is not generally possible to derive continuity on $\mathcal{M}_1$ by a simple convexity and local boundedness argument, and we had to give an explicit argument for Theorem 10. Furthermore, in the next section we show an example of the weak-discontinuity in β_α.

C. Continuity in P_X: topological input space

Our next goal will be to extend continuity of β_α on $\mathcal{M}_1(\mathsf{A})$ to weaker topologies. One possible choice would be to investigate the topology of pointwise convergence on all measurable sets, known as strong topology or τ-topology, cf. [20]. In this topology $P_n \to P$ if

$$P_n[E] \to P[E] \tag{59}$$

for any measurable set $E \subset \mathsf{A}$. The advantage of this definition is that it does not put any topological assumptions on the input space A itself. There are, however, several disadvantages. First, requirement (59) although much weaker than $||P_n - P|| \to 0$ is still very strong. For example, the sequence $\mathcal{N}(0, 1/n)$ of shrinking Gaussians does not converge to δ_0, a Dirac-delta at zero. The second problem is that typically the majority of τ-open sets does not belong to the σ-algebra $\mathcal{F}$ generated by

$$B_{E,I} = \{P_X : P_X[E] \in I\}\,, \tag{60}$$

where E is a measurable subset of A and I – an open subset of $[0,1]$[3]. The importance of $\mathcal{F}$ is that then a measurable map $P_{X|W} : \mathsf{W} \to \mathcal{M}_1(\mathsf{A})$ is precisely equivalent to defining a random transformation $P_{X|W} : \mathsf{W} \to \mathsf{A}$. Thus since $\tau \not\subseteq \mathcal{F}$ we cannot even guarantee that a τ-continuous function $F : \mathcal{M}(\mathsf{A}) \to \mathbb{R}$ induces a measurable map

$$w \mapsto F(P_{X|W=w}) \tag{61}$$

on W.

To resolve these problems we consider a much weaker notion of convergence, whose definition requires that the input space A itself be topological. The weak (or, more properly, weak-$*$) topology on $\mathcal{M}_1(\mathsf{A})$ is defined as the weakest topology under which the maps

$$P_X \mapsto \int_{\mathsf{A}} f(x) P_X(dx)\,,$$

are continuous for any continuous bounded f. In the case when A is Polish, the Borel σ-algebra of this topology coincides with σ-algebra $\mathcal{F}$ and $\mathcal{N}(0, 1/n) \to \delta_0$.

Is (49) a continuous function in the weak topology? The answer is negative:

Example (weak-discontinuity of β_α). Let $\mathsf{A} = \mathbb{R}$, B – arbitrary space with three probability distributions $P_0 \ne P_1$ and Q_Y on it. Then, consider

$$P_{Y|X}[\cdot|x] = P_0[\cdot]1\{x=0\} + P_1[\cdot]1\{x \ne 0\}\,.$$

[3] A simple argument shows that in the case when A is Polish, the τ-topology has cardinality at least $2^{\mathbb{R}}$ while $|\mathcal{F}| = |\mathbb{R}|$.

Let P_{X_n} be a uniform distribution on $[-1/n, 1/n]$. Then, clearly $P_{X_n} \to P_X$, where $P_X = \delta_0$ – a Dirac delta at 0. Thus, we have for any $\alpha \in [0,1]$:

$$\beta_\alpha(P_{X_nY}, P_{X_n}Q_Y) = \beta_\alpha(P_1, Q_Y) \not\to \beta_\alpha(P_{XY}, P_XQ_Y) = \beta_\alpha(P_0, Q_Y) \quad (62)$$

This example demonstrates, of course, that in order for β_α to be weakly continuous, we need to put some continuity requirements on the kernel $P_{Y|X}$ itself. This is done in the following:

Theorem 12: For every P_X let P_{XY} and Q_{XY} denote the joint distributions (35)-(36) on $\mathsf{A} \times \mathsf{B}$. Denote by $f'(y|x)$ the μ-density of the absolutely continuous component of $Q_{Y|X=x}$ in the Lebesgue decomposition of the latter. Assume that for any $\gamma \geq 0$ the function

$$x \mapsto \ell(x, \gamma) = \int_{\mathsf{B}} |\gamma\rho(y|x) - f'(y|x)|^+ \mu(dy) \quad (63)$$

is continuous. Then in the weak topology on $\mathcal{M}_1(\mathsf{A})$ the function

$$P_X \mapsto \beta_\alpha(P_{XY}, Q_{XY})$$

is continuous for $\alpha \in [0,1)$ and lower semicontinuous for $\alpha = 1$.

Proof: Denote the Fenchel-Legendre conjugate of β_α as

$$\beta^*_\gamma(P,Q) \stackrel{\triangle}{=} \sup_{0\leq\alpha\leq 1} \gamma\alpha - \beta_\alpha(P,Q)\,. \quad (64)$$

By the general Fenchel-Legendre duality and continuity-convexity of $\alpha \mapsto \beta_\alpha$ we have

$$\beta_\alpha(P,Q) = \sup_{\gamma\geq 0} \gamma\alpha - \beta^*_\gamma(P,Q)\,. \quad (65)$$

From the definition (5) we derive, as usual replacing Q with $Q \ll P$ if necessary,

$$\beta^*_\gamma(P,Q) = \int_{\mathsf{W}} \left|\gamma - \frac{dQ}{dP}\right|^+ P(dw)\,.$$

Thus, in our context we get

$$\beta^*_\gamma(P_{XY}, Q_{XY}) = \int_{\mathsf{A}} \ell(x,\gamma) P_X(dx)\,, \quad (66)$$

which is weakly continuous in P_X by assumption on $\ell(x,\gamma)$. Then lower-semicontinuity of (12) follows by characterization (65).

To show continuity for $\alpha < 1$ denote by $\mathcal{G}$ the space of all maps $g : [0,1) \to [0,1]$ corresponding to $\alpha \mapsto \beta_\alpha$ for some (P,Q):

$$\mathcal{G} \stackrel{\triangle}{=} \{f : \exists (P,Q) : f(\alpha) = \beta_\alpha(P,Q)\}\,,$$

and by $\mathcal{G}^*$ the space of all maps $g^* : \mathbb{R}^+ \to \mathbb{R}^+$ corresponding to $\gamma \mapsto \beta^*_\gamma$ for some (P,Q):

$$\mathcal{G}^* \stackrel{\triangle}{=} \{f^* : \exists (P,Q) : f^*(\gamma) = \beta^*_\gamma(P,Q)\}\,.$$

Endow both $\mathcal{G}$ and $\mathcal{G}^*$ with the topologies of pointwise convergence. Then, we can write map (12) as composition:

$$\mathcal{M}_1(\mathsf{A}) \xrightarrow{\beta^*_\gamma} \mathcal{G}^* \xrightarrow{T} \mathcal{G}\,,$$

where the first map is $P_X \mapsto \beta^*_\gamma(P_{XY}, Q_{XY})$ and the second is given by (65). Weak continuity of the first map follows from (66) and continuity of the second map from the following: ∎

Lemma 13: Operator $T : \mathcal{G}^* \to \mathcal{G}$ defined by (65) is continuous.

Proof: Consider $\alpha < 1$ and let $f = T(f^*)$, then

$$f(\alpha) \stackrel{\triangle}{=} \sup_{\gamma\geq 0} \gamma\alpha - f^*(\gamma) \quad (67)$$
$$= \max_{\gamma\geq 0} \gamma\alpha - f^*(\gamma) \quad (68)$$
$$= \max_{0\leq\gamma\leq\frac{1}{1-\alpha}} \gamma\alpha - f^*(\gamma)\,, \quad (69)$$

where (68) follows from the fact that the supremum must be achieved by any γ which defines a line touching the graph of f at α (i.e., γ is a subgradient of f at α), and (69) is because the slope γ cannot exceed $\frac{1-f(\alpha)}{1-\alpha}$ for otherwise $f(1-) \stackrel{\triangle}{=} \lim_{\alpha\nearrow 1} f(\alpha) > 1$.

On the other hand every $f^* \in \mathcal{G}^*$ is a convex conjugate of some $f \in \mathcal{G}$. Thus if we take $\gamma > \gamma_1$ and let α_* be the maximizer in the definition

$$f^*(\gamma) = \max_\alpha \gamma\alpha - f(\alpha)\,,$$

then we have

$$f^*(\gamma) - f^*(\gamma_1) \leq \gamma\alpha_* - f(\alpha_*) - (\gamma_1\alpha_* - f(\alpha_*)) \quad (70)$$
$$= (\gamma - \gamma_1)\alpha_* \quad (71)$$
$$\leq \gamma - \gamma_1\,, \quad (72)$$

where (70) is by taking a suboptimal $\alpha = \alpha_*$ for $f^*(\gamma_1)$ and (72) is because $0 \leq \alpha_* \leq 1$. Thus, every function in $\mathcal{G}^*$ is Lipschitz with constant 1. Moreover, since $f^*(0) = 0$, we also have

$$f^*(\gamma) \leq \gamma\,, \forall \gamma \geq 0\,.$$

Then by Arzela-Ascoli theorem, the pointwise convergence in $\mathcal{G}^*$ coincides with the topology of uniform convergence on compacts. By representation (69) the operator $T : \mathcal{G}^* \to \mathcal{G}$ is continuous in the latter. ∎

Finally, before closing this section we demonstrate that in the conditions of Theorem 12 there indeed can be a discontinuity at $\alpha = 1$.

Example (discontinuity at $\alpha = 1$). Let $\mathsf{A} = \mathsf{B} = \mathbb{R}$ and let the random transformation $P_{Y|X}$ be defined via

$$Y = XW\,,$$

where $W \sim \mathcal{N}(0,1)$ is standard Gaussian. Let $Q_Y = \mathcal{N}(0,1)$ and

$$\mu(dy) = \delta_0(dy) + dy\,,$$

where dy stands for a Lebesgue measure. Conditions of Theorem 12 are satisfied since the function

$$\ell(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} \int_{\mathsf{B}} \left|\frac{\gamma}{|x|} e^{-\frac{y^2}{2x^2}} - e^{-\frac{y^2}{2}}\right|^+ dy\,, & x \neq 0 \\ \gamma\,, & x = 0 \end{cases}$$

is continuous, which is verified by applying dominated convergence for $x_n \to x \neq 0$ and an explicit calculation

for $x_n \to 0$. On the other hand, consider the sequence $P_{X_n} = \mathcal{N}(0, 1/n) \to P_X = \delta_0$. We have

$$P_{X_n Y} \sim P_{X_n} Q_Y \sim \mathrm{Leb}(\mathbb{R}^2)\,,$$

where Leb denotes a Lebesgue measure. Thus

$$\beta_1(P_{X_n Y}, P_{X_n} Q_Y) = 1\,,$$

whereas

$$\beta_1(P_{XY}, P_X Q_Y) = 0\,,$$

achieved by a simple test $\{Y = 0\}$.

IV. Maximin converse

In this section we demonstrate that computation of the inner optimization in the maximin version (10) can be significantly simplified. By Theorem 8 we know that $\bigcup_{P_X} \mathcal{R}(P_{XY}, P_X Q_Y)$ is a convex set. It turns out that its extremal points correspond to the extremal measures on A:

Theorem 14: The union of $\mathcal{R}(P_{XY}, P_X Q_Y)$ taken over all distributions on A equals the convex hull, $\mathrm{co}(\cdot)$, of the union over all single-point measures:

$$\bigcup_{P_X} \mathcal{R}(P_{XY}, P_X Q_Y) = \mathrm{co}\left(\bigcup_{x \in \mathsf{A}} \mathcal{R}(P_{Y|X=x}, Q_Y)\right)\,. \quad (73)$$

Consequently,

$$\inf_{P_X} \beta_\alpha(P_{XY}, P_X Q_Y) = (\alpha \mapsto \inf_{x \in \mathsf{A}} \beta_\alpha(P_{Y|X=x}, Q_Y)^{**}\,, \quad (74)$$

where $(\cdot)^{**}$ denotes the operation of taking a convex envelope of a function (double Fenchel-Legendre conjugation).

Proof: First, notice that (74) follows from (73) since the functions appearing on both sides of (74) are the lower boundaries of the closures of the corresponding sets in (73). Next, we show

$$\bigcup_{P_X} \mathcal{R}(P_{XY}, P_X Q_Y) \subseteq \mathrm{co}\left(\bigcup_{x \in \mathsf{A}} \mathcal{R}(P_{Y|X=x}, Q_Y)\right)\,. \quad (75)$$

Indeed, consider any test $P_{Z|XY}$ and distribution P_X. Define

$$\alpha(x) \triangleq \int_{\mathsf{B}} P_{Z|XY}(1|x,y) dP_{Y|X=x}(y)\,, \quad (76)$$

$$\beta(x) \triangleq \int_{\mathsf{B}} P_{Z|XY}(1|x,y) dQ_Y(y)\,. \quad (77)$$

Clearly we have

$$(\alpha(x), \beta(x)) \in \mathcal{R}(P_{Y|X=x}, Q_Y)\,, \quad (78)$$

by the definition of $\mathcal{R}(P_{Y|X=x}, Q_Y)$. Averaging (78) over P_X we prove (75).

Conversely, consider any point

$$(\alpha, \beta) \in \mathrm{co}\left(\bigcup_{x \in \mathsf{A}} \mathcal{R}(P_{Y|X=x}, Q_Y)\right)\,.$$

By Caratheodory's theorem there exist $x_i \in \mathsf{A}, \lambda_i \in [0,1]$ and tests $P_{Z_i|Y}$ for each $i = 1, 2, 3$ such that $\sum_i \lambda_i = 1$ and

$$\sum_{i=1}^3 \lambda_i P[Z_i = 1 | X = x_i] = \alpha\,, \quad (79)$$

$$\sum_{i=1}^3 \lambda_i Q[Z_i = 1] = \beta\,, \quad (80)$$

where we adopted the notation (19)-(20). Thus the test

$$P_{Z|XY}(1|x,y) = \begin{cases} P_{Z_i|Y}(1|y)\,, & x = x_i, i = 1, 2 \text{ or } 3\,, \\ 0\,, & \text{otherwise.} \end{cases}$$

proves that (α, β) belongs to $\mathcal{R}(P_{XY}, P_X Q_Y)$ with

$$P_X[\cdot] = \sum_{i=1}^3 \lambda_i 1\{x_i \in \cdot\}\,.$$

■

V. Saddle point

The function $\beta_\alpha(P_{XY}, P_X Q_Y)$ is clearly concave in Q_Y and was shown to be convex in P_X by Theorem 6. Thus, it is natural to expect that the sup and inf in (9)-(10) are interchangeable. In this section we prove this under various assumptions.

A. Compact A

If the spaces A and B are finite then the infima and suprema in (9)-(10) are achievable and we have by the minimax theorem and continuity of β_α (Corollary 11):

$$\min_{P_X} \max_{Q_Y} \beta_\alpha(P_{XY}, P_X Q_Y) = \max_{Q_Y} \min_{P_X} \beta_\alpha(P_{XY}, P_X Q_Y)\,, \quad (81)$$

i.e. the function $(P_X, Q_Y) \mapsto \beta_\alpha(P_{XY}, P_X Q_Y)$ has *a saddle point* (P_X^*, Q_Y^*) found by solving the outer optimizations in (81).

We next extend this result to a slightly more general setting:

Theorem 15: Let A be compact and the random transformation $P_{Y|X}$ satisfy conditions of Theorem 12 for any $Q_Y \ll \mu$. Then for any $\alpha \in [0,1]$ we have

$$\min_{P_X} \sup_{Q_Y} \beta_\alpha(P_{XY}, P_X Q_Y) = \sup_{Q_Y} \min_{P_X} \beta_\alpha(P_{XY}, P_X Q_Y)\,. \quad (82)$$

Proof: As shown in the proof of Theorem 4 we may restrict to $Q_Y \ll \mu$ on both sides of (82). Since μ is σ-finite, there is $Q_Y \sim \mu$ and hence for $\alpha = 1$ both sides of (82) are equal to 1. For $\alpha < 1$ the result follows by Fan's minimax theorem [15] whose conditions are satisfied by concavity in Q_Y (obvious), convexity in P_X (Theorem 6) and continuity in P_X (Theorem 12). ■

Conditions of Theorem 15 may be verified with the help of the following:

Proposition 16: Let A be a first-countable topological space and a random transformation $P_{Y|X}$ be such that Radon-Nikodym derivatives $\rho(\cdot|x)$ in (14) satisfy:

1) $x \mapsto \rho(y|x)$ is continuous for μ-almost all y and

2) for every x there is a neighborhood U of x and μ-integrable function g such that

$$\rho(y|x) \le g(y) \qquad \forall x \in U, y \in \mathsf{B}\,. \tag{83}$$

Then for any measurable function $q : \mathsf{B} \to \mathbb{R}$ the map

$$x \mapsto \ell(x,q) \triangleq \int_{\mathsf{B}} |\rho(y|x) - q(y)|^+ \mu(dy) \tag{84}$$

is continuous.

Proof: To show that (84) is continuous simply apply the dominated convergence theorem in the neighborhood U majorizing $|\rho(y|x) - q(y)|^+$ by $g(y)$ via (83). ■

B. Non-compact A

Next, we replace the condition of compactness on A in Theorem 15 with local compactness (at the expense of additional assumptions on $P_{Y|X}$). Recall that a function f on a Hausdorff topological space A is said to converge to a at infinity if for every ϵ there is a compact $K_\epsilon \subseteq \mathsf{A}$ such that

$$\sup_{x \notin K_\epsilon} |f(x) - a| < \epsilon\,.$$

Definition 2: A random transformation $P_{Y|X} : \mathsf{A} \to \mathsf{B}$ satisfies the regularity assumptions if

1) A is a second-countable locally compact Hausdorff topological space;
2) for every $q \in L_1(\mathsf{B}, \mu)$ the map $\ell(x,q)$, see (84), is continuous in x and converges to 1 at infinity.

Topological conditions on A are satisfied for any open subset of a compact Polish space. Continuity of (84) can be verified via Proposition 16. Regarding the convergence at infinity the following is a simple criterion:

Proposition 17: If there exist sequences of compact sets $K_n \subseteq \mathsf{A}$ and monotonically increasing measurable sets $B_n \nearrow \mathsf{B}$ such that

$$\sup_{x \notin K_n} P_{Y|X}(B_n|x) \to 0 \qquad n \to \infty\,. \tag{85}$$

then $\ell(x,q)$, see (84), converges to 1 as $x \to \infty$ for any $q \in L_1(\mathsf{B}, \mu)$.

Example: If $\mathsf{A} = \mathsf{B} = \mathbb{R}^d$, μ is Lebesgue and $\rho(y|x) = (2\pi)^{-d/2} e^{-||y-x||^2/2}$ we can take $K_n = \{x : ||x|| \le 2n\}$, $B_n = \{y : ||y|| \le n\}$.

Proof: Consider the chain:

$$\begin{aligned}
&\left| \int_{\mathsf{B}} \min\{\rho(y|x), q(y)\} \mu(dy) \right| \\
&\le \int_{B_n} \rho(y|x)\mu(dy) + \int_{B_n^c} |q(y)|\mu(dy) && (86)\\
&= P_{Y|X}(B_n|x) + \int_{B_n^c} |q(y)|\mu(dy) && (87)
\end{aligned}$$

and thus

$$\begin{aligned}
&\sup_{x \notin K_n} |\ell(x,q) - 1| \\
&= \sup_{x \notin K_n} \left| \int_{\mathsf{B}} \mu(dy) \min\{\rho(y|x), q(y)\} \right| && (88)\\
&= \sup_{x \notin K_n} P_{Y|X}(B_n|x) + \int_{B_n^c} |q(y)| d\mu(y)\,, && (89)
\end{aligned}$$

which converges to zero by (85) and $B_n \nearrow \mathsf{B}$ as $n \to \infty$. ■

Theorem 18: For random transformation $P_{Y|X}$ satisfying Definition 2 we have for all $0 \le \alpha \le 1$:

$$\inf_{P_X} \sup_{Q_Y} \beta_\alpha(P_{XY}, P_X Q_Y) = \sup_{Q_Y} \inf_{P_X} \beta_\alpha(P_{XY}, P_X Q_Y)\,. \tag{90}$$

Consequently,

$$\left[\bigcup_{P_X} \bigcap_{Q_Y} \mathcal{R}(P_{XY}, P_X Q_Y) \right] = \left[\bigcap_{Q_Y} \bigcup_{P_X} \mathcal{R}(P_{XY}, P_X Q_Y) \right] \tag{91}$$

where $[\cdot]$ denotes the closure.

Proof: Denote

$$\begin{aligned}
b_1(\alpha) &= \inf_{P_X} \sup_{Q_Y} \beta_\alpha(P_{XY}, P_X Q_Y)\,, && (92)\\
b_2(\alpha) &= \sup_{Q_Y} \inf_{P_X} \beta_\alpha(P_{XY}, P_X Q_Y)\,. && (93)
\end{aligned}$$

By Theorems 9 and 14 both functions are convex, non-decreasing on $[0,1]$. Thus, it is enough to show that their convex conjugates match. Since clearly $b_1(\alpha) \ge b_2(\alpha)$ it is enough to show for every $\gamma > 0$:

$$\max_{0 \le \alpha \le 1} \alpha - \gamma b_1(\alpha) \ge \max_{0 \le \alpha \le 1} \alpha - \gamma b_2(\alpha) \tag{94}$$

Consider the left-hand side first:

$$\begin{aligned}
&\max_{0 \le \alpha \le 1} \alpha - \gamma b_1(\alpha) \\
&= \sup_{P_X} \max_{0 \le \alpha \le 1} \alpha - \gamma \sup_{Q_Y} \beta_\alpha(P_{XY}, P_X Q_Y) && (95)\\
&= \sup_{P_X} \max_{0 \le \alpha \le 1} \alpha - \gamma \tilde{\beta}_\alpha(P_X, P_{Y|X}) && (96)\\
&= \sup_{P_X} \max_{0 \le \alpha \le 1} \max_{P_{Z|XY} : P[Z=1] \ge \alpha} \alpha - \gamma \sup_{Q_Y} P_X Q_Y[Z=1] && (97)\\
&= \sup_{P_X} \max_{P_{Z|XY}} P_{XY}[Z=1] - \gamma \sup_{Q_Y} P_X Q_Y[Z=1] && (98)\\
&= \sup_{P_X} \max_{P_{Z|XY}} \inf_{Q_Y} P_{XY}[Z=1] - \gamma P_X Q_Y[Z=1] && (99)\\
&= \sup_{P_X} \inf_{Q_Y} \max_{P_{Z|XY}} P_{XY}[Z=1] - \gamma P_X Q_Y[Z=1] && (100)\\
&= \sup_{P_X} \inf_{q} \max_{P_{Z|XY}} \Big[P_{XY}[Z=1] \\
&\quad - \gamma \int_{\mathsf{A}} P_X(dx) \int_{\mathsf{B}} q(y) P_{Z|XY}(1|xy) \mu(dy) \Big] && (101)\\
&= \sup_{P_X} \inf_{q} \int_{\mathsf{A}} \ell(x, \gamma q) P_X(dx)\,, && (102)
\end{aligned}$$

where (95) is by definition, (96) is by (31), (97) is by (28), (98) is by merging the two optimizations, (100) is by a minimax theorem of Ky Fan [15], (101) is by taking $dQ_Y = q(y)d\mu$ with

$$q \in L_1(\mathsf{B}, \mu) : q \ge 0, ||q|| = 1\,,$$

which is done without loss of generality as argued in the proof of Theorem 4; and (102) is by solving a simple optimization over $P_{Z|XY}$ and the definition of $\ell(x,q)$ in (84).

For the right-hand side of (94) we have

$$\begin{aligned}
&\max_{0\le\alpha\le 1} \alpha - \gamma b_2(\alpha) \\
&= \max_{0\le\alpha\le 1} \inf_{Q_Y} \sup_{P_X} \alpha - \gamma\beta_\alpha(P_{XY}, P_X Q_Y) && (103)\\
&\le \inf_{Q_Y} \max_{0\le\alpha\le 1} \sup_{P_X} \alpha - \gamma\beta_\alpha(P_{XY}, P_X Q_Y) && (104)\\
&= \inf_{Q_Y} \sup_{P_X} \max_{0\le\alpha\le 1} \alpha - \gamma\beta_\alpha(P_{XY}, P_X Q_Y) && (105)\\
&= \inf_{q} \sup_{P_X} \int_{\mathsf{A}} \ell(x,\gamma q) P_X(dx)\,, && (106)
\end{aligned}$$

where (103) is by definition of b_2 in (93), (104) is by the general interchanging of max and inf, (105)-(106) is by the same argument as in (100)-(102).

Thus, (94) will follow once we show that the sup and inf in (102) and (106) are interchangeable. To that end, we employ the regularity conditions, which also guarantee that $\sup_{P_X}$ is in fact a max, and a minimax theorem of Fan.

Denote the Banach space of all regular σ-additive measures on A by $\mathcal{M}_{reg}(\mathsf{A})$, cf. [21, Definition III.5.10], and by $C_0(\mathsf{A})$ the space of all continuous functions tending to 0 at infinity. By [21, Theorem IV.6.3] and a simple one-point (Alexandroff) compactification argument, $\mathcal{M}_{reg}$ is the continuous dual of $C_0(\mathsf{A})$. The weakest topology on $\mathcal{M}_{reg}$ under which all elements of $C_0(\mathsf{A})$ are continuous is called weak-$*$ topology (not to be confused with the topology of weak convergence of measures defined by $C(\mathsf{A})$). Topological assumptions on A imply it is a normal space and thus (Urysohn lemma) any finite measure on A is regular. Consequently $\mathcal{M}_1$ is a convex subset of $\mathcal{M}_{reg}$, which is closed in the topology of total variation but in general is not weak-$*$ closed. The weak-$*$ closure of $\mathcal{M}_1$ is the set of all positive measures not exceeding 1 in total variation:

$$\mathcal{M}_{\le 1}^{+} \stackrel{\triangle}{=} \{\lambda : \lambda[\mathsf{A}] \le 1, \lambda \ge 0\}\,,$$

which is weak-$*$ compact by Banach-Alaoglu theorem.

We now argue that the extension of the domain from $\mathcal{M}_1$ to $\mathcal{M}_{\le 1}^{+}$ in (102) and (106) is immaterial. Indeed, take any $\nu \in \mathcal{M}_{\le 1}^{+}$ with $\nu[\mathsf{A}] = a \in (0,1]$. Then by non-negativity of $\ell(x,q)$ we have

$$\int_{\mathsf{A}} \ell(x,\gamma q)\nu(dx) \le \int_{\mathsf{A}} \ell(x,\gamma q)\tilde{\nu}(dx)\,,$$

where $\tilde{\nu} = \frac{1}{a}\nu$. Hence to every choice in $\mathcal{M}_{\le 1}^{+}$ there exists a better or equal choice in $\mathcal{M}_1$:

$$\sup_{P_X} \int_{\mathsf{A}} \ell(x,\gamma q) P_X(dx) = \max_{\nu\in\mathcal{M}_{\le 1}^{+}} \int_{\mathsf{A}} \ell(x,\gamma q)\nu(dx) \tag{107}$$

$$\sup_{P_X} \inf_{q} \int_{\mathsf{A}} \ell(x,\gamma q) P_X(dx) = \max_{\nu\in\mathcal{M}_{\le 1}^{+}} \inf_{q} \int_{\mathsf{A}} \ell(x,\gamma q)\nu(dx)\,. \tag{108}$$

Thus, by the minimax theorem of Ky Fan [15] we get

$$\inf_{q} \max_{\nu\in\mathcal{M}_{\le 1}^{+}} \int_{\mathsf{A}} \ell(x,\gamma q)\nu(dx) = \max_{\nu\in\mathcal{M}_{\le 1}^{+}} \inf_{q} \int_{\mathsf{A}} \ell(x,\gamma q)\nu(dx)\,,$$

completing the proof of (94).

Finally, (91) follows from (90) by the symmetry of the regions. ∎

VI. Computing saddle point

Computing the distributions (P_X, Q_Y) achieving the saddle point (81) is in general a hard problem. It can be significantly simplified if the random transformation possesses some symmetries. In this section we define such symmetries and demonstrate how they help in computing the value of the minimax problem.

A. General symmetry considerations

Definition 3: A pair of measurable maps $f = (f_i, f_o)$ is a symmetry of $P_{Y|X}$ if

$$P_{Y|X}(f_o^{-1}(E)|f_i(x)) = P_{Y|X}(E|x)\,,$$

for all measurable $E \subset \mathsf{B}$ and $x \in \mathsf{A}$. Two symmetries f and g can be composed to produce another symmetry as

$$(g_i, g_o) \circ (f_i, f_o) \stackrel{\triangle}{=} (g_i \circ f_i, f_o \circ g_o)\,. \tag{109}$$

A symmetry group G of $P_{Y|X}$ is any collection of invertible symmetries (automorphisms) closed under the group operation (109).

Note that both components of an automorphism $f = (f_i, f_o)$ are bimeasurable bijections, that is $f_i, f_i^{-1}, f_o, f_o^{-1}$ are all measurable and well-defined functions.

Naturally, every symmetry group G possesses a canonical left action on $\mathsf{A}\times\mathsf{B}$ defined as

$$g\cdot(x,y) \stackrel{\triangle}{=} (g_i(x), g_o^{-1}(y))\,. \tag{110}$$

Since the action on $\mathsf{A}\times\mathsf{B}$ splits into actions on A and B, we will abuse notation slightly and write

$$g\cdot(x,y) \stackrel{\triangle}{=} (g\,x, g\,y)\,.$$

For the cases of infinite A, B we need to impose certain additional regularity conditions:

Definition 4: A symmetry group G is called regular if it possesses a left-invariant Haar probability measure ν such that the group action (110)

$$G \times \mathsf{A} \times \mathsf{B} \to \mathsf{A} \times \mathsf{B}$$

is measurable.

Note that under the regularity assumption the action (110) also defines left-action of G on $\mathcal{M}_1(\mathsf{A})$ and $\mathcal{M}_1(\mathsf{B})$ according to

$$(gP_X)[E] \stackrel{\triangle}{=} P_X[g^{-1}E]\,, \tag{111}$$

$$(gQ_Y)[E] \stackrel{\triangle}{=} Q_Y[g^{-1}E]\,, \tag{112}$$

or, in words, if $X \sim P_X$ then $gX \sim gP_X$, and similarly for Y and gY. For every distribution P_X we define an averaged distribution $\bar{P}_X$ as

$$\bar{P}_X[E] \stackrel{\triangle}{=} \int_G P_X[g^{-1}E]\nu(dg)\,, \tag{113}$$

which is the distribution of random variable gX when $g \sim \nu$ and $X \sim P_X$. The measure $\bar{P}_X$ is G-invariant, in the sense that $g\bar{P}_X = \bar{P}_X$. Indeed, by left-invariance of ν we have for every bounded function f

$$\int_G f(g)\nu(dg) = \int_G f(hg)\nu(dg) \qquad \forall h \in G\,,$$

and therefore

$$\bar{P}_X[h^{-1}E] = \int_G P_X[(hg)^{-1}E]\nu(dg) = \bar{P}_X[E]\,.$$

Similarly one defines $\bar{Q}_Y$:

$$\bar{Q}_Y[E] \stackrel{\triangle}{=} \int_G Q_Y[g^{-1}E]\nu(dg)\,, \tag{114}$$

which is also G-invariant: $g\bar{Q}_Y = \bar{Q}_Y$.

The main property of the action of G may be rephrased as follows: For arbitrary $\phi : \mathsf{A}\times\mathsf{B}\to\mathbb{R}$ we have

$$\begin{aligned}&\int_{\mathsf{A}}\int_{\mathsf{B}} \phi(x,y)P_{Y|X}(dy|x)(gP_X)(dx)\\ &= \int_{\mathsf{A}}\int_{\mathsf{B}} \phi(gx,gy)P_{Y|X}(dy|x)P_X(dx)\,.\end{aligned} \tag{115}$$

In other words, if the pair (X,Y) is generated by taking $X\sim P_X$ and applying $P_{Y|X}$, then the pair (gX,gY) has marginal distribution gP_X but conditional kernel is still $P_{Y|X}$. For finite A,B this is equivalent to

$$P_{Y|X}(gy|gx) = P_{Y|X}(y|x)\,, \tag{116}$$

which may also be taken as the definition of the automorphism. In terms of the G-action on $\mathcal{M}_1(\mathsf{B})$ we may also say:

$$gP_{Y|X=x} = P_{Y|X=gx} \qquad \forall g\in G, x\in\mathsf{A}\,. \tag{117}$$

Proposition 19: Fix P_X, Q_Y and $g\in G$ and denote $P'_X = gP_X$, $Q'_Y = gQ_Y$. Then

$$\beta_\alpha(P'_{XY}, P'_XQ'_Y) = \beta_\alpha(P_{XY},P_XQ_Y)\,, \tag{118}$$

$$\tilde{\beta}_\alpha(P'_X, P_{Y|X}) = \tilde{\beta}_\alpha(P_X,P_{Y|X})\,, \tag{119}$$

$$\inf_{P_X}\beta_\alpha(P_{XY},P_XQ'_Y) = \inf_{P_X}\beta_\alpha(P_{XY},P_XQ_Y)\,. \tag{120}$$

Proof: All statements are proved by a straightforward application of (115). For example, to show (118) it is sufficient to verify

$$\beta_\alpha(P'_{XY},P'_XQ'_Y)\ge\beta_\alpha(P_{XY},P_XQ_Y)\,, \tag{121}$$

since the reverse inequality follows by applying the argument with $g\to g^{-1}$. Let $P_{Z'|XY}$ be the test achieving $\beta_\alpha(P'_{XY},P'_XQ'_Y)$. Then define

$$P_{Z|XY}(1|x,y) = P_{Z'|XY}(1|gx,gy)$$

and apply (115) to show

$$P_{XY}[Z=1] = P'_{XY}[Z'=1]\,.$$

On the other hand,

$$\begin{aligned}&P_XQ_Y[Z=1]\\ &= \int_{\mathsf{A}}\int_{\mathsf{B}} P_{Z'|XY}(1|gx,gy)Q_Y(dy)P_X(dx) &(122)\\ &= \int_{\mathsf{A}}\int_{\mathsf{B}} P_{Z'|XY}(1|x,y)(gQ_Y)(dy)(gP_X)(dx)\,, &(123)\end{aligned}$$

which follows by a standard change of variable formula and (111)-(112). (119) and (120) are shown similarly. ∎

The main result of this section is the following:

Theorem 20: Let G be a regular group of symmetries of $P_{Y|X}$. Then the infima and suprema in both (9) and (10) can be restricted to G-invariant distributions, namely:

$$\forall g\in G:\forall E\subset\mathsf{B}: Q_Y[g^{-1}E] = Q_Y[E]\,, \tag{124}$$

$$\forall g\in G:\forall E\subset\mathsf{A}: P_X[g^{-1}E] = P_X[E]\,. \tag{125}$$

Moreover, whenever P_X and Q_Y are such, the optimal test $P_{Z|XY}$ achieving $\beta_\alpha(P_{XY},P_XQ_Y)$ can be chosen to be constant on the orbits of G-action on $\mathsf{A}\times\mathsf{B}$. Similarly, whenever P_X is $G-$invariant, there exists an optimal P_X-balanced G-invariant test achieving $\tilde{\beta}_\alpha(P_X,P_{Y|X})$.

Remark: For example, in DMC G can be chosen to be the symmetric group, in which case the orbits on $\mathsf{A}\times\mathsf{B}$ are the joint types and the optimization problem becomes simpler, see [6, Section III.B].

Proof: The following claims are being made:

1) Outer optimization in (9):
$$\begin{aligned}&\inf_{P_X}\sup_{Q_Y}\beta_\alpha(P_{XY},P_XQ_Y)\\ &= \inf_{P_X-\text{sat. (125)}}\sup_{Q_Y}\beta_\alpha(P_{XY},P_XQ_Y)\end{aligned} \tag{126}$$
2) Inner optimization in (10) subject to P_X satisfying (125):
$$\sup_{Q_Y}\beta_\alpha(P_{XY},P_XQ_Y) = \sup_{Q_Y-\text{sat. (124)}}\beta_\alpha(P_{XY},P_XQ_Y) \tag{127}$$
3) Tests for G-invariant P_X and Q_Y:
$$\beta_\alpha(P_{XY},P_XQ_Y) = \inf_{P_{Z|XY}} P_XQ_Y[Z=1]\,, \tag{128}$$
where $P_{Z|XY}$ satisfies
$$P_{XY}[Z=1]\ge\alpha$$
and
$$P_{Z|XY}(1|x,y) = P_{Z|XY}(1|gx,gy)$$
for all $x\in\mathsf{A}, y\in\mathsf{B}, g\in G$.
4) A similar set of claims for (10).

A very simple method to show (126)-(128) would be the following. First notice that by (118) and Theorem 4 we have that the function

$$f(P_X) = \sup_{Q_Y}\beta_\alpha(P_{XY},P_XQ_Y) \tag{129}$$

is constant on the orbits of G. Therefore, by invoking convexity of β_α (Theorem 6) and applying the Jensen inequality we obtain:

$$f(P_X) = \int_G f(gP_X)\nu(dg)\le f(\bar{P}_X)\,,$$

where ν is the Haar measure on G, $\bar{P}_X$ is the distribution of gX when $g\sim\nu$ and $X\sim P_X$. Since obviously $\bar{P}_X$ is G-invariant (126) follows. Similarly, one shows (127), (128) and analogous claims for (10).

Unfortunately, the proofs as above (with exception of that for (128)) contain a subtle gap: it is not known whether f defined by (129) is measurable on $\mathcal{M}_1$. Notice that because of the remark (61), Theorem 10 does not help. Fortunately, it

is not hard to find an explicit proof for these claims without invoking Jensen's inequality.

For example, we show (126), which is equivalent to (Theorem 4)

$$\tilde{\beta}_\alpha(P_X) \geq \tilde{\beta}_\alpha(\bar{P}_X)\,, \tag{130}$$

where for the remainder of the proof we omit the second argument of $\tilde{\beta}_\alpha$. Indeed, assume to the contrary that there is $\epsilon > 0$ such that:

$$\tilde{\beta}_\alpha(P_X) < \tilde{\beta}_\alpha(\bar{P}_X) - \epsilon\,. \tag{131}$$

First, by Corollary 11 we have for some small ϵ_1:

$$\tilde{\beta}_\alpha((1-\epsilon_1)P_X + \epsilon_1\bar{P}_X) < \tilde{\beta}_\alpha(\bar{P}_X) - \epsilon/2\,.$$

Thus, perhaps by replacing P_X with $(1-\epsilon_1)P_X + \epsilon_1\bar{P}_X$ we may assume without loss of generality that $P_X \ll \bar{P}_X$. Denote

$$\psi(x) \triangleq \frac{dP_X}{d\bar{P}_X}(x)\,.$$

Next, we observe that

$$(gP_X)[E] = \int_E \psi(g^{-1}x)\bar{P}_X(dx)\,. \tag{132}$$

Thus, functions $\psi(g^{-1}x)$ are the $\bar{P}_X$-densities of gP_X. Therefore applying Fubini's theorem

$$\begin{aligned}\bar{P}_X[E] &= \int_G (gP_X)[E]\nu(dg) &(133)\\ &= \int_G\int_E \psi(g^{-1}x)\nu(dg)\bar{P}_X(dx) &(134)\end{aligned}$$

we conclude that

$$\int_G \psi(g^{-1}x)\nu(dg) = 1 \tag{135}$$

for $\bar{P}_X$-almost all x.

Next, consider a test $P_{Z|XY}$ achieving $\tilde{\beta}_\alpha(P_X)$ and let

$$P_{\bar{Z}|X,Y}(1|x,y) = \int_G \psi(g^{-1}x)P_{Z|XY}(1|g^{-1}x,g^{-1}y)\nu(dg)\,. \tag{136}$$

By (135) the right-hands side of (136) does not exceed 1 and therefore defines a valid probability kernel. We have then

$$\begin{aligned}&\bar{P}_{XY}[\bar{Z}=1]\\ &= \int_{\mathsf{A}}\int_{\mathsf{B}}\int_G P_{Z|XY}(1|g^{-1}x,g^{-1}y)\\ &\quad\cdot\psi(g^{-1}x)\nu(dg)P_{Y|X}(dy|x)\bar{P}_X(dx) &(137)\\ &= \int_G \nu(dg)\int_{\mathsf{A}}\int_{\mathsf{B}} P_{Z|XY}(1|g^{-1}x,g^{-1}y)\\ &\quad\cdot P_{Y|X}(dy|x)(gP_X)(dx) &(138)\\ &= \int_G \nu(dg)\int_{\mathsf{A}}\int_{\mathsf{B}} P_{Z|XY}(1|x,y)\nu(dg)P_{Y|X}(dy|x)P_X(dx) &(139)\\ &= \int_G \nu(dg)P_{XY}[Z=1] &(140)\\ &\geq \alpha\,, &(141)\end{aligned}$$

where in (137) we denote $\bar{P}_{XY} = \bar{P}_X P_{Y|X}$, (138) is by (132), (139) is by (115) with $\phi(x,y) = P_{Z|XY}(1|g^{-1}x,g^{-1}y)$, and (141) is by assumption on $P_{Z|XY}$.

On the other hand,

$$\begin{aligned}&\sup_{Q_Y}\bar{P}_X Q_Y[\bar{Z}=1]\\ &= \sup_y \int_{\mathsf{A}} P_{\bar{Z}|XY}(1|x,y)\bar{P}_X(dx) &(142)\\ &= \sup_y \int_{\mathsf{A}}\int_G P_{Z|XY}(1|g^{-1}x,g^{-1}y)\psi(g^{-1}x)\nu(dg)\bar{P}_X(dx) &(143)\\ &= \sup_y \int_G \nu(dg)\int_{\mathsf{A}} P_{Z|XY}(1|g^{-1}x,g^{-1}y)(gP_X)(dx) &(144)\\ &= \sup_y \int_G \nu(dg)\int_{\mathsf{A}} P_{Z|XY}(1|x,g^{-1}y)P_X(dx) &(145)\\ &\leq \int_G \nu(dg)\sup_y\int_{\mathsf{A}} P_{Z|XY}(1|x,g^{-1}y)P_X(dx) &(146)\\ &= \int_G \nu(dg)\sup_y\int_{\mathsf{A}} P_{Z|XY}(1|x,y)P_X(dx) &(147)\\ &= \int_G \nu(dg)\tilde{\beta}_\alpha(P_X) = \tilde{\beta}_\alpha(P_X)\,, &(148)\end{aligned}$$

where (142) is by (22), (143) is by (136), (144) is by (132), (145) is by a change of variable formula, (146) is possible since we show next that the function under the integration over G is measurable (in fact, constant), (147) follows since $g^{-1}:\mathsf{B}\to\mathsf{B}$ is a bijection and (148) is by the assumption that $P_{Z|XY}$ achieves $\tilde{\beta}_\alpha(P_X)$. Hence, (148) implies (130) and therefore (131) cannot hold.

The measurability assumptions in the proofs of (127) and the analogous claims for (10) can be worked around in a similar fashion. ■

B. Symmetric channels

As Theorem 20 shows, the larger the G-orbits in A (or B) are, the easier the solution of the saddle-point problem (81) becomes. The extreme cases deserve a special definition:

Definition 5: The random transformation $P_{Y|X}$ is called input-symmetric (output-symmetric) if there exists a regular group of symmetries G acting transitively on A (B).

Theorem 21: If the channel is input-symmetric (resp. output-symmetric), then the saddle-point in (81) is achieved by the uniform P_X (resp. Q_Y).

Proof: We will show that under the assumptions there is only one G-invariant distribution, which may be defined via (113) or (114) starting from an arbitrary P_X or Q_Y. Indeed, consider the case of input symmetry and assume there are two G-invariant input distributions P_1 and P_2. Let

$$P_0 = \frac{1}{2}P_1 + \frac{1}{2}P_2$$

and let $\psi_1 = \frac{dP_1}{dP_0}$, be the P_0-densities of P_1. The G-invariance of P_0, P_1 and P_2, equivalently, states that for any bounded f

$$\int_G f(x)P_j(dx) = \int_G f(gx)P_j(dx) \qquad j=0,1,2\,. \tag{149}$$

Applying (149) to P_1 and rewriting in terms of P_0 we get:

$$\int_G f(x)\psi_1(x)P_0(dx) = \int_G f(gx)\psi_1(x)P_0(dx) \tag{150}$$

$$= \int_G f(x)\psi_1(g^{-1}x)P_0(dx)\,, \tag{151}$$

where in (151) we applied G-invariance property (149) of P_0 for g^{-1}. Since (151) holds for all f we conclude

$$\psi_1(x) = \psi_1(g^{-1}x)$$

for P_0-almost all x and all $g \in G$. Since G acts transitively on A we conclude that ψ_1 is a constant, indeed a unity:

$$\psi_1(x) = 1\,,$$

and hence $P_1 = P_2 = P_0$. ∎

We mention relations of these definitions to other concepts of symmetry which have previously appeared in the literature. We restrict the following discussion to the case of finite A, B and thus $P_{Y|X}$ is just a $|\mathsf{A}| \times |\mathsf{B}|$ stochastic matrix, or a DMC:

- $P_{Y|X}$ is a *group-noise channel* if $\mathsf{A} = \mathsf{B}$ is a group and $P_{Y|X}$ acts by composing X with a noise variable Z:
$$Y = X \circ Z\,,$$
where $\circ$ is a group operation and Z is independent of X.
- $P_{Y|X}$ is called *Dobrushin-symmetric* if every row of $P_{Y|X}$ is a permutation of the first one and every column of $P_{Y|X}$ is a permutation of the first one; see [22].
- $P_{Y|X}$ is called *Gallager-symmetric* if the output alphabet B can be split into a disjoint union of sub-alphabets such that restricted to each sub-alphabet $P_{Y|X}$ has the Dobrushin property: every row (every column) is a permutation of the first row (column); see [23, Section 4.5].
- for convenience, say that the channel is *square* if $|\mathsf{A}| = |\mathsf{B}|$.

We demonstrate some of the relationship between these various notions of symmetry:

1) Note that it is an easy consequence of the definitions that any input-symmetric (resp. output-symmetric) channel's $P_{Y|X}$ has all rows (resp. columns) – permutations of the first row (resp. column). Hence,
$$\text{input-symmetric, output-symmetric} \Longrightarrow \text{Dobrushin} \tag{152}$$

2) Group-noise channels satisfy all other definitions of symmetry:
$$\text{group-noise} \Longrightarrow \text{square, input/output-symmetric} \tag{153}$$
$$\Longrightarrow \text{Dobrushin, Gallager} \tag{154}$$

3) Since Gallager symmetry implies all rows are permutations of the first one, while output symmetry implies the same statement for columns we have
$$\text{Gallager, output-symmetric} \Longrightarrow \text{Dobrushin}$$

4) Clearly, not every Dobrushin-symmetric channel is square. One may wonder, however, whether every square Dobrushin channel is a group-noise channel. This is not so. Indeed, according to [24] the latin squares that are Cayley tables are precisely the ones in which composition of two rows (as permutations) gives another row. An example of the latin square which is not a Cayley table is the following:
$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 1 & 3 \\ 3 & 1 & 2 & 5 & 4 \\ 4 & 3 & 5 & 2 & 1 \\ 5 & 4 & 1 & 3 & 2 \end{pmatrix}. \tag{155}$$
Thus, by multiplying this matrix by $\frac{1}{15}$ we obtain a counter-example:
$$\text{Dobrushin, square} \not\Longrightarrow \text{group-noise}$$
In fact, this channel is not even input-symmetric. Indeed, suppose there is $g \in G$ such that $g4 = 1$ (on A). Then, applying (116) with $x = 4$ we figure out that on B the action of g must be:
$$1 \mapsto 4, 2 \mapsto 3, 3 \mapsto 5, 4 \mapsto 2, 5 \mapsto 1\,.$$
But then we have
$$gP_{Y|X=1} = \begin{pmatrix} 5 & 4 & 2 & 1 & 3 \end{pmatrix} \cdot \frac{1}{15}\,,$$
which by a simple inspection does not match any of the rows in (155). Thus, (117) cannot hold for $x = 1$. We conclude:
$$\text{Dobrushin, square} \not\Longrightarrow \text{input-symmetric}$$
Similarly, if there were $g \in G$ such that $g2 = 1$ (on B), then on A it would act as
$$1 \mapsto 2, 2 \mapsto 5, 3 \mapsto 1, 4 \mapsto 3, 5 \mapsto 4\,,$$
which implies via (116) that $P_{Y|X}(g1|x)$ is not a column of (155). Thus:
$$\text{Dobrushin, square} \not\Longrightarrow \text{output-symmetric}$$

5) Clearly, not every input-symmetric channel is Dobrushin (e.g., BEC). One may even find a counter-example in the class of square channels:
$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 2 & 4 \\ 4 & 2 & 3 & 1 \\ 4 & 3 & 2 & 1 \end{pmatrix} \cdot \frac{1}{10} \tag{156}$$
This shows:
$$\text{input-symmetric, square} \not\Longrightarrow \text{Dobrushin}$$

6) Channel (156) also demonstrates:
$$\text{Gallager-symmetric, square} \not\Longrightarrow \text{Dobrushin}\,.$$

7) Example (156) naturally raises the question of whether every input-symmetric channel is Gallager symmetric. The answer is positive: by splitting B into the orbits of G we see that a subchannel $\mathsf{A} \to \{\text{orbit}\}$ is input and output symmetric. Thus by (152) we have:
$$\text{input-symmetric} \Longrightarrow \text{Gallager-symmetric} \tag{157}$$

8) As previous argument shows, input-symmetry is more restrictive than Gallager symmetry. It turns out, however, one may define a notion of a weakly input symmetric channel [2, Definition 4], which is close in spirit to the definition of input symmetry (in the sense of implying that all inputs have equivalent coding properties), while being also more general than Gallager's definition; see discussion in [10, Section 3.4.5].
9) Other definitions that have appeared in the literature may also be recast in terms of requirements on the action of G on A or B. For example, Hof et al [25] define the DMC to be symmetric if A is an abelian group and there is a set of bijections $T_x : \mathsf{B} \to \mathsf{B}, x \in \mathsf{A}$ such that

$$P_{Y|X}(T_{x_2-x_1}(y)|x_2) = P_{Y|X}(y|x_1)\,.$$

It is easy to see that, the element that acts by adding x_0 on A and by $T_{x_0}^{-1}$ on B forms a channel symmetry $(\cdot + x_0, T_{x_0}^{-1}(\cdot))$. This collection can be completed to form a group (under composition (109)) by adding elements σ that act trivially on A and permute B such that

$$P_{Y|X}(\sigma(y)|x) = P_{Y|X}(y|x) \forall x \in \mathsf{A}, y \in \mathsf{B}\,.$$

Thus, we see that symmetry of [25] is a special case of input symmetry, when the action of G is addition in the abelian group A.

A pictorial representation of these relationships between the notions of symmetry is given schematically on Fig. 2.

C. Binary symmetric channel (BSC)

Recall that the $BSC(n, \delta)$ of blocklength n and crossover probability δ has the binary input and output alphabets, $\mathsf{A} = \mathsf{B} = \mathbb{F}_2^n$, and transition probabilities

$$P_{Y^n|X^n}(y^n|x^n) = \delta^{|y^n - x^n|}(1-\delta)^{n-|y^n-x^n|}\,, \qquad (158)$$

where $|z^n|$ denotes the Hamming weight of the binary vector z^n. Consider the group $G = \mathbb{F}_2^n \rtimes S_n$ generated by symmetries of two kinds:

1) translation by $v \in \mathbb{F}_2^n$:

$$f_i(x^n) = x^n + v \qquad (159)$$
$$f_o(y^n) = y^n - v\,. \qquad (160)$$

2) permutation by $\sigma \in S_n$ – a group of all bijections $\{1, \ldots, n\} \to \{1, \ldots, n\}$:

$$f_i(x^n) = (x_{\sigma(1)}, \ldots, x_{\sigma(n)})\,, \qquad (161)$$
$$f_o(y^n) = (y_{\sigma(1)}, \ldots, y_{\sigma(n)}) \qquad (162)$$

It is easy to see that group G acts transitively on both A and B, and thus by Theorem 21 we have:

Theorem 22: Uniform distributions P_X and Q_Y are the saddle point in (81) for the BSC. The value of the saddle point is

$$\min_{P_{X^n}} \max_{Q_{Y^n}} \beta_\alpha(P_{X^nY^n}, P_{X^n}Q_{Y^n}) = (1-\lambda)\beta_L + \lambda\beta_{L+1}\,,$$

where β_ℓ, $\lambda \in [0, 1)$ and the integer L are found from

$$\beta_\ell = \sum_{k=0}^{\ell} \binom{n}{k} 2^{-n} \qquad (163)$$
$$\alpha = (1-\lambda)\alpha_L + \lambda\alpha_{L+1} \qquad (164)$$
$$\alpha_\ell = \sum_{k=0}^{\ell-1} \binom{n}{k} (1-\delta)^{n-k}\delta^k\,. \qquad (165)$$

Remark: The resulting minimax channel coding converse coincides with the classical sphere packing bound, e.g. [1, Theorem 35].

D. Binary erasure channel (BEC)

Recall that $BEC(n, \delta)$ for blocklength n and erasure probability δ is defined as follows: the input alphabet $\mathsf{A} = \mathbb{F}_2^n$, the output alphabet $\mathsf{B} = \{0, e, 1\}^n$, and the transition probabilities are

$$P_{Y^n|X^n}(y^n|x^n) = \begin{cases} \left(\frac{\delta}{1-\delta}\right)^{\mathsf{e}(y^n)}(1-\delta)^n, & (x^n, y^n) - \text{compatible}\,, \\ 0, & \text{otherwise}\,, \end{cases} \qquad (166)$$

where (x^n, y^n) is called compatible if $x_i = y_i$ whenever $y_i \neq e$ and

$$\mathsf{e}(y^n) = \#\{j : y_j = e\}\,.$$

Consider the same group G as for the BSC, except that in the definition (160) of translation by v on the output space, the arithmetic on $\{0, e, 1\}$ is extended from $\mathbb{F}_2^n$ as $0 + e = e$, $1 + e = e$.

Theorem 23: The saddle point in (81) for the BEC is:

$$P^*_{X^n}(x^n) = 2^{-n} \qquad (167)$$
$$Q^*_{Y^n}(y^n) = \lambda\left(\frac{\delta}{1-\delta}\right)^{\mathsf{e}(y^n)}(1-\delta)^n \cdot 1\{\mathsf{e}(y^n) \geq u\}\,, \qquad (168)$$

where the parameter $u \in \mathbb{R}$ depends on α and λ is a normalization factor:

$$\lambda^{-1} = \sum_{e \geq u} 2^{n-e}\binom{n}{e}\delta^e(1-\delta)^{n-e}\,.$$

The value of the saddle point can be represented parametrically as

$$\min_{P_{X^n}} \max_{Q_{Y^n}} \beta_\alpha(P_{X^nY^n}, P_{X^n}Q_{Y^n}) = 2^{u-n}\,, \qquad (169)$$

where

$$\alpha = \sum_{e=0}^{n} \binom{n}{e}\delta^e(1-\delta)^{n-e}2^{-|e-u|^+} \qquad (170)$$

for all $u \in \mathbb{R}$.

Remark: A simple inspection reveals that the resulting channel coding converse bound implied by (7) and (169) coincides exactly with the tight finite-blocklength converse [1, Theorem 38], obtained there by an ad-hoc (BEC-specific) argument.

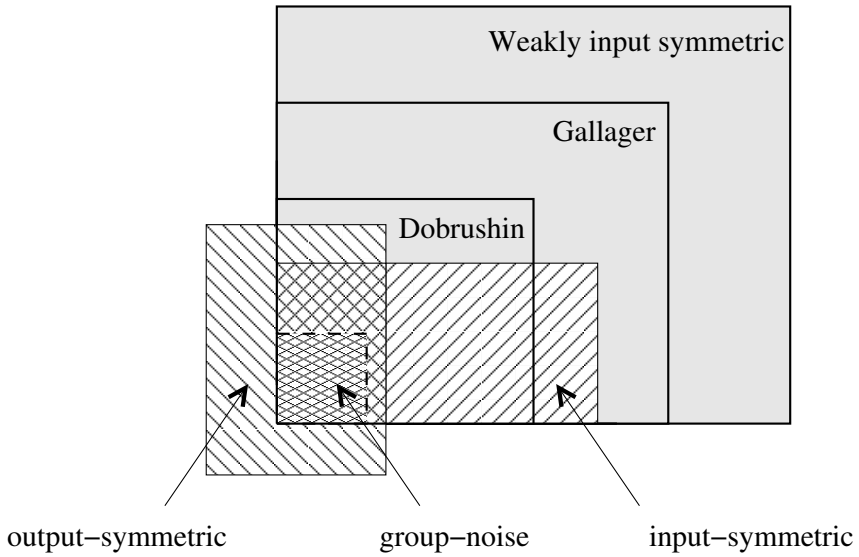

Fig. 2: Schematic representation of inclusions of various classes of channels

Proof: Optimality of (167) immediately follows from Theorem 21. We next compute the value of the saddle point by evaluating $\tilde{\beta}_\alpha(P^*_{X^n}, P_{Y^n|X^n})$. First, it is clear that the most general form of the test achieving $\tilde{\beta}_\alpha$ is:

$$P_{Z|XY}(1|x^n, y^n) = \begin{cases} 0, & (x^n, y^n) - \text{not compatible}\,, \\ f(\mathsf{e}(y^n)) & \text{otherwise} \end{cases} \tag{171}$$

where $f : \{0, \ldots, n\} \to [0, 1]$ is some function. On the other hand, by Theorem 5 function f can further be constrained to be constant over y^n so that

$$\sum_{x^n} 2^{-n} f(\mathsf{e}(y^n)) = \text{const}\,,$$

where summation is over all x^n compatible with a given y^n. Thus, we find that

$$f^*(e) = 2^{-|e-u|^+}\,, \tag{172}$$

for some u. Thus the test is uniquely specified by (171)-(172), resulting in

$$\tilde{\beta}_\alpha(P^*_{X^n}, P_{Y^n|X^n}) = 2^{-|e-u|^+}\,, \tag{173}$$

where α is found from (170). By Theorem 4 we conclude

$$\min_{P^n_X} \max_{Q_{Y^n}} \beta_\alpha(P_{X^nY^n}, P_{X^n}Q_{Y^n}) = 2^{-|e-u|^+}\,.$$

We are left to show that in the dual problem

$$\max_{Q_{Y^n}} \min_{P_{X^n}} \beta_\alpha(P_{X^nY^n}, P_{X^n}Q_{Y^n}) \tag{174}$$

the outer maximization is solved by (168). Note that any Q_{Y^n} that satisfies

$$\beta_\alpha(P^*_{X^nY^n}, P^*_{X^n}Q_{Y^n}) \geq \tilde{\beta}_\alpha(P^*_{X^n}, P_{Y^n|X^n})\,, \tag{175}$$

will automatically be the optimal one since the reverse inequality (which always holds) shows one must have in fact equality in (175).

First, we show how the form (168) of the distribution can be derived. By Theorem 20 it is sufficient to restrict attention to

$$Q_{Y^n}(y^n) = q(\mathsf{e}(y^n))\,, \tag{176}$$

where $q : \{0, \ldots, n\} \to [0, 1]$ is a function satisfying the normalization requirement:

$$\sum_{e=0}^{n} 2^{n-e} \binom{n}{e} q(e) = 1\,.$$

For any such Q_{Y^n} the minimizing P_{X^n} in (174) is given by the uniform $P^*_{X^n}$ (Theorem 20). By definition of β_α and (171) we have

$$\beta_\alpha(P_{X^nY^n}, P^*_{X^n}Q_{Y^n}) = \min \sum_{e=0}^{n} \binom{n}{e} q(e) f(e)\,, \tag{177}$$

where minimum is taken over all f such that

$$\sum_{e=0}^{n} \binom{n}{e} \delta^e (1-\delta)^{n-e} f(e) \geq \alpha\,.$$

It is natural to look for Q_{Y^n} such that the optimizing f in (177) were given by (172). Then by the Neyman-Pearson lemma, it is clear that we must have

$$q(e) = \lambda \delta^e (1-\delta)^{n-e}$$

for all $e \geq u$ and some λ. It is natural to complete the definition of $q(e)$ by taking it to be zero for $e < u$, which results in (168).

Finally, to show (175) with Q_{Y^n} given by (168) consider the test

$$P_{Z|XY}(1|x^n, y^n) = \begin{cases} 0, & (x^n, y^n) - \text{not compatible}\,, \\ \tau & \mathsf{e}(y^n) \geq u\,, \\ 1 & \mathsf{e}(y^n) < u \end{cases} \tag{178}$$

with τ chosen to satisfy

$$P^*_{X^nY^n}[Z = 1] = \alpha$$

where α is given by (170). This test is optimal by Neyman-Pearson lemma and it achieves

$$P^*_{X^n}Q^*_{Y^n}[Z=1] = 2^{u-n}\,,$$

which is shown by a direct verification. Thus by (173) the (175) follows. ∎

E. General discrete memoryless channel (DMC)

In the previous section we have seen an example that the optimal distribution Q_{Y^n} may not be a product distribution. For an arbitrary DMC, by the action of the permutation group S_n and Theorem 20 one may restrict attention to exchangeable distributions P_{X^n} and Q_{Y^n}. In this section we demonstrate, however, that it is safe to further restrict Q_{Y^n} to a product distributions at least as far as the error-exponent asymptotic is concerned.

We follow the notation of [26], in particular $\mathsf{A} = \mathcal{X}^n, \mathsf{B} = \mathcal{Y}^n$, where $|\mathcal{X}|, |\mathcal{Y}| < \infty$ and the random transformation is

$$P_{Y^n|X^n}(y^n|x^n) = \prod_{j=1}^n W(y_j|x_j)\,,$$

where $W : \mathcal{X} \to \mathcal{Y}$ is a fixed stochastic matrix. The sphere-packing exponent at rate R is defined as

$$E_{sp}(R) \stackrel{\triangle}{=} \max_P \min_{V: I(P,V)\le R} D(V||W|P)\,,$$

where P ranges over all distributions on $\mathcal{X}$ and V over all stochastic matrices $V : \mathcal{X} \to \mathcal{Y}$, see [26, Chapter 10].

Denote by $\epsilon_{mc}(n,R)$ the smallest ϵ satisfying the minimax converse (7):

$$\epsilon_{mc}(n,R) = \min\left\{\epsilon\colon \inf_{P_{X^n}} \sup_{Q_{Y^n}} \beta_{1-\epsilon}\left(P_{X^nY^n}, P_{X^n}Q_{Y^n}\right) \le \exp\{-nR\}\right\}. \tag{179}$$

Theorem 24: For any DMC W there exist sequences $\delta_n \to 0$ and $\delta'_n \to 0$ such that for all rates $R > 0$ for which $E_{sp}(R) < \infty$ we have

$$\epsilon_{mc}(n, R+\delta'_n) = \exp\{-n(E_{sp}(R)+\delta_n)\}\,,$$

while for all other rates

$$\epsilon_{mc}(n, R+\delta'_n) = 0\,,$$

for all n sufficiently large.

Remark: Since at low rates the sphere packing bound on the error exponent is known to be non-tight [16], and since the Poor-Verdú bound [17] is a consequence of (7), see [10, Section 2.7.3], Theorem 24 settles in the negative the conjecture about the tightness of the Poor-Verdú bound on the error exponent [17]. For the BEC this has been shown previously in [27].

Proof: First we show

$$\epsilon_{mc}(n, R+\delta'_n) \ge \exp\{-n(E_{sp}(R)+\delta_n)\}\,, \tag{180}$$

for a suitably chosen $\delta_n, \delta'_n \to 0$. The proof of (180) shows that the sphere-packing error-exponent can be derived from the minimax converse by taking Q_{Y^n} to be a product distribution, cf. [10, Section 2.7.3]. Then it is sufficient to show that

$$\sup_{Q_{Y^n}=(Q_Y)^n} \beta_{1-\epsilon_n}(P_{X^nY^n}, P_{X^n}Q_{Y^n}) \le \exp\{-n(R+\delta'_n)\} \tag{181}$$

implies

$$\epsilon_n \ge \exp\{-n(E_{sp}(R)+\delta_n)\}\,, \tag{182}$$

where we restricted to product distributions Q_{Y^n} and P_{X^n} corresponds to the optimal distribution in (179).

Since the symmetric group S_n is a natural symmetry group for a DMC of blocklength n, then according to Theorem 20 P_{X^n} is a convex combination

$$P_{X^n} = \sum_j \lambda_j P^{(j)}_{X^n}\,, \qquad \sum_j \lambda_j = 1, \lambda_j \ge 0 \tag{183}$$

where j ranges over all n-types on $\mathcal{X}$ and $P^{(j)}_{X^n}$ is a distribution uniform on the j-th type. If decomposition (183) consists of a single non-zero term, then (181) (even with $\delta'_n = 0$) implies (182) by a standard argument [28]. In general, since the number of different types is bounded by $n^{|\mathcal{X}|-1}$, there is j_0 with $\lambda_{j_0} \ge \frac{1}{n^{|\mathcal{X}|-1}}$, and thus the general case follows from the following self-evident result:

Lemma 25: Let $P_X = \sum_j \lambda_j P_{X_j}$ be a convex combination of P_{X_j} with $\lambda_j > 0$. Then for all Q_Y and j we have

$$\beta_{1-\epsilon}(P_{XY}, P_XQ_Y) \ge \lambda_j \beta_{1-\epsilon\lambda_j^{-1}}(P_{X_jY}, P_{X_j}Q_Y)\,.$$

Furthermore, if supports of P_{X_j} are pairwise disjoint then

$$\beta_{1-\epsilon}(P_{XY}, P_XQ_Y) = \inf_{\sum_j \lambda_j\epsilon_j = \epsilon} \sum_j \lambda_j \beta_{1-\epsilon_j}(P_{X_jY}, P_{X_j}Q_Y)\,.$$

To prove the converse of (180), we notice that by Theorem 4

$$\epsilon_{mc}(n,R) = \min\left\{\epsilon\colon \inf_{P_{X^n}} \tilde\beta_{1-\epsilon}\left(P_{X^n}, P_{Y^n|X^n}\right) \le \exp\{-nR\}\right\}. \tag{184}$$

Thus, it is sufficient to construct a P_{X^n} and one test $P_{Z|X^nY^n}$ which achieves

$$\sup_{y^n} \sum_{x^n} P_{X^n}(x^n) P_{Z|X^nY^n}(1|x^n,y^n) \le \exp\{-n(R+\delta'_n)\}\,, \tag{185}$$

$$P_{X^nY^n}[Z=0] \le \exp\{-n(E_{sp}(R)+\delta_n)\} \tag{186}$$

for some $\delta_n \to 0$ and $\delta'_n \to 0$.

Recall that, [26, Problem 10.28] and [29], for any rate $R > 0$ for which $E_{sp}(R) < \infty$ there exists a positive integer ℓ and a sequence of codebooks $\mathcal{C}$ list-decodable to a constant list size ℓ with probability of error

$$\epsilon_n \le \exp\{-n(E_{sp}(R)+\delta_n)\}\,, \tag{187}$$

and of asymptotic rate R:

$$|\mathcal{C}| = \exp\{nR + o(n)\}\,.$$

To each codebook $\mathcal{C}$ we define a distribution P_{X^n}

$$P_{X^n}(x^n) = \exp\{-nR\}1\{x^n \in \mathcal{C}\}\,,$$

and the test

$$P_{Z|X^n,Y^n}(1|x^n,y^n) = 1\{x^n \in L(y^n)\}\,,$$

where $L(y^n)$ is the list output by the decoder. Elementary calculation then shows that in (185) we have

$$\sup_{y^n}\sum_{x^n} P_{X^n}(x^n)P_{Z|X^nY^n}(1|x^n,y^n) = \frac{\ell}{|\mathcal{C}|} = \exp\{-nR+\delta_n'\}\,, \tag{188}$$

for a suitably chosen $\delta_n' \to 0$. Hence (180) holds. Similarly, for rates with $E_{sp}(R) = \infty$ there exist zero-error constant list-size codes implying we have $\epsilon_n = 0$ in (187) and the right-hand side of (186). ∎

F. Additive white Gaussian noise (AWGN) channel

The AWGN channel $AWGN(n,P)$ is given by $\mathsf{A} = \mathbb{R}^n$ and $\mathsf{B} = \mathbb{R}^n$ and $P_{Y^n|X^n}$ acts by adding a white Gaussian noise:

$$Y^n = X^n + Z^n\,, \tag{189}$$

where $Z^n \sim \mathcal{N}(0,\mathbf{I}_n)$ – is the isotropic standard normal vector. We impose an *equal-power constraint* on the codebook: each codeword $c_i, i = 1,\dots,M$ must satisfy

$$||c_i||^2 = nP\,. \tag{190}$$

By a standard $n \to n+1$ argument this power constraint can be assumed without loss of generality, e.g. [1, Lemma 39].

Regardless of the location of the M codewords on the power sphere, it is clear that the optimal (maximum likelihood) decoder operates on the basis of $\frac{y^n}{||y^n||}$ only. Thus, we may replace $P_{Y^n|X^n}$ with an equivalent random transformation $P_{B|A}: \mathbb{S}^{n-1} \to \mathbb{S}^{n-1}$:

$$B = \frac{\sqrt{nP}A + Z^n}{||A+Z^n||}\,, \tag{191}$$

where the input A and the output B are elements of $\mathbb{S}^{n-1}$, an $(n-1)$-dimensional sphere embedded canonically into $\mathbb{R}^n$. A regular group of symmetries G can be taken to be $SO(n)$ – the special orthogonal group, which acts in a standard manner on both the input and the output $\mathbb{S}^{n-1}$. Since this action is transitive, Theorem 21 implies that for the equivalent channel (191) the saddle point is achieved by the uniform distributions on the sphere. The resulting minimax converse bound is precisely the Shannon's cone-packing [30].

VII. Discussion

We conclude with several observations regarding the results we have obtained.

First, we have shown that the optimization over the input distributions in the minimax converse, Theorem 1, is a convex problem which is further simplified by the channel symmetries present in many practical communication channels. Thus not only does the minimax converse strengthen known information-spectrum bounds, see [10, Section 2.7.3], but it also simplifies the calculation. In particular, we have demonstrated that for symmetric channels one may restrict attention to memoryless input distributions (in both the information-spectrum bounds or the minimax converse). For general memoryless channels, one may restrict to exchangeable distributions.

Second, in all of the examples considered in the paper the optimal input distribution turned out to coincide with the distribution yielding (e.g., via random coding) the best known achievability bounds. Therefore, one naturally expects that in cases where the saddle-point input distribution is non-product, we may hope to improve non-asymptotic achievability bounds by considering non-product input distributions.

Third, the example of BEC (Section VI-D) demonstrated that the saddle-point output distribution may be non-product. Interestingly, BEC is one of a few examples of channels with zero in the logarithmic term in the expansion, cf. [1]:

$$\log M^*(n,\epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(1)\,, \quad n \to \infty \tag{192}$$

where $M^*(n,\epsilon)$ is the maximal cardinality of the code of blocklength n and error probability ϵ, C is the channel capacity and V – the channel dispersion. Note that the behavior of β_α for product distributions is given by, e.g. [10, (2.71)],

$$\log\beta_\alpha(P^n,Q^n) = -nD(P||Q) - \sqrt{nV(P||Q)}Q^{-1}(\alpha) - \frac{1}{2}\log n + O(1)\,, \tag{193}$$

implying that an upper-bound obtained from (10):

$$\log M^*(n,\epsilon) \le \sup_{P_{X^n}} -\log\beta_{1-\epsilon}(P_{X^nY^n}, P_{X^n}Q_{Y^n})$$

cannot yield a zero $\log n$ term whenever Q_{Y^n} is a product distribution. However, since we have shown that the exact minimax converse for BEC coincides with the (BEC-specific) converse used in [1] to show (192) we conclude that Theorem 1 may still be used to show tight estimates for the $\log n$ term even in case when this term is $0\cdot\log n$ and that in such cases the optimal Q_{Y^n} is necessarily non-product. For more on the $\log n$ term in expansions (192) we refer to [10, Section 3.4.5] and [7].

Overall, we conclude that studying the saddle point (81) provides interesting new insights regarding the structure and performance of optimal channel codes.

Acknowledgment

We are grateful to R. Nasser for the comments.

References

[1] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[2] ——, "Feedback in the non-asymptotic regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4903–4925, Aug. 2011.

[3] ——, "Minimum energy to send k bits with and without feedback," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4880–4902, Aug. 2011.

[4] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *CoRR*, vol. abs/1102.3944, 2011. [Online]. Available: http://arxiv.org/abs/1102.3944

[5] P. Moulin, "Finite-blocklength universal coding for multiple-access channels," in *DARPA ITMANET meeting*, Stanford, CA, Jan. 2011.

[6] W. Matthews, "A linear program for the finite block length converse of Polyanskiy-Poor-Verdú via non-signalling codes," *IEEE Trans. Inf. Theory*, vol. 58, no. 12, pp. 7036 – 7044, Dec. 2012.

[7] P. Moulin, "The log-volume of optimal codes for memoryless channels, within a few nats," in *2012 Inf. Theory and Appl. Workshop (ITA)*, San Diego, CA, Feb. 2012.
[8] Y. Polyanskiy, "Asynchronous communication: exact synchronization, universality and dispersion," *IEEE Trans. Inf. Theory*, 2012, to appear. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/async.pdf
[9] Y. Polyanskiy and S. Verdú, "Empirical distribution of good channel codes with non-vanishing error probability," preprint. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/optcodes_journal.pdf
[10] Y. Polyanskiy, "Channel coding: non-asymptotic fundamental limits," Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 2010, available: http://www.mit.edu/~ypol.
[11] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, 1994.
[12] E. Lehmann and J. Romano, *Testing statistical hypotheses*, 3rd ed. New York: Springer Verlag, 2005.
[13] A. Wald, *Statistical decision functions.* New York: Wiley, 1950.
[14] L. LeCam, "An extension of Wald's theory of statistical decision functions," *Ann. Math. Stat.*, vol. 26, no. 1, pp. 69–81, 1955.
[15] K. Fan, "Minimax theorems," *Proc. Nat. Acad. Sci.*, vol. 39, pp. 42–47, 1953.
[16] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels I," *Inf. Contr.*, vol. 10, pp. 65–103, 1967.
[17] H. V. Poor and S. Verdú, "A lower bound on the error probability in multihypothesis testing," *IEEE Trans. Inf. Theory*, vol. 41, no. 6, pp. 1992–1993, 1995.
[18] E. Çinlar, *Probability and Stochastics*. New York: Springer Verlag, 2011.
[19] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton University Press, 1996.
[20] I. Csiszár, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Probab.*, vol. 12, no. 3, pp. 768–793, 1984.
[21] N. Dunford and J. Schwartz, *Linear Operators: General theory*. New York: Interscience Publishers, 1958, vol. 1.
[22] R. L. Dobrushin, "Asymptotic bounds on error probability for transmission over DMC with symmetric transition probabilities," *Theor. Probability Appl.*, vol. 7, pp. 283–311, 1962.
[23] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[24] M.-K. Siu, "Which latin squares are cayley tables?" *Amer. Math. Monthly*, vol. 98, no. 7, pp. 625–627, Aug. 1991.
[25] E. Hof, I. Sason, and S. Shamai, "Performance bounds for nonbinary linear block codes over memoryless symmetric channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 3, pp. 977–996, Mar. 2009.
[26] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
[27] F. Alajaji, P.-N. Chen, and Z. Rached, "A note on the Poor-Verdú upper bound for the channel reliability function," *IEEE Trans. Inf. Theory*, vol. 48, no. 1, pp. 309–313, jan 2002.
[28] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 405–417, 1974.
[29] P. Elias, "List decoding for noisy channels," MIT, Tech. Rep. RLE-TR-335, Sep. 1957.
[30] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell Syst. Tech. J.*, vol. 38, pp. 611–656, 1959.

Yury Polyanskiy (S'08-M'10) received the M.S. degree (Hons.) in applied mathematics and physics from the Moscow Institute of Physics and Technology, Moscow, Russia in 2005 and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ in 2010. In 2000-2005 he lead development of the embedded software in the Department of Surface Oilfield Equipment, Borets Company LLC (Moscow). In 2011 Dr. Polyanskiy joined MIT as an Assistant Professor of Electrical Engineering and Computer Science (EECS), and a member of Laboratory of Information and Decision Systems.

His research interests include information theory, error-correcting codes, wireless communication and the theory of random processes. Over the years Dr. Polyanskiy won the 2011 Best Paper Award from IEEE Information Theory Society, the Best Student Paper Awards at the 2008 and 2010 IEEE International Symposiums on Information Theory (ISIT). His final year of graduate studies was supported by a Princeton University Honorific Dodds Fellowship (2009-2010). In 2012 Yury was selected to hold a Robert J. Shillman (1974) Career Development Professorship of EECS.