

The Mutual Information In The Vicinity of Capacity-Achieving Input Distributions

Barış Nakiboğlu, *Member, IEEE*, and Hao-Chung Cheng, *Member, IEEE*,

Abstract—On small neighborhoods of the capacity-achieving input distributions, the decrease of the mutual information with the distance to the capacity-achieving input distributions is bounded below by a linear function of the square of the distance to the capacity-achieving input distributions for all channels with (possibly multiple) linear constraints and finite input sets using an identity due to Topsøe and Pinsker’s inequality. Counter examples demonstrating non-existence of such a quadratic bound are provided for the case of infinite many linear constraints and the case of infinite input sets. Using a Taylor series approximation, rather than Pinsker’s inequality, the exact characterization of the slowest decrease of the mutual information with the distance to the capacity-achieving input distributions is determined on small neighborhoods of the capacity-achieving input distributions. Analogous results are established for classical-quantum channels whose output density operators are defined on a separable Hilbert spaces. Implications of these observations for the channel coding problem and applications of the proof technique to related problems are discussed.

I. INTRODUCTION

In his seminal paper [1], Strassen proved for channels with finite input and output sets that there exist positive constants γ and δ for which the mutual information satisfies

$$I(p; W) \leq C - \gamma \|p - p_*\|^2 \quad \text{if } \|p - p_*\| \leq \delta \quad (1)$$

where p_* is the projection of p to the set of all capacity-achieving input distributions Π in the underlying Euclidean space, and hence $\|p - p_*\|$ is the distance of p to Π . Strassen’s brief and elegant argument relies implicitly on the fact that for any $p \notin \Pi$, the direction $p - p_*$ cannot be simultaneously orthogonal to the gradient of mutual information at p_* , i.e., orthogonal to $D(W\|q_W)$, and in the kernel of the linear transformation relating the input distributions to the output distributions, i.e., in \mathcal{K}_W . We believe one of the claims in Strassen’s proof, which holds trivially for some channels, requires a more nuanced justification to be valid for all channels with finite input and output alphabets. Nevertheless, the claim can be established as is using polyhedral convexity as we discuss in more detail in Appendix A.

Strassen’s bound (1), plays an important role in establishing sharp impossibility results for the channel coding theorem, see [1]–[3]. Determining an explicit expression for (δ, γ) pair for which (1) holds is also worthwhile because of this role.

B. Nakiboğlu is with the Department of Electrical and Electronics Engineering Middle East Technical University 06800 Ankara, Turkey (✉ 0000-0001-7737-5423).

H-C. Cheng is with Department of Electrical Engineering, Graduate Institute of Communication Engineering, and Department of Mathematics, National Taiwan University Taipei 10617, Taiwan (R.O.C.) and with Hon Hai (Foxconn) Quantum Computing Centre (✉ 0000-0003-4499-4679)

One of the claims of Polyanskiy, Poor, and Verdú in [2] is to establish (1) with an explicit coefficient γ . They apply an orthogonal decomposition to assert $p - p_* = v_0 + v^\perp$, where v_0 is the projection of $p - p_*$ to \mathcal{K}_W . Then they argue $\langle v_0, D(W\|q_W) \rangle \leq -\Gamma \|v_0\|$ for some $\Gamma > 0$, see [2, (500)]. This claim, however, is wrong for some p ’s on certain channels as we demonstrate through a particular channel in Appendix B.

In our judgment, the issue overlooked in [2] is the following. The projection of $p - p_*$ to the subspace of \mathcal{K}_W can have a non-zero component that is also orthogonal to $D(W\|q_W)$; the principle used by Strassen in [1] asserts merely that this component cannot be the $p - p_*$ vector itself. This principle can be strengthened using polyhedral convexity to assert that the angle between the $p - p_*$ vector and the subspace of \mathcal{K}_W that is orthogonal to $D(W\|q_W)$ cannot be less than a positive constant, determined by the channel. In §III, we establish this fact for the case with multiple linear constraints. In §IV, we use this observation together with Pinsker’s inequality to prove (1) with explicit expressions for γ and δ for channels with finitely many linear constraints and finite input sets using an orthogonal decomposition, similar to [2]. Using the trace norm in place of the total variation norm the proof presented in §IV applies to classical-quantum channels with finite input sets whose density operators are defined on a separable Hilbert spaces, as later demonstrate in §VI-A. The orthogonal decomposition idea itself can be strengthened by considering the orthogonal decomposition to a closed convex cone and its polar cone, via Moreau’s decomposition theorem. In §V, we employ such a decomposition together with a Taylor series expansion to determine the order and the coefficient of the leading term characterizing the qualitative behavior of the slowest decay of the mutual information in the vicinity of the capacity-achieving input distributions. In other words, we determine the largest γ_1 satisfying $I(p; W) \leq C - \gamma_1 \|p - p_*\|$ for all p and for the cases when the largest γ_1 is zero, we determine the best γ coefficient for Strassen’s bound in (1) for the cases when the slowest decrease is quadratic. In §VI-B we extend this result to classical-quantum channels with finite input sets.

Recently in [4], Cao and Tomamichel presented a proof of (1), in the spirit of [1]. First the cone generated by the vectors $p - p_*$ for $p \notin \Pi$ is proved to be closed, and then a second-order Taylor series expansion for the parametric family of functions $\{I(p_* + \tau(p - p_*); W)\}_{p \notin \Pi}$ at $\tau = 0$ with a uniform approximation error term for all $p \notin \Pi$ is obtained. Then (1) is established using the extreme value theorem, the fact that $p - p_*$ cannot be an element of \mathcal{K}_W that is orthogonal to $D(W\|q_W)$, and the Taylor series expansion. Cao and Tomamichel, later generalized their analysis to the case with

finitely many linear constraints, in [5].

In §II, we introduce our notation and review fundamental observations about the Kullback–Leibler divergence, mutual information, Shannon capacity and Shannon center.

In §III, we review essential concepts and results from convex analysis and prove the positivity of the aforementioned minimum angle. In §IV we prove (1) for any channel with possibly multiple linear constraints and a finite input set using the Pinsker’s inequality and the minimum angle. Unlike [1], [2], [4], [5], we will not need to assume that the channel has a finite output set. In §IV, we demonstrate the necessity of finiteness of the input set and finiteness of the number of linear constraints for (1) by providing two finite output set channels violating (1). Example 1 describes a channel with three input letters and infinitely many linear constraints for which quadratic decrease does not hold. Example 2 describes a channel with a countably infinite input set for which the capacity can be achieved by input distributions that are bounded away from Π .

In §V, we use Moreau’s decomposition theorem together with a Taylor series expansion to characterize the slowest decay of the mutual information with the distance to the capacity-achieving input distributions by determining the order and the coefficient of the leading term of its Taylor expansion. Example 3 describes a channel with a finite input set and a countably infinite output set for which one cannot use the Taylor series expansion to establish (1); nevertheless, one can establish (1) using Pinsker’s inequality as in §IV.

In §VI, we first recall the quantum information-theoretic framework and review the fundamental observations about quantum information-theoretic quantities in a way analogous to our discussion in §II. Then in §VI-A we prove (1) for any classical-quantum channels with possibly multiple linear constraints, a finite input set, and a separable Hilbert space using the quantum Pinsker’s inequality and the minimum angle, similar to §IV. In §VI-B we characterize the slowest decay of the quantum mutual information around the capacity-achieving input distributions for the above mentioned channel in a way analogous to §VI-B.

In §VII, we discuss the implications of the analysis presented and possible applications of the proof techniques to some related problems.

II. INFORMATION THEORETIC PRELIMINARIES

We denote the set of all probability mass functions on countable subsets of a set \mathcal{X} by $\mathcal{P}(\mathcal{X})$ and the set of probability measures on a measurable space $(\mathcal{Y}, \mathcal{Y})$ by $\mathcal{P}(\mathcal{Y})$. A $w \in \mathcal{P}(\mathcal{Y})$ is said to be absolutely continuous in a $q \in \mathcal{P}(\mathcal{Y})$, i.e., $w \prec q$, if $w(\mathcal{E}) = 0$ for all $\mathcal{E} \in \mathcal{Y}$ satisfying $q(\mathcal{E}) = 0$.

The Kullback–Leibler divergence $D(w\|q)$ is defined for any $w, q \in \mathcal{P}(\mathcal{Y})$ as

$$D(w\|q) := \begin{cases} \int \left(\frac{dw}{dq} \ln \frac{dw}{dq} \right) dq & w \prec q \\ \infty & w \not\prec q \end{cases}. \quad (2)$$

The Kullback–Leibler divergence is a non-negative function and $D(w\|q) = 0$ iff $w = q$. Furthermore, the Kullback–Leibler

is bounded from below in terms of the total variation norm via Pinsker’s inequality, [6],

$$D(w\|q) \geq \frac{1}{2} \|w - q\|_1^2. \quad (3)$$

where $\|\cdot\|_1$ is the total variation norm defined for all signed measures μ on $(\mathcal{Y}, \mathcal{Y})$, i.e. for all $\mu \in \mathcal{M}(\mathcal{Y})$, as

$$\|\mu\|_1 := \int \left| \frac{d\mu}{d\nu} \right| d\nu,$$

where ν is any reference measure satisfying $\mu \prec \nu$. On the other hand the Kullback–Leibler divergence is bounded above by χ^2 divergence, see [7, Theorem 5.1], [8, Theorem 5],

$$\chi^2(w\|q) \geq \ln(1 + \chi^2(w\|q)) \geq D(w\|q) \quad (4)$$

where χ^α divergence is introduced by Vajda, see [9, p. 246], [10], [11]. For $\alpha > 1$ case χ^α divergence between finite signed measure w (i.e., $w \in \mathcal{M}(\mathcal{Y})$) satisfying $\|w\|_1 < \infty$ and probability measure q (i.e., $q \in \mathcal{P}(\mathcal{Y})$) is defined as

$$\chi^\alpha(w\|q) := \begin{cases} \int \left| \frac{dw}{dq} - 1 \right|^\alpha dq & w \prec q \\ \infty & w \not\prec q \end{cases}. \quad (5)$$

Note that $\chi^\alpha(w\|q) \geq 0$ and the equality holds iff $w = q$. Using a standard Taylor series analysis $D(w\|q)$ can be bounded in terms of $\chi^2(w\|q)$ and $\chi^3(w\|q)$ as follows

$$\frac{1}{2} \chi^3(w\|q) \geq D(w\|q) - \frac{1}{2} \chi^2(w\|q) \geq -\frac{1}{6} \chi^3(w\|q) \quad (6)$$

provided that $\chi^3(w\|q) < \infty$; see Appendix C for a proof.

A channel W is a $\mathcal{P}(\mathcal{Y})$ valued function defined on the input set \mathcal{X} , where \mathcal{Y} is the σ -algebra of the output space $(\mathcal{Y}, \mathcal{Y})$, i.e., a channel is a function of the form $W : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$. For any $W : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$ and $p \in \mathcal{P}(\mathcal{X})$ we define the conditional Kullback–Leibler divergence $D(W\|q|p)$ as

$$D(W\|q|p) := \sum_x p(x) D(W(x)\|q) \quad \forall p \in \mathcal{P}(\mathcal{X}).$$

For any channel $W : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$ and $p \in \mathcal{P}(\mathcal{X})$, the mutual information $I(p; W)$ is defined as

$$I(p; W) := D(W\|q_p|p), \quad (7)$$

where $q_p \in \mathcal{P}(\mathcal{Y})$ is the output distribution induced by the input distribution p , for any $p \in \mathcal{P}(\mathcal{X})$, which is defined more generally for any $v : \mathcal{X} \rightarrow \mathbb{R}$ with a countable support satisfying $\sum_x |v(x)| < \infty$ as

$$q_v := \sum_x v(x) W(x). \quad (8)$$

The following identity, due to Topsøe [12], can be confirmed by substitution

$$D(W\|q|p) = I(p; W) + D(q_p\|q) \quad (9)$$

for all $p \in \mathcal{P}(\mathcal{X})$ and $q \in \mathcal{P}(\mathcal{Y})$.

For any convex constraint set $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$, the Shannon capacity $C_{\mathcal{A}}$ and the set of all capacity-achieving input distributions in \mathcal{A} , i.e., $\Pi_{\mathcal{A}}$, are defined as

$$C_{\mathcal{A}} := \sup_{p \in \mathcal{A}} I(p; W), \quad (10)$$

$$\Pi_{\mathcal{A}} := \{p \in \mathcal{A} : I(p; W) = C_{\mathcal{A}}\}. \quad (11)$$

With a slight abuse of notation, we denote $C_{\mathcal{P}(X)}$ and $\Pi_{\mathcal{P}(X)}$ by C_W and Π_W .

If $C_A < \infty$ then, by [13], [14], there exists a unique Shannon center $q_A \in \mathcal{P}(\mathcal{Y})$ satisfying,

$$D(W \| q_A | p) \leq C_A \quad \forall p \in \mathcal{A}. \quad (12)$$

Furthermore, $D(q_{p_*} \| q_A) = 0$ and thus $q_{p_*} = q_A$ for any $p_* \in \Pi_A$ by (3) and (9).

For the rest of this section, we assume that the input set \mathcal{X} is finite and the constraint set \mathcal{A} is closed. Then $C_A < \infty$ because $C_A \leq \ln |\mathcal{X}|$ and thus a unique Shannon center q_A exists. Furthermore, as a result of the extreme value theorem, the supremum in (10) is achieved, i.e. $\Pi_A \neq \emptyset$, because $I(p; W)$ is continuous in p and \mathcal{A} is closed and bounded, i.e., compact. Furthermore, Π_A is a closed set because it is the preimage of a closed set, for a continuous function.

We interpret real valued functions on a finite set \mathcal{X} as the elements of a Euclidean vector space $\mathbb{R}^{\mathcal{X}}$. For any $|\mathcal{X}|$ -by- $|\mathcal{X}|$ positive semidefinite matrix A , we define the inner product $\langle \cdot, \cdot \rangle_A : \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}} \rightarrow \mathbb{R}$ and the norm $\|\cdot\|_A : \mathbb{R}^{\mathcal{X}} \rightarrow \mathbb{R}_{\geq 0}$ as

$$\begin{aligned} \langle f, g \rangle_A &:= f^T A g & \forall f, g \in \mathbb{R}^{\mathcal{X}}, \\ \|f\|_A &:= \sqrt{\langle f, f \rangle_A} & \forall f \in \mathbb{R}^{\mathcal{X}}. \end{aligned}$$

When A is the identity matrix, we denote the inner product and the norm by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$, respectively.

The non-negativity of the mutual information, (9), and (12), imply $D(q_p \| q_A) \leq C_A < \infty$. Thus for any $p_* \in \Pi_A$ and $p \in \mathcal{A}$ as a result of (9) we have

$$\begin{aligned} I(p; W) &= D(W \| q_A | p) - D(q_p \| q_A) \\ &= I(p_*; W) + \langle p - p_*, D(W \| q_A) \rangle - D(q_p \| q_A) \\ &= C_A + \langle p - p_*, D(W \| q_A) \rangle - D(q_p \| q_A), \end{aligned} \quad (13)$$

where $D(W \| q_A)$ is a column vector whose rows are $D(W(x) \| q_A)$'s. The second term in (13) is non-positive by (12) and its kernel is \mathcal{K}_A^d defined as follows

$$\mathcal{K}_A^d := \{v \in \mathbb{R}^{\mathcal{X}} : \langle v, D(W \| q_A) \rangle = 0\}. \quad (14)$$

The third term is non-positive by (3) and its kernel is the Kernel of the channel, i.e., \mathcal{K}_W defined¹ in the following, because $q_p = q_{p-p_*} + q_A$ and the Kullback–Leibler divergence is zero iff its arguments are equal;

$$\mathcal{K}_W := \left\{ v \in \mathbb{R}^{\mathcal{X}} : \left\| \sum_x v(x) W(x) \right\|_1 = 0 \right\}. \quad (15)$$

The Shannon center not only allows us rewrite (13), but also allows us to characterize Π_A as the elements of \mathcal{A} satisfying certain linear constraints. To see why first note that $q_p = q_A$ and $D(W \| q_A | p) = C_A$ imply $I(p; W) = C_A$ by (7). The existence of a unique Shannon center implies that the converse statement is true as well, i.e., $I(p; W) = C_A$ implies $q_p = q_A$ and $D(W \| q_A | p) = C_A$. Hence,

$$\Pi_A = \mathcal{A} \cap \mathcal{S}_A, \quad (16)$$

where \mathcal{S}_A is an affine subset of $\mathbb{R}^{\mathcal{X}}$ defined below

$$\mathcal{S}_A := \{v \in \mathbb{R}^{\mathcal{X}} : \langle v, D(W \| q_A) \rangle = C_A \text{ and } q_v = q_A\}, \quad (17)$$

¹Note that the total variation norm can be replaced by any norm on $\mathcal{M}(\mathcal{Y})$.

for q_v is defined in (8).

We define δ neighborhood Π_A^δ of the set of all capacity-achieving input distributions Π_A as

$$\Pi_A^\delta := \{p \in \mathcal{A} : \min_{p_* \in \Pi_A} \|p - p_*\| \leq \delta\}. \quad (18)$$

Note that we can use minimum instead of infimum in the definition because $\|\cdot\|$ is a continuous function and Π_A is a closed and bounded, i.e., compact, set.

III. PRELIMINARIES ON CONVEX ANALYSIS

Let \mathcal{A} be a closed convex subset of the Euclidean space \mathbb{R}^n . Then by [15, Proposition A.5.2.1], the *tangent cone* of \mathcal{A} at $p_* \in \mathcal{A}$ is a closed convex cone that can be expressed as the closure of the cone generated by $\{p - p_* : p \in \mathcal{A}\}$:

$$\mathcal{T}_A(p_*) = \text{cl}(\text{cone}(\mathcal{A} - p_*)). \quad (19)$$

The *normal cone* of \mathcal{A} at a point $p_* \in \mathcal{A}$ is

$$\mathcal{N}_A(p_*) := \{s \in \mathbb{R}^n : \langle s, p - p_* \rangle \leq 0, \forall p \in \mathcal{A}\}. \quad (20)$$

Thus the normal cone $\mathcal{N}_A(p_*)$ is a closed convex cone, as well. Furthermore,

$$\mathcal{T}_A(p_*) \cap \mathcal{N}_A(p_*) = \{0\} \quad \forall p_* \in \mathcal{A}. \quad (21)$$

because the normal cone is the polar of the tangent cone, i.e., $\mathcal{N}_A(p_*) = \mathcal{T}_A(p_*)^\circ$, by [15, Proposition A.5.2.4], where the polar of a convex cone \mathcal{C} is defined as, [15, A.3.2.1]

$$\mathcal{C}^\circ := \{s \in \mathbb{R}^n : \langle s, v \rangle \leq 0, \forall v \in \mathcal{C}\}. \quad (22)$$

Let Π be a closed convex set in \mathbb{R}^n , then the *projection* of a point $p \in \mathbb{R}^n$ onto Π is the unique point $P_\Pi(p)$ satisfying

$$P_\Pi(p) = \arg \min_{p_* \in \Pi} \|p - p_*\| \quad \forall p \in \mathbb{R}^n,$$

see [15, p. 46]. Then by [15, Theorem A.3.1.1]

$$p_* = P_\Pi(p) \iff \langle p - p_*, s - p_* \rangle \leq 0 \quad \forall s \in \Pi.$$

In other words, $p_* = P_\Pi(p)$ iff $p - p_* \in \mathcal{N}_\Pi(p_*)$ for the normal cone defined in (20). On the other hand, if $\Pi \subset \mathcal{A}$ for a closed convex set \mathcal{A} , then $p - p_* \in \mathcal{T}_A(p_*)$ for the tangent cone defined in (19) for all $p \in \mathcal{A}$, as well. Thus

$$p_* = P_\Pi(p) \iff p - p_* \in \mathcal{N}_\Pi^A(p_*) \quad \forall p \in \mathcal{A}. \quad (23)$$

where $\mathcal{N}_\Pi^A(p_*)$ is defined as

$$\mathcal{N}_\Pi^A(p_*) := \mathcal{T}_A(p_*) \cap \mathcal{N}_\Pi(p_*). \quad (24)$$

For all $p \in \mathcal{A}$, a necessary and sufficient condition for $p_* = P_\Pi(p)$ is $p - p_* \in \mathcal{N}_\Pi^A(p_*)$. This, however, does not ensure the existence of a $p \in \mathcal{A}$ satisfying $p = p_* + \tau v$ for a $\tau > 0$ for all $v \in \mathcal{N}_\Pi^A(p_*)$ because v might not be a feasible direction at p_* for \mathcal{A} , i.e., v might not be an element of $\text{cone}(\mathcal{A} - p_*)$. If $\mathcal{T}_A(p_*) = \text{cone}(\mathcal{A} - p_*)$ for a $p_* \in \Pi$, then for all $v \in \mathcal{N}_\Pi^A(p_*)$, there exists $\tau > 0$ satisfying $p_* + \tau v \in \mathcal{A}$. The polyhedral convexity discussed in the following ensures that $\mathcal{T}_A(p_*) = \text{cone}(\mathcal{A} - p_*)$ for all $p_* \in \mathcal{A}$.

Let us define \mathcal{N}_Π^A as the union of all $\mathcal{N}_\Pi^A(p_*)$'s for $p_* \in \Pi$ defined in (24), i.e.,

$$\mathcal{N}_\Pi^A := \bigcup_{p_* \in \Pi} \mathcal{N}_\Pi^A(p_*). \quad (25)$$

A. Polyhedral Convexity And The Minimum Angle

Any closed convex set in \mathbb{R}^n can be expressed as the intersection closed half spaces, see [15, §A.4.2.b]; when this description can be done with a finitely many half spaces they are called *polyhedral*. In other words, a closed convex set $\mathcal{A} \subset \mathbb{R}^n$ is *polyhedral* iff there exists a finite index set $\mathcal{J}_{\mathcal{A}}$, vectors $\{f_i \in \mathbb{R}^n\}_{i \in \mathcal{J}_{\mathcal{A}}}$, and constants $\{b_i \in \mathbb{R}\}_{i \in \mathcal{J}_{\mathcal{A}}}$ such that

$$\mathcal{A} = \{p \in \mathbb{R}^n : \langle f_i, p \rangle \leq b_i \quad \forall i \in \mathcal{J}_{\mathcal{A}}\}. \quad (26)$$

We denote the set of active constraints at p_* by $\mathcal{J}_{\mathcal{A}}(p_*)$, i.e.,

$$\mathcal{J}_{\mathcal{A}}(p_*) := \{i \in \mathcal{J}_{\mathcal{A}} : \langle f_i, p_* \rangle = b_i\} \quad \forall p_* \in \mathcal{A}. \quad (27)$$

Then the tangent cone and the normal cone at any $p_* \in \mathcal{A}$ can be characterized via $\mathcal{J}_{\mathcal{A}}(p_*)$ as follows, see [15, p. 67],

$$\mathcal{T}_{\mathcal{A}}(p_*) = \{p \in \mathbb{R}^n : \langle f_i, p \rangle \leq 0 \quad \forall i \in \mathcal{J}_{\mathcal{A}}(p_*)\}, \quad (28)$$

$$\mathcal{N}_{\mathcal{A}}(p_*) = \text{cone}(\{f_i : i \in \mathcal{J}_{\mathcal{A}}(p_*)\}). \quad (29)$$

Thus both $\mathcal{T}_{\mathcal{A}}(p_*)$ and $\mathcal{N}_{\mathcal{A}}(p_*)$ are closed convex polyhedral sets, as well.

\mathcal{S} is an affine subspace iff there exists a finite index set $\mathcal{J}_{\mathcal{S}}$, vectors $\{f_i\}_{i \in \mathcal{J}_{\mathcal{S}}}$, and constants $\{b_i\}_{i \in \mathcal{J}_{\mathcal{S}}}$ such that

$$\mathcal{S} = \{p \in \mathbb{R}^n : \langle f_i, p \rangle = b_i \quad \forall i \in \mathcal{J}_{\mathcal{S}}\}. \quad (30)$$

Thus an affine subspace \mathcal{S} can be interpreted as a closed convex polyhedral set for which all constraints are active at all points $p_* \in \mathcal{S}$. Hence, the tangent cone and the normal cone will not change from one point of \mathcal{S} to the next and they can be denoted by $\mathcal{T}_{\mathcal{S}}$ and $\mathcal{N}_{\mathcal{S}}$ instead of $\mathcal{T}_{\mathcal{S}}(p_*)$ and $\mathcal{N}_{\mathcal{S}}(p_*)$. If \mathcal{S} is non-empty then $\mathcal{T}_{\mathcal{S}}$ and $\mathcal{N}_{\mathcal{S}}$ are

$$\mathcal{T}_{\mathcal{S}} = \{p \in \mathbb{R}^n : \langle f_i, p \rangle = 0 \quad \forall i \in \mathcal{J}_{\mathcal{S}}\}, \quad (31)$$

$$\mathcal{N}_{\mathcal{S}} = \text{span}(\{f_i : i \in \mathcal{J}_{\mathcal{S}}\}), \quad (32)$$

where $\text{span}(\{f_i : i \in \mathcal{J}_{\mathcal{S}}\})$ is the subspace spanned by f_i vectors for $i \in \mathcal{J}_{\mathcal{S}}$.

Lemma 1. *Let \mathcal{A} be a closed convex polyhedral subset of \mathbb{R}^n , \mathcal{S} be an affine subspace, Π be their intersection, and θ be the minimum angle between $\mathcal{T}_{\mathcal{S}}$ and $\mathcal{N}_{\Pi}^{\mathcal{A}}$, i.e.,*

$$\Pi := \mathcal{A} \cap \mathcal{S}, \quad (33)$$

$$\theta := \begin{cases} \frac{\pi}{2} & \mathcal{N}_{\Pi}^{\mathcal{A}} = \{0\} \\ \inf_{v \in \mathcal{N}_{\Pi}^{\mathcal{A}} : \|v\|=1} \arccos \|P_{\mathcal{T}_{\mathcal{S}}}(v)\| & \mathcal{N}_{\Pi}^{\mathcal{A}} \neq \{0\} \end{cases}. \quad (34)$$

Then θ is a positive angle that is equal to the minimum angle between $\mathcal{T}_{\mathcal{S}}$ and $\mathcal{N}_{\Pi}^{\mathcal{A}}(p_*)$ for some $p_* \in \Pi$, which is uniquely determined by the active constraints at p_* for \mathcal{A} and \mathcal{S} , i.e. by $\{f_i\}_{i \in \mathcal{J}_{\mathcal{A}}(p_*)}$ and $\{f_i\}_{i \in \mathcal{J}_{\mathcal{S}}}$. Furthermore,

$$\mathcal{T}_{\Pi}(p_*) = \mathcal{T}_{\mathcal{A}}(p_*) \cap \mathcal{T}_{\mathcal{S}} \quad \forall p_* \in \Pi, \quad (35)$$

$$\mathcal{N}_{\Pi}(p_*) = \mathcal{N}_{\mathcal{A}}(p_*) + \mathcal{N}_{\mathcal{S}} \quad \forall p_* \in \Pi, \quad (36)$$

$$\mathcal{N}_{\Pi}^{\mathcal{A}}(p_*) \cap \mathcal{T}_{\mathcal{S}} = \{0\} \quad \forall p_* \in \Pi. \quad (37)$$

Proof of Lemma 1. Note that Π is a closed convex polyhedral set because any affine subspace is a closed convex polyhedral set and intersection of two closed convex polyhedral sets is again a closed convex polyhedral set. Furthermore,

$$\mathcal{J}_{\Pi}(p_*) = \mathcal{J}_{\mathcal{A}}(p_*) \cup \mathcal{J}_{\mathcal{S}}(p_*) \quad \forall p_* \in \Pi. \quad (38)$$

(35) follows from (28), (31), and (38). The identity in (36) follows from (29), (32), and (38). Furthermore, (37) follows from (24) and (35) because $\mathcal{T}_{\Pi}(p_*) \cap \mathcal{N}_{\Pi}^{\mathcal{A}}(p_*) = \{0\}$ by (21).

Using (25), we can express θ defined in (34) as

$$\theta = \inf_{p_* \in \Pi} \theta(p_*), \quad (39)$$

where $\theta(p_*)$ is the minimum angle between $\mathcal{T}_{\mathcal{S}}$ and $\mathcal{N}_{\Pi}^{\mathcal{A}}(p_*)$, i.e.

$$\theta(p_*) := \begin{cases} \frac{\pi}{2} & \mathcal{N}_{\Pi}^{\mathcal{A}}(p_*) = \{0\} \\ \inf_{v \in \mathcal{N}_{\Pi}^{\mathcal{A}}(p_*) : \|v\|=1} \arccos \|P_{\mathcal{T}_{\mathcal{S}}}(v)\| & \mathcal{N}_{\Pi}^{\mathcal{A}}(p_*) \neq \{0\} \end{cases}. \quad (40)$$

Let us proceed with establishing the positivity of $\theta(p_*)$. If $\mathcal{N}_{\Pi}^{\mathcal{A}}(p_*) = \{0\}$, then $\theta(p_*) = \frac{\pi}{2}$; else $v = P_{\mathcal{T}_{\mathcal{S}}}(v) + P_{\mathcal{N}_{\mathcal{S}}}(v)$ and $\|P_{\mathcal{N}_{\mathcal{S}}}(v)\| \neq 0$ for any $v \in \mathcal{N}_{\Pi}^{\mathcal{A}}(p_*)$ satisfying $\|v\| > 0$, because $v \notin \mathcal{T}_{\mathcal{S}}$ by (37). Thus $\|P_{\mathcal{T}_{\mathcal{S}}}(v)\| < \|v\|$ whenever $v \in \mathcal{N}_{\Pi}^{\mathcal{A}}(p_*)$ and $\|v\| > 0$. On the other hand,

$$\sup_{v \in \mathcal{N}_{\Pi}^{\mathcal{A}}(p_*) : \|v\|=1} \|P_{\mathcal{T}_{\mathcal{S}}}(v)\| = \max_{v \in \mathcal{N}_{\Pi}^{\mathcal{A}}(p_*) : \|v\|=1} \|P_{\mathcal{T}_{\mathcal{S}}}(v)\|,$$

by the extreme value theorem because norm and projection are continuous and supremum is over a compact set. Thus $\theta(p_*) > 0$ for all $p_* \in \Pi$.

There are only finitely many distinct possible $\mathcal{T}_{\mathcal{A}}(p_*)$ sets for $p_* \in \mathcal{A}$ and finitely many distinct possible $\mathcal{N}_{\Pi}^{\mathcal{A}}(p_*)$ sets for $p_* \in \Pi$ because both \mathcal{A} and Π are polyhedral. Thus there are only finitely many distinct $\theta(p_*)$ values for $p_* \in \Pi$. Consequently the infimum in (39) is a minimum, θ is positive, and $\theta = \theta(p_*)$ for some $p_* \in \Pi$. On the other hand the angle $\theta(p_*)$, defined in (40), is uniquely determined by the active constraints at p_* for \mathcal{A} and \mathcal{S} , i.e. by $\{f_i\}_{i \in \mathcal{J}_{\mathcal{A}}(p_*)}$ and $\{f_i\}_{i \in \mathcal{J}_{\mathcal{S}}}$. \square

B. Moreau's Decomposition Theorem: Projection To A Closed Convex Cone and Its Polar

A linear subspace \mathcal{S} of \mathbb{R}^n and the linear subspace \mathcal{S}^{\perp} defines an orthogonal decomposition for vectors in \mathbb{R}^n . The closed convex cones and their polar cones enjoy an analogous property commonly known as Moreau's decomposition theorem.

Lemma 2 ([15, Theorem A.3.2.5]). *Let \mathcal{C} be a closed convex cone. For the three elements v , v_1 , and v_2 in \mathbb{R}^n , the properties below are equivalent:*

- (i) $v = v_1 + v_2$ with $v_1 \in \mathcal{C}$, $v_2 \in \mathcal{C}^{\circ}$, and $\langle v_1, v_2 \rangle = 0$;
- (ii) $v_1 = P_{\mathcal{C}}(v)$ and $v_2 = P_{\mathcal{C}^{\circ}}(v)$.

IV. A SIMPLE AND GENERAL PROOF VIA PINSKER'S INEQUALITY

In this section we will establish the quadratic decay of the mutual information on small neighborhoods of the capacity-achieving input distributions using Pinsker's inequality given in (3) together with Lemma 1. For any $p \in \mathcal{A}$ and $p_* \in \Pi_{\mathcal{A}}$, we can bound $I(p; W)$ from above using (3), (8), and (13),

$$I(p; W) \leq C_{\mathcal{A}} + \langle p - p_*, D(W \| q_{\mathcal{A}}) \rangle - \frac{1}{2} \|q_{p-p_*}\|_1^2. \quad (41)$$

If p_* is the projection of p to $\Pi_{\mathcal{A}}$ then $p - p_*$ is in $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$ by (23). If \mathcal{X} is a finite set and \mathcal{A} is polyhedral, then $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$ is a closed convex polyhedral cone for each $p_* \in \Pi_{\mathcal{A}}$ because both $\mathcal{T}_{\mathcal{A}}(p_*)$ and $\mathcal{N}_{\Pi_{\mathcal{A}}}(p_*)$ are closed convex polyhedral cones for each $p_* \in \Pi_{\mathcal{A}}$ and $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$ is their intersection. On the other hand, there are only finitely many distinct $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$'s for $p_* \in \Pi_{\mathcal{A}}$ because there are only finitely many distinct $\mathcal{T}_{\mathcal{A}}(p_*)$'s and $\mathcal{N}_{\Pi_{\mathcal{A}}}(p_*)$'s for $p_* \in \Pi_{\mathcal{A}}$. Thus $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}$, defined as the union $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$'s for $p_* \in \Pi_{\mathcal{A}}$ in (25), is a closed cone, as well. Since $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}$ is closed there is a minimum angle between vectors in $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}$ and $\mathcal{T}_{\mathcal{S}_{\mathcal{A}}}$ which is equal to the subspace of the intersection of the kernels of the last two terms in (41). This minimum angle has to be positive because otherwise one could move from p_* in this direction and stay in $\Pi_{\mathcal{A}}$ to reach a point that is closer to p as a result of (13), for some p . Thus the norm of $\mathbb{P}_{\mathcal{T}_{\mathcal{S}_{\mathcal{A}}}}(p - p_*)$ is guaranteed to be larger than a certain fraction of $\|p - p_*\|$. This observation, which essentially is what is established in Lemma 1, is at the core of bound presented in Theorem 1.

We will need the following bound on $\|q_v\|$ in terms of $\|v\|$ to obtain explicit approximation error terms.

$$\begin{aligned} \|q_v\|_1 &= \left\| \sum_x v(x) W(x) \right\|_1 \\ &\leq \sum_x \|v(x) W(x)\|_1 \\ &= \sum_x |v(x)| \|W(x)\|_1 \\ &= \|v\|_1 \\ &\leq \|v\| \cdot \sqrt{|\mathcal{X}|} \quad \forall v \in \mathbb{R}^{\mathcal{X}}, \end{aligned} \quad (42)$$

where the first inequality follows from the triangle inequality, and the second inequality follows from the general upper bound on the ℓ^1 norm in terms of the ℓ^2 norm for $\mathbb{R}^{\mathcal{X}}$.

Theorem 1. *Let $W : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$ be a channel with a finite input set \mathcal{X} and $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$ be a closed convex polyhedral constraint set, i.e., a constraint set that can be characterized by a finite number of linear constraints, then*

$$I(p; W) \leq C_{\mathcal{A}} - \gamma \|p - \mathbb{P}_{\Pi_{\mathcal{A}}}(p)\|^2 \quad \forall p \in \Pi_{\mathcal{A}}^{\delta}, \quad (43)$$

for the set $\Pi_{\mathcal{A}}^{\delta}$ defined in (18) and positive constants β , γ , and δ defined as follows

$$\beta := \begin{cases} \frac{\pi}{2} & \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}} = \{0\} \\ \inf_{v \in \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}} : \|v\|=1} \arccos \left\| \mathbb{P}_{\mathcal{T}_{\mathcal{S}_{\mathcal{A}}}}(v) \right\| & \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}} \neq \{0\}, \end{cases} \quad (44a)$$

$$\gamma := \frac{\sin^2 \beta}{2} \inf_{v \in \mathcal{K}_{\mathcal{A}}^d \cap \mathcal{N}_{\mathcal{S}_{\mathcal{A}}} : \|v\|=1} \|q_v\|_1^2, \quad (44b)$$

$$\delta := \left(|\mathcal{X}| + \frac{\gamma}{\sin^2 \beta} \right)^{-1} \|D(W \| q_{\mathcal{A}})\|. \quad (44c)$$

Proof of Theorem 1. Since (16) holds for any closed convex constraint set \mathcal{A} , affine subspace $\mathcal{S}_{\mathcal{A}}$ defined in (17), and $\Pi_{\mathcal{A}}$ defined in (11), the hypotheses of Lemma 1 holds for $(\mathcal{A}, \mathcal{S}, \Pi) \rightarrow (\mathcal{A}, \mathcal{S}_{\mathcal{A}}, \Pi_{\mathcal{A}})$. Thus β defined in (44a) (i.e., the minimum angle between $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}$ and $\mathcal{T}_{\mathcal{S}_{\mathcal{A}}}$), is positive because \mathcal{A}

is polyhedral (i.e., \mathcal{A} is determined by finite number of linear constraints). Consequently,

$$\left\| \mathbb{P}_{\mathcal{T}_{\mathcal{S}_{\mathcal{A}}}}(v) \right\| \leq \|v\| \cos \beta \quad \forall v \in \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}, \quad (45)$$

Let $v \in \mathbb{R}^{\mathcal{X}}$ be $p - \mathbb{P}_{\Pi_{\mathcal{A}}}(p)$, and v_1, v_2, v_3 be its projections to the orthogonal subspace $\mathcal{T}_{\mathcal{S}_{\mathcal{A}}}$, $\mathcal{K}_{\mathcal{A}}^d \cap \mathcal{N}_{\mathcal{S}_{\mathcal{A}}}$, and $\{\tau D(W \| q_{\mathcal{A}}) : \tau \in \mathbb{R}\}$:

$$v := p - \mathbb{P}_{\Pi_{\mathcal{A}}}(p), \quad (46a)$$

$$v_1 := \mathbb{P}_{\mathcal{T}_{\mathcal{S}_{\mathcal{A}}}}(v), \quad (46b)$$

$$v_2 := \mathbb{P}_{\mathcal{K}_{\mathcal{A}}^d \cap \mathcal{N}_{\mathcal{S}_{\mathcal{A}}}}(v) \quad (46c)$$

$$v_3 := \frac{\langle v, D(W \| q_{\mathcal{A}}) \rangle}{\|D(W \| q_{\mathcal{A}})\|^2} D(W \| q_{\mathcal{A}}). \quad (46d)$$

Note that $\text{span}(\mathcal{T}_{\mathcal{S}_{\mathcal{A}}}, \mathcal{K}_{\mathcal{A}}^d \cap \mathcal{N}_{\mathcal{S}_{\mathcal{A}}}, D(W \| q_{\mathcal{A}})) = \mathbb{R}^{\mathcal{X}}$. Thus

$$v = v_1 + v_2 + v_3. \quad (47)$$

Thus the upper bound on $I(p; W)$ for any $p \in \mathcal{A}$ in (41) is

$$I(p; W) \leq C_{\mathcal{A}} + \langle v, D(W \| q_{\mathcal{A}}) \rangle - \frac{1}{2} \|q_v\|_1^2. \quad (48)$$

Let us proceed with bounding the terms in (48). Note that the sign of the inner product $\langle v, D(W \| q_{\mathcal{A}}) \rangle$ cannot be positive because otherwise (12) would be violated. Thus

$$\begin{aligned} \langle v, D(W \| q_{\mathcal{A}}) \rangle &= \langle v_3, D(W \| q_{\mathcal{A}}) \rangle \\ &= -\|v_3\| \|D(W \| q_{\mathcal{A}})\|. \end{aligned} \quad (49)$$

On the other hand,

$$\begin{aligned} \|q_v\|_1^2 &= \|q_{v_2} + q_{v_3}\|_1^2 \\ &\stackrel{(a)}{\geq} (\|q_{v_2}\|_1 - \|q_{v_3}\|_1)^2 \\ &\geq \|q_{v_2}\|_1^2 - 2\|q_{v_2}\|_1 \cdot \|q_{v_3}\|_1 \\ &\stackrel{(b)}{\geq} \|q_{v_2}\|_1^2 - 2|\mathcal{X}| \cdot \|v_2\| \cdot \|v_3\| \\ &\stackrel{(c)}{\geq} \frac{2\gamma}{\sin^2 \beta} \|v_2\|^2 - 2|\mathcal{X}| \cdot \|v_2\| \cdot \|v_3\| \\ &= \frac{2\gamma}{\sin^2 \beta} \|v_2 + v_3\|^2 - 2 \left(\frac{\gamma \|v_3\|}{\sin^2 \beta} + |\mathcal{X}| \|v_2\| \right) \|v_3\| \\ &\stackrel{(d)}{\geq} \frac{2\gamma}{\sin^2 \beta} \|v_2 + v_3\|^2 - 2\|v\| \frac{\|D(W \| q_{\mathcal{A}})\|}{\delta} \|v_3\| \\ &\stackrel{(e)}{\geq} 2\gamma \|v\|^2 - 2\|v\| \frac{\|D(W \| q_{\mathcal{A}})\|}{\delta} \|v_3\|, \end{aligned} \quad (50)$$

where (a) follows from the triangle inequality, (b) follows from (42), (c) follows from the definition of γ given in (44b), (d) follows from (44c) and $\|v_2\| \vee \|v_3\| \leq \|v\|$, and (e) follows from (45), which implies $\|v_2 + v_3\| \geq \|v\| \sin \beta$.

(43) holds for all $p \in \Pi_{\mathcal{A}}^{\delta}$ as a result of (48), (49), and (50).

We are left with establishing the positivity of γ . Note that $\{v \in \mathcal{K}_{\mathcal{A}}^d \cap \mathcal{N}_{\mathcal{S}_{\mathcal{A}}} : \|v\| = 1\}$ is a closed and bounded set, i.e., a compact set, thus the infimum in the definition of γ given in (44b) is a minimum, i.e., it is achieved by some v_* . If the minimum value in (44b) is zero then $v_* \in \mathcal{K}_W$ by (8) and (15); on the other hand $v_* \in \mathcal{K}_{\mathcal{A}}^d$, for $\mathcal{K}_{\mathcal{A}}^d$ defined in (14), by hypothesis. Thus $v_* \in \mathcal{T}_{\mathcal{S}_{\mathcal{A}}}$ because

$$\mathcal{T}_{\mathcal{S}_{\mathcal{A}}} = \mathcal{K}_{\mathcal{A}}^d \cap \mathcal{K}_W, \quad (51)$$

as a result of (17) and (31). This, however, is a contradiction because $v_* \in \mathcal{N}_{\mathcal{S}_{\mathcal{A}}}$ by hypothesis. Hence γ is positive. \square

Theorem 1 assumes \mathcal{A} to be polyhedral and input set \mathcal{X} to be finite; both of these assumptions are necessary to establish a quadratic bound on the worst case decrease of the mutual information with the Euclidean distance to $\Pi_{\mathcal{A}}$. Example 1 in the following describes a channel with a finite input set and a convex constraint set \mathcal{A} that is not polyhedral for which the decrease of the mutual information with the distance to $\Pi_{\mathcal{A}}$ is proportional to the fourth power of the distance to $\Pi_{\mathcal{A}}$, which is much slower. Example 2 describes a channel with countably infinite input set and finite output set for which, if

$$I(p; W) \leq C_W - f(\|p - P_{\Pi}(p)\|) \quad \forall p \in \Pi^\delta, \quad (52)$$

for some $f: \mathbb{R} \rightarrow \mathbb{R}$, then $f(z) = 0$ for all $z \in [0, \delta]$.

Example 1. Let the channel W with three input letters and two output letters and convex constraint set \mathcal{A} be

$$W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad \mathcal{A} = \left\{ p \in \mathcal{P}(\mathcal{X}) : \|p - s\| \leq \frac{1}{2\sqrt{6}} \right\}$$

$$s = \left[\frac{2}{3} \quad \frac{1}{6} \quad \frac{1}{6} \right]^T$$

Then $C_{\mathcal{A}} = \ln 2$, $\Pi_{\mathcal{A}} = \left\{ \left[\frac{1}{2} \quad \frac{1}{4} \quad \frac{1}{4} \right]^T \right\}$, and $q_{\mathcal{A}} = \left[\frac{1}{2} \quad \frac{1}{2} \right]^T$. Furthermore, the boundary \mathcal{A} can be described parametrically as follows $\partial\mathcal{A} = \{p_\tau : \tau \in (-\pi, \pi)\}$, where the parametric family of input distributions p_τ is

$$p_\tau = \frac{1}{6} \begin{bmatrix} 4 - \cos \tau \\ 1 + \sin(\frac{\pi}{6} + \tau) \\ 1 + \sin(\frac{\pi}{6} - \tau) \end{bmatrix} = s + \frac{1}{12} \begin{bmatrix} -2 \cos \tau \\ \cos \tau + \sqrt{3} \sin \tau \\ \cos \tau - \sqrt{3} \sin \tau \end{bmatrix}.$$

One can confirm by substitution the following closed-form expressions for $\|p_\tau - p_0\|$ and $I(p_\tau; W)$

$$I(p_\tau; W) = \ln 2 - D(q_{p_\tau} \| q_{p_0}), \quad q_{p_\tau} = \frac{1}{6} \begin{bmatrix} 3 + 2 \sin^2 \frac{\tau}{2} \\ 3 - 2 \sin^2 \frac{\tau}{2} \end{bmatrix},$$

$$\|p_\tau - p_0\| = \frac{1}{\sqrt{6}} \left| \sin \frac{\tau}{2} \right| \quad \Pi_{\mathcal{A}} = \{p_0\}.$$

Thus

$$\lim_{\tau \downarrow 0} \frac{C_{\mathcal{A}} - I(p_\tau; W)}{\|p_\tau - P_{\Pi_{\mathcal{A}}}(p_\tau)\|^4} = 8$$

Hence, the decrease of mutual information with the distance from $\Pi_{\mathcal{A}}$ is proportional with fourth power of the distance, rather than the second power for the points on $\partial\mathcal{A}$, i.e., on the boundary of \mathcal{A} . Thus (1) does not hold for any positive constants γ and δ .

Example 2. Let us consider the channel W whose input set is the set of all non-zero integers and whose output set is has only two elements.

$$W(1) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad W(-1) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$$W(x) = \frac{1}{2} \begin{bmatrix} 1 + \tanh x \\ 1 - \tanh x \end{bmatrix} \quad \text{for } x \notin \{-1, 1\},$$

Then $C_W = \ln 2$ and $\Pi = \{p_1\}$ where p_i is the uniform distribution on the input letter i and $-i$ for all $i \in \mathbb{Z}_+$. Then $\|p_i - P_{\Pi}(p_i)\| = 1$ for all $i \in \mathbb{Z}_+$ and $I(p_i; W) \uparrow C_W$. Thus as a result of the concavity of mutual information in the input distribution, and Jensen's inequality, we have

$$I((1 - \tau)p_1 + \tau p_i; W) \geq (1 - \tau)C_W + \tau I(p_i; W).$$

On the other hand,

$$\|(1 - \tau)p_1 + \tau p_i - P_{\Pi}((1 - \tau)p_1 + \tau p_i)\| = \tau$$

because $\Pi = \{p_1\}$. Thus (52) holds for a $\delta > 0$ iff $f(z) = 0$ for all $z \in [0, \delta]$.

V. EXACT CHARACTERIZATION

The positivity of the minimum angle between $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}$, i.e., the cone of directions that one can move away from $\Pi_{\mathcal{A}}$, and the subspace of the intersection of the kernel of the channel and the kernel of the gradient of mutual information, i.e., β defined in (44a), is sufficient to establish the quadratic decrease of the mutual information in directions pointing away from $\Pi_{\mathcal{A}}$. One can even determine whether the slowest decay is linear or quadratic in the distance to $\Pi_{\mathcal{A}}$ using the extreme value theorem and the fact that $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}$ is closed. To determine the tightest coefficient in quadratic decrease case, however, the minimum angle idea by itself is not sufficient; projections to closed convex cones via Moreau's decomposition theorem rather than projections to subspaces needs to be considered. Recall that $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$ is a closed convex polyhedral cone for each $p_* \in \Pi_{\mathcal{A}}$, whenever \mathcal{X} is a finite set and \mathcal{A} is polyhedral. As we have discussed in §IV, $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}$ defined in (25) as the union of $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$'s for $p_* \in \Pi_{\mathcal{A}}$, is a closed cone. However, $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}$ is not necessarily convex. Hence we can apply Moreau's decomposition theorem, i.e. Lemma 2, to each $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$ separately but not necessarily to $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}$ itself.

We will use the minimum angle idea by invoking Lemma 1 in our analysis in this section too, though in a more nuanced manner. Let $\Upsilon(p_*)$ be

$$\Upsilon(p_*) := \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*) \cap \mathcal{K}_{\mathcal{A}}^d \quad \forall p_* \in \Pi_{\mathcal{A}}. \quad (53)$$

Let us assume that $\Upsilon(p_*) \neq \{0\}$ for a $p_* \in \Pi_{\mathcal{A}}$. Then any $v \in \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$ can be decomposed into two orthogonal components: v_* in $\Upsilon(p_*)$, and $v - v_*$ in $\Upsilon(p_*)^\circ$ by Lemma 2 because $\Upsilon(p_*)$ is a closed convex cone. Applying Lemma 1 for the case $(\mathcal{A}, \mathcal{S}, \Pi) \rightarrow (\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*), \mathcal{K}_{\mathcal{A}}^d, \Upsilon(p_*))$, we can assert the positivity of the minimum angle $\phi(p_*)$ between $v - v_*$ and $\mathcal{T}_{\mathcal{K}_{\mathcal{A}}^d}$ for all $v \in \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$, whenever $\Upsilon(p_*) \neq \{0\}$. Thus using the fact that $\mathcal{T}_{\mathcal{K}_{\mathcal{A}}^d} = \mathcal{K}_{\mathcal{A}}^d$ we can conclude that

$$\left\| P_{\mathcal{K}_{\mathcal{A}}^d}(v - v_*) \right\| \leq \|v - v_*\| \cos(\phi(p_*)) \quad \forall v \in \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*) \quad (54)$$

where $v_* = P_{\Upsilon(p_*)}(v)$, and $\phi(p_*)$ is defined as

$$\phi(p_*) := \begin{cases} \frac{\pi}{2} & \text{if } \mathcal{B}(p_*) = \{0\} \\ \min_{u \in \mathcal{B}(p_*): \|u\|=1} \arccos \left\| P_{\mathcal{K}_{\mathcal{A}}^d}(u) \right\| & \text{if } \mathcal{B}(p_*) \neq \{0\} \end{cases} \quad (55)$$

for $\mathcal{B}(p_*) := \mathcal{N}_{\Upsilon(p_*)}^{\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)}$ whenever $\Upsilon(p_*) \neq \{0\}$.

As we did in §IV, we will establish (1) by invoking (13) first. Instead of bounding $D(q_p \| q_{\mathcal{A}})$ using the Pinsker's inequality given in (3), however, we will use (6) together with Hilbert spaces structure induced by the existence and uniqueness of the Shannon center on the minimal affine subspace of \mathcal{A} .

A. A Taylor Series Expansion Of Kullback–Leibler Divergence

For any channel $W : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$ and convex constraint set $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$ satisfying $C_{\mathcal{A}} < \infty$, let us define the subset of the input set $\mathcal{X}_{\mathcal{A}}$ and the extended real valued function $\Lambda_{\mathcal{A}} : \mathcal{X}_{\mathcal{A}} \times \mathcal{X}_{\mathcal{A}} \rightarrow [0, \infty]$ as

$$\mathcal{X}_{\mathcal{A}} := \{x \in \mathcal{X} : \exists p \in \mathcal{A} \text{ s.t. } p(x) > 0\}, \quad (56)$$

$$\Lambda_{\mathcal{A}}(x, z) := \int \left(\frac{dW(x)}{dq_{\mathcal{A}}} \right) \left(\frac{dW(z)}{dq_{\mathcal{A}}} \right) dq_{\mathcal{A}} \quad \forall x, z \in \mathcal{X}_{\mathcal{A}}. \quad (57)$$

(2) and (12) imply $W(x) \prec_{q_{\mathcal{A}}}$, and hence the existence of the Radon-Nikodym derivative $\frac{dW(x)}{dq_{\mathcal{A}}}$ for all $x \in \mathcal{X}_{\mathcal{A}}$. This, however, does not imply the finiteness of $\Lambda_{\mathcal{A}}(x, z)$ as demonstrated by Example 3 for a channel with a finite input set and countable output set. This is possible because finiteness of $C_{\mathcal{A}}$ implies the finiteness of $D(q_p \| q_{\mathcal{A}})$ for all $p \in \mathcal{A}$ via (3), (9), and (12), but not the finiteness of $\chi^2(q_p \| q_{\mathcal{A}})$ for all $p \in \mathcal{A}$ and for all $p \in \mathcal{P}(\mathcal{X}_{\mathcal{A}})$ we have

$$\sum_{x, z} p(x) \Lambda_{\mathcal{A}}(x, z) p(z) = 1 + \chi^2(q_p \| q_{\mathcal{A}}). \quad (58)$$

When $\mathcal{X}_{\mathcal{A}}$ is a finite set and $\max_{x, z} \Lambda_{\mathcal{A}}(x, z)$ is finite, then $\Lambda_{\mathcal{A}}$ is a positive semi-definite matrix because

$$v^T \Lambda_{\mathcal{A}} v = \int \left(\frac{dq_v}{dq_{\mathcal{A}}} \right)^2 dq_{\mathcal{A}} \quad \forall v \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}.$$

Thus $\Lambda_{\mathcal{A}}$ defines an inner product on $\mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$. Furthermore, the resulting norm is related to the χ^2 divergence as follows:

$$\chi^2(q_p \| q_{\mathcal{A}}) = \|p - p_*\|_{\Lambda_{\mathcal{A}}}^2 \quad (59)$$

for all $p \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ and $p_* \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ satisfying $q_{p_*} = q_{\mathcal{A}}$ for q_{p_*} is defined in (8).

On the other hand, for all $p \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ and $p_* \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ satisfying $q_{p_*} = q_{\mathcal{A}}$, the Cauchy–Schwarz inequality implies

$$\begin{aligned} \left| \frac{dq_p}{dq_{\mathcal{A}}} - 1 \right| &= \left| \sum_x (p(x) - p_*(x)) \frac{dW(x)}{dq_{\mathcal{A}}} \right| \\ &\leq \|p - p_*\| \sqrt{\sum_{x \in \mathcal{X}_{\mathcal{A}}} \left(\frac{dW(x)}{dq_{\mathcal{A}}} \right)^2}. \end{aligned}$$

Thus

$$\chi^3(q_p \| q_{\mathcal{A}}) \leq \|p - p_*\|^3 (\kappa_{\mathcal{A}})^3 \quad (60)$$

for all $p \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ and $p_* \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$, where $\kappa_{\mathcal{A}}$ is defined as follows

$$\kappa_{\mathcal{A}} := \sqrt[3]{\int \left(\sum_{x \in \mathcal{X}_{\mathcal{A}}} \left(\frac{dW(x)}{dq_{\mathcal{A}}} \right)^2 \right)^{3/2} dq_{\mathcal{A}}}. \quad (61)$$

Applying (6) for $w = q_p$ and $q = q_{\mathcal{A}}$, and invoking (59) and (60), we get the following lemma.

Lemma 3. For any $W : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$ with a finite input set \mathcal{X} and convex constraint set $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$ satisfying $\kappa_{\mathcal{A}} < \infty$, we have

$$D(q_p \| q_{\mathcal{A}}) \geq \frac{1}{2} \|p - p_*\|_{\Lambda_{\mathcal{A}}}^2 - \frac{\kappa_{\mathcal{A}}^3}{6} \|p - p_*\|^3, \quad (62a)$$

$$D(q_p \| q_{\mathcal{A}}) \leq \frac{1}{2} \|p - p_*\|_{\Lambda_{\mathcal{A}}}^2 + \frac{\kappa_{\mathcal{A}}^3}{2} \|p - p_*\|^3. \quad (62b)$$

Using Jensen’s inequality and the convexity of function $z^{3/2}$ in z we can bound $\kappa_{\mathcal{A}}$ from below by $\sqrt{\text{Tr}[\Lambda_{\mathcal{A}}]}$, where $\text{Tr}[\Lambda_{\mathcal{A}}]$ is the trace of the $|\mathcal{X}_{\mathcal{A}}|$ -by- $|\mathcal{X}_{\mathcal{A}}|$ matrix $\Lambda_{\mathcal{A}}$. On the

other hand, we can bound $\kappa_{\mathcal{A}}$ from above using the general bound on ℓ^2 in terms of ℓ^3 norm, i.e. $\|v\|_2 \leq |\mathcal{X}_{\mathcal{A}}|^{1/6} \|v\|_3$. Thus

$$\sqrt[3]{\sqrt{|\mathcal{X}_{\mathcal{A}}|} \sum_{x \in \mathcal{X}_{\mathcal{A}}} \int \left(\frac{dW(x)}{dq_{\mathcal{A}}} \right)^3 dq_{\mathcal{A}}} \geq \kappa_{\mathcal{A}} \geq \sqrt{\text{Tr}[\Lambda_{\mathcal{A}}]}. \quad (63)$$

For channels with finite input and output sets $\kappa_{\mathcal{A}} < \infty$, i.e., the hypothesis of Lemma 3 is always satisfied. For channels with finite input sets and infinite output sets, however, even $\Lambda_{\mathcal{A}}(x, z)$ can be infinite for some $x, z \in \mathcal{X}_{\mathcal{A}}$.

Example 3. Let the discrete channel $W : \mathcal{X} \rightarrow \mathcal{P}(\mathbb{Z}_+)$ with the finite input set $\mathcal{X} = \{0, 1, \dots, n\}$ be

$$W(y|x) = \begin{cases} \frac{(y-n)^{-2}}{\zeta(2)} \mathbb{1}_{\{y>n\}} & \text{if } x=0 \\ \frac{1}{2} \mathbb{1}_{\{y=x\}} + \frac{(y-n)^{-3}}{2\zeta(3)} \mathbb{1}_{\{y>n\}} & \text{if } x \in \{1, 2, \dots, n\} \end{cases}$$

where $\zeta(s) := \sum_{y \in \mathbb{Z}_+} y^{-s}$, i.e., the Riemann zeta function.

If $n \geq \left(\frac{2\zeta(3)}{\zeta(2)} \right)^2 e^{2 \sum_{y \in \mathbb{Z}_+} \frac{y^{-2}}{\zeta(2)} \ln y}$, then we have

$$C_W = \ln \sqrt{n},$$

$$q_W(y) = \frac{1}{2n} \mathbb{1}_{\{y \leq n\}} + \frac{(y-n)^{-3}}{2\zeta(3)} \mathbb{1}_{\{y > n\}},$$

$$D(W(\cdot|0) \| q_W) = \ln \frac{2\zeta(3)}{\zeta(2)} + \sum_{y \in \mathbb{Z}_+} \frac{y^{-2}}{\zeta(2)} \ln y \leq C_W.$$

The diagonal entry of the matrix Λ_W corresponding to the input letter 0 itself is infinite:

$$\begin{aligned} \Lambda_W(0, 0) &= \sum_{y \in \mathbb{Z}_+} q_W(y) \left(\frac{W(y|0)}{q_W(y)} \right)^2 \\ &\geq \frac{2\zeta(3)}{(\zeta(2))^2} \sum_{y > n} \frac{1}{y-n} \\ &= \infty. \end{aligned}$$

Then κ_W is infinite, as well because $\kappa_W \geq \sqrt{\text{Tr}[\Lambda_W]}$. Thus for this channel Lemma 3 is mute.

In our analysis we will need an operator-norm bound analogous to (42). We bound $\|v\|_{\Lambda_{\mathcal{A}}}$ above by the product of $\|v\|$ and the trace of $\Lambda_{\mathcal{A}}$ using the Cauchy–Schwarz inequality:

$$\begin{aligned} \|v\|_{\Lambda_{\mathcal{A}}}^2 &= \int \left(\sum_x v(x) \frac{dW(x)}{dq_{\mathcal{A}}} \right)^2 dq_{\mathcal{A}} \\ &\leq \int \|v\|^2 \left(\sum_{x \in \mathcal{X}_{\mathcal{A}}} \left(\frac{dW(x)}{dq_{\mathcal{A}}} \right)^2 \right) dq_{\mathcal{A}} \\ &= \|v\|^2 \text{Tr}[\Lambda_{\mathcal{A}}]. \end{aligned} \quad (64)$$

The following discussion is not used in the rest of the paper; nevertheless (67) is just too beautiful to ignore in the name of utilitarianism. Note that (59) and $\chi^2(0 \| q_{\mathcal{A}}) = 1$ implies,

$$\|p_*\|_{\Lambda_{\mathcal{A}}}^2 = 1 \quad (65)$$

for all $p_* \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ satisfying $q_{p_*} = q_{\mathcal{A}}$. On the other hand,

$$\|p\|_{\Lambda_{\mathcal{A}}}^2 = 1 + \chi^2(q_p \| q_{\mathcal{A}}) \quad (66)$$

for all $p \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ satisfying $\langle p, \mathbb{1} \rangle = 1$, where $\mathbb{1} \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ is the vector whose entries are all ones. Equations (59), (65), and (66) gives us the following ‘‘Pythagorean Theorem’’

$$\|p\|_{\Lambda_{\mathcal{A}}}^2 = \|p_*\|_{\Lambda_{\mathcal{A}}}^2 + \|p - p_*\|_{\Lambda_{\mathcal{A}}}^2 \quad (67)$$

for all $p \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ satisfying $\langle p, \mathbb{1} \rangle = 1$, and $p_* \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ satisfying $q_{p_*} = q_{\mathcal{A}}$.

B. Exact Characterization Of The Slowest Decay

Theorem 2. For a channel $W : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$ with a finite input set \mathcal{X} , a closed convex polyhedral constraint set $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$ satisfying $\kappa_{\mathcal{A}} < \infty$, let γ_1 be

$$\gamma_1 := \min_{v \in \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}} : \|v\|=1} -\langle v, D(W \| q_{\mathcal{A}}) \rangle \quad (68)$$

for $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}$ and $\kappa_{\mathcal{A}}$ defined in (25) and (61). Then

$$I(p; W) \leq C_{\mathcal{A}} - \gamma_1 \|v_p\| \quad \forall p \in \mathcal{A}. \quad (69)$$

where $v_p := p - \text{P}_{\Pi_{\mathcal{A}}}(p)$ and there exists a $p \in \mathcal{A} \setminus \Pi_{\mathcal{A}}$ satisfying

$$I(p + \tau v_p; W) \geq C_{\mathcal{A}} - \gamma_1 \|\tau v_p\| - \text{Tr}[A_{\mathcal{A}}] \|\tau v_p\|^2 \quad (70)$$

for all $\tau \in (0, 1)$.

Furthermore, if $\gamma_1 = 0$, then

$$I(p; W) \leq C_{\mathcal{A}} - \gamma_2 \|v_p\|^2 + \frac{\kappa_{\mathcal{A}}^3}{6} \|v_p\|^3 \quad \forall p \in \Pi_{\mathcal{A}}^{\delta} \quad (71)$$

for positive constants γ_2 and δ defined in terms of $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$, $\Upsilon(p_*)$, and $\phi(p_*)$ defined in (24), (53), and (55), as follows

$$\gamma_2 := \frac{1}{2} \min_{v \in \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}} \cap \mathcal{K}_{\mathcal{A}}^d : \|v\|=1} \|v\|_{\mathcal{A}_{\mathcal{A}}}^2, \quad (72a)$$

$$\delta := \min_{p_* \in \Pi_{\mathcal{A}}} \delta_{p_*}, \quad (72b)$$

$$\delta_{p_*} := \begin{cases} \min_{v \in \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*) : \|v\|=1} -\frac{\langle v, D(W \| q_{\mathcal{A}}) \rangle}{\gamma_2} & \text{if } \Upsilon(p_*) = \{0\} \\ \frac{\sin(\phi(p_*))}{\text{Tr}[A_{\mathcal{A}}] + \gamma_2} \|D(W \| q_{\mathcal{A}})\| & \text{if } \Upsilon(p_*) \neq \{0\} \end{cases}, \quad (72c)$$

and there exists a $p \in \mathcal{A} \setminus \Pi_{\mathcal{A}}$ satisfying for all $\tau \in (0, 1)$

$$I(p + \tau v_p; W) \geq C_{\mathcal{A}} - \gamma_2 \|\tau v_p\|^2 - \frac{\kappa_{\mathcal{A}}^3}{2} \|\tau v_p\|^3. \quad (73)$$

Proof of Theorem 2. First note that (68) can be stated as a minimum rather than an infimum by the extreme value theorem because $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}$ is closed and thus the minimization in (68) is that of a continuous function over a closed and bounded (i.e., compact) set.

Note that $v_p \in \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$ by (23) where p_* is the projection of an $p \in \mathcal{A}$ onto $\Pi_{\mathcal{A}}$, i.e., $p_* = \text{P}_{\Pi_{\mathcal{A}}}(p)$. Then (13) and $D(q_p \| q_{\mathcal{A}}) \geq 0$ imply

$$\begin{aligned} I(p; W) &\leq C_{\mathcal{A}} + \langle v_p, D(W \| q_{\mathcal{A}}) \rangle \\ &\leq C_{\mathcal{A}} + \|v_p\| \max_{v \in \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*) : \|v\|=1} \langle v, D(W \| q_{\mathcal{A}}) \rangle, \end{aligned} \quad (74)$$

for all $p \in \mathcal{A}$. Then (69) holds by (25).

Let v^* be a minimizer for minimization defining γ_1 in (68). Then there exists a $p^* \in \Pi_{\mathcal{A}}$ satisfying $v^* \in \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p^*)$ by (25). Furthermore, there exists a $p^* \in \mathcal{A} \setminus \Pi_{\mathcal{A}}$ such that $\text{P}_{\Pi_{\mathcal{A}}}(p^*) = p^*$ by (23) and (24) because $\mathcal{T}_{\mathcal{A}}(p^*) = \text{cone}(\mathcal{A} - p^*)$ as a result of polyhedral convexity of \mathcal{A} . Then

$$\begin{aligned} I(p^*; W) &= C_{\mathcal{A}} + \langle v^*, D(W \| q_{\mathcal{A}}) \rangle - D(q_{p^*} \| q_{\mathcal{A}}) && \text{by (13),} \\ &= C_{\mathcal{A}} - \gamma_1 \|v^*\| - D(q_{p^*} \| q_{\mathcal{A}}) && \text{by (68),} \\ &\geq C_{\mathcal{A}} - \gamma_1 \|v^*\| - \chi^2(q_{p^*} \| q_{\mathcal{A}}) && \text{by (4),} \\ &= C_{\mathcal{A}} - \gamma_1 \|v^*\| - \|v^*\|_{\mathcal{A}_{\mathcal{A}}}^2 && \text{by (59),} \\ &\geq C_{\mathcal{A}} - \gamma_1 \|v^*\| - \text{Tr}[A_{\mathcal{A}}] \|v^*\|^2 && \text{by (64).} \end{aligned}$$

Then (70) holds for $p = p^*$ by Jensen's inequality and the concavity of $I(p; W)$ in p .

Let us proceed with the claims for $\gamma_1 = 0$ case. First note that, $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}} \cap \mathcal{K}_{\mathcal{A}}^d \neq \{0\}$ because $\langle v^*, D(W \| q_{\mathcal{A}}) \rangle = 0$. Since $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}$ and $\mathcal{K}_{\mathcal{A}}^d$ are closed so is $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}} \cap \mathcal{K}_{\mathcal{A}}^d$. Thus (72a) can be stated as a minimum rather than an infimum by the extreme value theorem. Let v^\dagger be the minimizer of (72a). Then there exists a $p^\dagger \in \Pi_{\mathcal{A}}$ satisfying $v^\dagger \in \mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p^\dagger) \cap \mathcal{K}_{\mathcal{A}}^d$ by (25). Furthermore, there exists a $p^\dagger \in \mathcal{A} \setminus \Pi_{\mathcal{A}}$ such that $\text{P}_{\Pi_{\mathcal{A}}}(p^\dagger) = p^*$ by (23) and (24) because $\mathcal{T}_{\mathcal{A}}(p^*) = \text{cone}(\mathcal{A} - p^*)$ as a result of polyhedral convexity of \mathcal{A} . Then

$$\begin{aligned} I(p^\dagger; W) &= C_{\mathcal{A}} + \langle v^\dagger, D(W \| q_{\mathcal{A}}) \rangle - D(q_{p^\dagger} \| q_{\mathcal{A}}) && \text{by (13),} \\ &= C_{\mathcal{A}} - D(q_{p^\dagger} \| q_{\mathcal{A}}) && \text{by } v^\dagger \in \mathcal{K}_{\mathcal{A}}^d \\ &\geq C_{\mathcal{A}} - \chi^2(q_{p^\dagger} \| q_{\mathcal{A}}) && \text{by (4),} \\ &= C_{\mathcal{A}} - \|v^\dagger\|_{\mathcal{A}_{\mathcal{A}}} && \text{by (59),} \\ &= C_{\mathcal{A}} - 2\gamma_2 \|v^\dagger\|^2 && \text{by (72a).} \end{aligned}$$

Then γ_2 is positive because otherwise $p^\dagger \in \Pi_{\mathcal{A}}$ would hold, but $p^\dagger \in \mathcal{A} \setminus \Pi_{\mathcal{A}}$ by construction. Invoking (62b) instead of (4) we get

$$\begin{aligned} I(p^\dagger; W) &\geq C_{\mathcal{A}} - \frac{1}{2} \|v^\dagger\|_{\mathcal{A}_{\mathcal{A}}} - \frac{\kappa_{\mathcal{A}}^3}{2} \|v^\dagger\|^3 \\ &= C_{\mathcal{A}} - \gamma_2 \|v^\dagger\|^2 - \frac{\kappa_{\mathcal{A}}^3}{2} \|v^\dagger\|^3 && \text{by (72a).} \end{aligned}$$

Then (73) holds for $p = p^\dagger$ by Jensen's inequality and the concavity of $I(p; W)$ in p .

Furthermore, δ_{p_*} is positive for all $p_* \in \Pi_{\mathcal{A}}$ by definition because $\phi(p_*)$ is positive whenever $\Upsilon(p_*) \neq \{0\}$ for $\Upsilon(p_*)$ defined in (53). On the other hand there are only finitely many distinct $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$ cones, and hence only finitely many distinct $\Upsilon(p_*)$ cones and δ_{p_*} values, for $p_* \in \Pi_{\mathcal{A}}$. Thus the minimization defining δ given in (72b) can be written as a minimum rather than an infimum and δ is positive whenever $\gamma_1 = 0$, as well.

When $\gamma_1 = 0$, there are two groups of p 's we need to consider those for which $\Upsilon(p_*) = \{0\}$ and the rest. For p 's for which $\Upsilon(p_*) = \{0\}$, the inequality (71) follows from (72b), (72c), and (74). Let us proceed with p 's for which $\Upsilon(p_*) \neq \{0\}$. Since $\Upsilon(p_*)$ is a closed convex cone the projection on $\Upsilon(p_*)$ and the projection on its polar cone $\Upsilon(p_*)^\circ$ form an orthogonal decomposition by Lemma 2, i.e.,

$$v_p = v_p^* + u_p \quad \text{and} \quad \langle v_p^*, u_p \rangle = 0, \quad (75)$$

where v_p^* and u_p are

$$v_p^* := \text{P}_{\Upsilon(p_*)}(v_p), \quad \text{and} \quad u_p := \text{P}_{\Upsilon(p_*)^\circ}(v_p). \quad (76)$$

Furthermore, $\mathcal{N}_{\Pi_{\mathcal{A}}}^{\mathcal{A}}(p_*)$ is not only a closed convex cone but also a polyhedral cone; thus we can apply Lemma 1 to assert that the angle between u_p and $\mathcal{K}_{\mathcal{A}}^d$ is bounded below by $\phi(p_*)$ defined in (55).

On the other hand $\langle v_p^*, D(W \| q_{\mathcal{A}}) \rangle = 0$ because $v_p^* \in \mathcal{K}_{\mathcal{A}}^d$ by construction. Thus (75) implies

$$\begin{aligned} \langle v_p, D(W \| q_{\mathcal{A}}) \rangle &= \langle u_p, D(W \| q_{\mathcal{A}}) \rangle \\ &\leq -\|D(W \| q_{\mathcal{A}})\| \cdot \|u_p\| \sin(\phi(p_*)). \end{aligned} \quad (77)$$

To see why the last inequality holds, first note that the angle between u_p and $D(W \| q_{\mathcal{A}})$ lies either in $[0, \frac{\pi}{2} - \phi(p_*)]$ or in

$[\frac{\pi}{2} + \phi(p_*) , \pi]$ because the minimum angle between u_p and $\mathcal{K}_{\mathcal{A}}^d$ is $\phi(p_*) \in [0, \frac{\pi}{2}]$. On the other hand, $\langle v_p, D(W \| q_{\mathcal{A}}) \rangle \leq 0$ by (12); thus the angle between u_p and $D(W \| q_{\mathcal{A}})$ has to lie in $[\frac{\pi}{2}, \pi]$. Thus the angle between u_p and $D(W \| q_{\mathcal{A}})$ lies in $[\frac{\pi}{2} + \phi(p_*), \pi]$ and its cosine is bounded from above by $-\sin(\phi(p_*))$.

Furthermore,

$$\begin{aligned} \|v_p\|_{\mathcal{A}, \mathcal{A}}^2 &= \|v_p^* + u_p\|_{\mathcal{A}, \mathcal{A}}^2, \\ &\stackrel{(a)}{\geq} \left(\|v_p^*\|_{\mathcal{A}, \mathcal{A}} - \|u_p\|_{\mathcal{A}, \mathcal{A}} \right)^2, \\ &\geq \|v_p^*\|_{\mathcal{A}, \mathcal{A}}^2 - 2 \|v_p^*\|_{\mathcal{A}, \mathcal{A}} \cdot \|u_p\|_{\mathcal{A}, \mathcal{A}}, \\ &\stackrel{(b)}{\geq} \|v_p^*\|_{\mathcal{A}, \mathcal{A}}^2 - 2 \operatorname{Tr}[A_{\mathcal{A}}] \|v_p^*\| \cdot \|u_p\|, \\ &\stackrel{(c)}{\geq} 2\gamma_2 \|v_p^*\|^2 - 2 \operatorname{Tr}[A_{\mathcal{A}}] \|v_p^*\| \cdot \|u_p\|, \\ &\stackrel{(d)}{=} 2\gamma_2 \|v_p\|^2 - 2 \left(\operatorname{Tr}[A_{\mathcal{A}}] \|v_p^*\| + \gamma_2 \|u_p\| \right) \|u_p\|, \\ &\stackrel{(e)}{\geq} 2\gamma_2 \|v\|^2 - 2 \frac{\|v\| \|D(W \| q_{\mathcal{A}})\| \sin(\phi(p_*))}{\delta_{p_*}} \|u_p\|, \end{aligned} \quad (78)$$

where (a) follows from the triangle inequality, (b) follows from (64), (c) follows from (72a), (d) follows from (75), (e) follows from $\|v_p^*\| \vee \|u_p\| \leq \|v_p\|$ and the definition of δ_{p_*} given in (72c). For p 's satisfying $\Upsilon(p_*) \neq \{0\}$, the inequality (71) follows from (13), (62a), (72b), (72c), (77), and (78). \square

VI. QUADRATIC DECAY FOR QUANTUM MUTUAL INFORMATION

In this section, we present the analysis for the Quantum Mutual Information between the input and the output of a classical-quantum channel with a finite input set. We will first introduce the quantum information-theoretic framework and quantities. In §VI-A, we extend the analysis in §IV to establish the quadratic decay for the quantum mutual information on classical-quantum channels whose Hilbert spaces at the output are separable. In VI-B, we characterize the slowest decay of quantum mutual information with the distance to the capacity-achieving input distributions on classical-quantum channels with finite-dimensional output Hilbert spaces.

Let \mathcal{H} be a separable Hilbert space, i.e., a complete inner product space that has a countable orthonormal basis. We denote the set of all bounded operators on \mathcal{H} , i.e., all continuous linear mappings of the form $T : \mathcal{H} \rightarrow \mathcal{H}$, by $\mathcal{L}(\mathcal{H})$. The operator absolute value $|T| \in \mathcal{L}(\mathcal{H})$ of a bounded linear operator T is defined in terms of its adjoint operator T^* as

$$|T| := \sqrt{T^* T} \quad \forall T \in \mathcal{L}(\mathcal{H}). \quad (79)$$

An operator T is selfadjoint iff $T^* = T$. We denote a non-commutative quotient for selfadjoint operator T and positive definite operator M as

$$\frac{T}{M} := M^{-\frac{1}{2}} T M^{-\frac{1}{2}}. \quad (80)$$

A gentle introduction to separable Hilbert spaces can be found in [16, Chapter 1].

We denote the set of all density operators, i.e., positive semidefinite operators with unit trace, on a separable Hilbert

space \mathcal{H} by $\mathcal{S}(\mathcal{H})$. The eigenvalues of a density operator in $\mathcal{S}(\mathcal{H})$ correspond to a probability mass function, [16, Theorem 2.5]. The *quantum relative entropy*, a quantum generalization of the Kullback–Leibler divergence, $D(\rho \| \sigma)$ is defined for any $\rho, \sigma \in \mathcal{S}(\mathcal{H})$ as, see [17],

$$D(\rho \| \sigma) := \begin{cases} \rho \prec \sigma \\ \infty & \rho \not\prec \sigma \end{cases}, \quad (81)$$

where Tr is the standard trace, and $\rho \prec \sigma$ means that the support of ρ is contained in that of σ . Furthermore, the quantum relative entropy, is bounded from below in terms of the trace-norm via quantum Pinsker's inequality [18, Theorem 3.1]:

$$D(q_p \| q_{\mathcal{A}}) \geq \frac{1}{2} \|q_p - q_{\mathcal{A}}\|_1^2, \quad (82)$$

where $\|\cdot\|_1$ is the trace-norm, i.e., the trace of the operator absolute value of a bounded operator:

$$\|T\|_1 := \operatorname{Tr}[|T|] \quad \forall T \in \mathcal{L}(\mathcal{H}). \quad (83)$$

On the other hand, the quantum relative entropy is bounded above by the quantum χ^2 divergence, see [19, Theorem 8] for a proof for finite dimensional Hilbert spaces,

$$\chi^2(\rho \| \sigma) \geq D(\rho \| \sigma) \quad (84)$$

where χ^α divergence is defined for $\alpha > 1$ as,

$$\chi^\alpha(\rho \| \sigma) := \begin{cases} (\alpha - 1) \int_0^\infty \operatorname{Tr} \left[\left(\frac{|\rho - \sigma|}{\sigma + uI} \right)^\alpha \right] du & \rho \prec \sigma \\ \infty & \rho \not\prec \sigma \end{cases}, \quad (85)$$

where I stands for the identity operator on \mathcal{H} .

When ρ and σ commute, i.e. when they have the same set of eigenvectors, the definition in (85) reduces to the one in (5), as expected.

We can also bound $D(\rho \| \sigma)$ in terms of $\chi^2(\rho \| \sigma)$ and $\chi^3(\rho \| \sigma)$ (as in (6) for the classical setting, but in a slightly different way) as follows

$$D(\rho \| \sigma) - \frac{1}{2} \chi^2(\rho \| \sigma) \geq -\frac{1}{6} \chi^3(\rho \| \sigma), \quad (86a)$$

$$D(\tilde{\rho}_\tau \| \sigma) - \frac{1}{2} \chi^2(\tilde{\rho}_\tau \| \sigma) \leq \frac{1}{6(1-\tau)^2} \chi^3(\tilde{\rho}_\tau \| \sigma), \quad (86b)$$

for all $\rho, \sigma \in \mathcal{S}(\mathcal{H})$, $\tau \in [0, 1)$, where $\tilde{\rho}_\tau := \tau \rho + (1 - \tau) \sigma$, provided that $\chi^3(\rho \| \sigma) < \infty$; see Appendix D for a proof.

A classical-quantum channel $W : \mathcal{X} \rightarrow \mathcal{S}(\mathcal{H})$ maps letters of the input alphabet \mathcal{X} to a density operator on the output Hilbert space \mathcal{H} . For any $W : \mathcal{X} \rightarrow \mathcal{S}(\mathcal{H})$ and $p \in \mathcal{P}(\mathcal{X})$, the mutual information $I(p; W)$ is defined as

$$I(p; W) := \sum_x p(x) D(W \| q_p), \quad (87)$$

where $q_p \in \mathcal{S}(\mathcal{H})$ is the output density operator induced by the input distribution p , for any $p \in \mathcal{P}(\mathcal{X})$, which is defined more generally for any $v : \mathcal{X} \rightarrow \mathbb{R}$ with a countable support satisfying $\sum_x |v(x)| < \infty$ as

$$q_v := \sum_x v(x) W(x). \quad (88)$$

Note that (9) can be confirmed for the quantum case by substitution using (88), instead of (8). Furthermore, all of the properties of the Shannon capacity and center discussed in §II hold for the classical-quantum channels, as well, see for

example [20, Theorem 2] discussing the case of image-additive quantum channels, which covers as a special case the classical to quantum channels, with a finite dimensional \mathcal{H} . Thus, (13) holds for classical-quantum channels, i.e., for any $p_* \in \Pi_{\mathcal{A}}$ and $p \in \mathcal{A}$,

$$I(p; W) = C_{\mathcal{A}} + \langle p - p_*, D(W \| q_{\mathcal{A}}) \rangle - D(q_p \| q_{\mathcal{A}}), \quad (89)$$

where $q_{\mathcal{A}} \in \mathcal{S}(\mathcal{H})$ is the Shannon center for the classical-quantum channel W for the convex constraint set \mathcal{A} , satisfying $q_{\mathcal{A}} = q_{p_*}$ for all $p_* \in \Pi_{\mathcal{A}}$.

Without loss of generality, we assume that $\mathcal{S}(\mathcal{H})$ equals to the union of the supports of all the channel outputs $W(x)$'s; otherwise, we may restrict the underlying Hilbert space to this union. Such a consideration ensures that the Shannon center $q_{\mathcal{A}}$ to have full support.

A. A Simple Proof for Separable Hilbert Spaces via Pinsker's inequality

Using (82) and (89), we can confirm that (41) for classical-quantum channels, as well. Thus for any $p \in \mathcal{A}$ and $p_* \in \Pi_{\mathcal{A}}$, we have

$$I(p; W) \leq C_{\mathcal{A}} + \langle p - p_*, D(W \| q_{\mathcal{A}}) \rangle - \frac{1}{2} \|q_p - p_*\|_1^2. \quad (90)$$

On the other hand, by following the reasoning as in (42), we can translate the trace-norm on the quantum output space back to the ℓ^2 norm on the classical input space:

$$\|q_v\|_1 \leq \|v\| \cdot \sqrt{|\mathcal{X}|} \quad \forall v \in \mathbb{R}^{\mathcal{X}}. \quad (91)$$

We can follow the argument in the proof of Theorem 1 given in §IV by invoking (90) and (91) in place of (41) and (42) to get the following result on the quadratic decay of the quantum mutual information for a classical-quantum channel.

Theorem 3. *Let $W : \mathcal{X} \rightarrow \mathcal{S}(\mathcal{H})$ be a classical-quantum channel with a finite input set \mathcal{X} and separable Hilbert space \mathcal{H} , and $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$ be a closed convex polyhedral constraint set, i.e. a constraint set that can be characterized by a finite number of linear constraints, then*

$$I(p; W) \leq C_{\mathcal{A}} - \gamma \|p - P_{\Pi_{\mathcal{A}}}(p)\|^2 \quad \forall p \in \Pi_{\mathcal{A}}^{\delta}, \quad (92)$$

for the set $\Pi_{\mathcal{A}}^{\delta}$ defined in (18) and positive constants β , γ , and δ defined in (44).

B. Exact Characterization Of The Slowest Decay

We define the *Bogoliubov–Kubo–Mori inner product* with respect to some positive definite operator $\sigma \in \mathcal{L}(\mathcal{H})$ on bounded operator space $\mathcal{L}(\mathcal{H})$ over field \mathbb{C} [21, §7.5] as

$$\langle \rho, \omega \rangle_{\text{BKM}}^{\sigma} := \int_0^{\infty} \text{Tr} \left[\frac{\rho^*}{\sigma + uI} \frac{\omega}{\sigma + uI} \right] du \quad \forall \rho, \omega \in \mathcal{L}(\mathcal{H}).$$

For any classical-quantum channel $W : \mathcal{X} \rightarrow \mathcal{S}(\mathcal{H})$ with a finite-dimensional \mathcal{H} and convex constraint set $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$, we define the set $\mathcal{X}_{\mathcal{A}}$ using (56) and the extended real valued function $\Lambda_{\mathcal{A}} : \mathcal{X}_{\mathcal{A}} \times \mathcal{X}_{\mathcal{A}} \rightarrow [0, \infty]$ via the Bogoliubov–Kubo–Mori inner product:

$$\Lambda_{\mathcal{A}}(x, z) := \langle W(x), W(z) \rangle_{\text{BKM}}^{q_{\mathcal{A}}} \quad \forall x, z \in \mathcal{X}_{\mathcal{A}}. \quad (93)$$

For the case when $W(x)$, $W(z)$, and $q_{\mathcal{A}}$ mutually commute the definition in (93) reduces to the one in (57) given in §V-A, as expected.

When $\mathcal{X}_{\mathcal{A}}$ is a finite set and $\max_{x,z} \Lambda_{\mathcal{A}}(x, z)$ is finite, then $\Lambda_{\mathcal{A}}$ is a positive semi-definite matrix because

$$v^T \Lambda_{\mathcal{A}} v = \int_0^{\infty} \text{Tr} \left[\left(\frac{q_v}{q_{\mathcal{A}} + uI} \right)^2 \right] du \geq 0 \quad \forall v \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}.$$

Thus $\Lambda_{\mathcal{A}}$ defines an inner product on $\mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ for classical-quantum channels, as well. Furthermore, the resulting norm is related to the quantum χ^2 divergence as follows:

$$\chi^2(q_p \| q_{\mathcal{A}}) = \|p - p_*\|_{\Lambda_{\mathcal{A}}}^2 \quad (94)$$

for all $p \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ and $p_* \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ satisfying $q_{p_*} = q_{\mathcal{A}}$ for q_{p_*} is defined in (88).

On the other hand we can bound the operator absolute value q_v for any v with finite $\|v\|$ as follows. First note that the formula of the operator perfect square and its positive semi-definiteness, imply the following operator inequality

$$\begin{aligned} (q_v)^2 &= \left(\sum_{x \in \mathcal{X}_{\mathcal{A}}} v(x) W(x) \right)^2 \\ &\leq \|v\|^2 \cdot \left(\sum_{x \in \mathcal{X}_{\mathcal{A}}} W(x)^2 \right). \end{aligned}$$

Since the square-root is operator monotone (see e.g., [21, §4]), we have

$$|q_v| \leq \|v\| \cdot \sqrt{\sum_{x \in \mathcal{X}_{\mathcal{A}}} W(x)^2}. \quad (95)$$

Then as a result of the monotonicity of the map $\text{Tr}[(\cdot)^3]$ we have

$$\chi^3(q_p \| q_{\mathcal{A}}) \leq \|p - p_*\|^3 (\kappa_{\mathcal{A}})^3, \quad (96)$$

for all $p \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$ and $p_* \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$, where $\kappa_{\mathcal{A}}$ is defined as follows

$$\kappa_{\mathcal{A}} := \sqrt[3]{\int_0^{\infty} \text{Tr} \left[\left(\frac{\sqrt{\sum_{x \in \mathcal{X}_{\mathcal{A}}} W(x)^2}}{q_{\mathcal{A}} + uI} \right)^3 \right] du}. \quad (97)$$

When $\{W(x)\}_{x \in \mathcal{X}_{\mathcal{A}}}$ mutually commute $\kappa_{\mathcal{A}}$ defined in (97) reduces to the one in (61).

As in §V-A, applying (86) for $\rho = q_p$ and $\sigma = q_{\mathcal{A}}$, and invoking (94) and (96), we obtain the following lemma for classical-quantum channels.

Lemma 4. *For any classical-quantum channel $W : \mathcal{X} \rightarrow \mathcal{S}(\mathcal{H})$ with a finite input set \mathcal{X} and finite-dimensional Hilbert space \mathcal{H} and convex constraint set $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$, we have*

$$D(q_p \| q_{\mathcal{A}}) - \frac{1}{2} \|v_p\|_{\Lambda_{\mathcal{A}}}^2 \geq -\frac{\kappa_{\mathcal{A}}^3}{6} \|v_p\|^3, \quad (98a)$$

$$D(q_{p+\tau v_p} \| q_{\mathcal{A}}) - \frac{1}{2} \|\tau v_p\|_{\Lambda_{\mathcal{A}}}^2 \leq \frac{\kappa_{\mathcal{A}}^3}{6(1-\tau)^2} \|\tau v_p\|^3, \quad (98b)$$

for $v_p := p - p_*$ all $p \in \mathcal{A}$, $p_* \in \Pi_{\mathcal{A}}$, and $\tau \in [0, 1]$.

In our analysis on classical-quantum channels, we will need an operator-norm bound analogous to (42), as well. To that end

we invoke (95), and bound $\|v\|_{A_{\mathcal{A}}}$ from above in terms of $\|v\|$ for an arbitrary $v \in \mathbb{R}^{\mathcal{X}_{\mathcal{A}}}$, as follows

$$\begin{aligned} \|v\|_{A_{\mathcal{A}}}^2 &= \int \text{Tr} \left[\left(\frac{q_v}{q_{\mathcal{A}} + uI} \right)^2 \right] du \\ &\leq \int \|v\|^2 \text{Tr} \left[\left(\frac{\sqrt{\sum_{x \in \mathcal{X}_{\mathcal{A}}} W(x)^2}}{q_{\mathcal{A}} + uI} \right)^2 \right] du \\ &= \|v\|^2 \text{Tr} [A_{\mathcal{A}}]. \end{aligned} \quad (99)$$

We apply the analysis of Theorem 2 given in §V by invoking Lemma 4, (84) and (99) in place of Lemma 3, (4), and (64) to obtain the following result of the exact characterization of the slowest decay for quantum mutual information on classical-quantum channels with finite-dimensional Hilbert space \mathcal{H} .

Theorem 4. *For a classical-quantum channel $W : \mathcal{X} \rightarrow \mathcal{S}(\mathcal{H})$ with a finite input set \mathcal{X} and a finite-dimensional Hilbert space \mathcal{H} , a closed convex polyhedral constraint set $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$,*

$$I(p; W) \leq C_{\mathcal{A}} - \gamma_1 \|v_p\| \quad \forall p \in \mathcal{A} \quad (100)$$

for γ_1 defined in (68), where $v_p := p - \text{P}_{\Pi_{\mathcal{A}}}(p)$, and there exists a $p \in \mathcal{A} \setminus \Pi_{\mathcal{A}}$ satisfying

$$I(p + \tau v_p; W) \geq C_{\mathcal{A}} - \gamma_1 \|\tau v_p\| - \text{Tr} [A_{\mathcal{A}}] \|\tau v_p\|^2 \quad (101)$$

for all $\tau \in (0, 1)$. Furthermore, if $\gamma_1 = 0$, then

$$I(p; W) \leq C_{\mathcal{A}} - \gamma_2 \|v_p\|^2 + \frac{\kappa_{\mathcal{A}}^3}{6} \|v_p\|^3 \quad \forall p \in \Pi_{\mathcal{A}}^{\delta} \quad (102)$$

for positive constants γ_2 , δ , $\kappa_{\mathcal{A}}$ defined in (72a), (72b), (97) for $A_{\mathcal{A}}$ given in (93), and there exists a $p \in \mathcal{A} \setminus \Pi_{\mathcal{A}}$ satisfying for all $\tau \in (0, 1)$

$$I(p + \tau v_p; W) \geq C_{\mathcal{A}} - \gamma_2 \|\tau v_p\|^2 - \frac{\kappa_{\mathcal{A}}}{6(1-\tau)^2} \|\tau v_p\|^3. \quad (103)$$

VII. DISCUSSION

We have extend Strassen's observation in (1) and bound the mutual information from above by a function that is decreasing quadratically with the distance to the set of all capacity-achieving input distributions, for channels with finite input sets and with a finite number of linear constraints in Theorem 1, using Pinsker's inequality (i.e., (3)), Topsøe identity (i.e., (9)), and positivity of the minimum angle implied by the polyhedral convexity (i.e., Lemma 1). We have also shown that same tools suffice to establish (1) for quantum mutual information on classical-quantum channels whose output Hilbert spaces are separable (possibly infinite dimensional) in Theorem 3. For general convex constraint sets that may not be expressed as a finite number of linear constraints (1) might not hold; Example 1 demonstrates it for a channel with three input letters and two output letters. If input set is infinite then the only function $f : [0, \delta] \rightarrow \mathbb{R}_{\geq 0}$ satisfying $I(p; W) \leq C - f(\|p - p_*\|)$ for all $\|p - p_*\| \leq \delta$ for a positive δ can be $f(z) = 0$; Example 2 demonstrates this for a channel with a countable input set and two output letters.

We have also determined the exact leading term in the Taylor series expansion of the slowest decay of the mutual information around the capacity-achieving input distributions

for channels with finite input sets and with a finite number of linear constraints in Theorem 1, using a Taylor series expansion of Kullback–Leibler divergence (i.e., (6)), Topsøe identity (i.e., (9)), and Moreau's decomposition theorem (i.e., Lemma 2). We have also shown that same tools suffice to determine the leading terms in the Taylor series expansion of the slowest decay of the quantum mutual information around capacity-achieving distributions on classical-quantum channels whose output Hilbert spaces are finite dimensional in Theorem 4. We have worked with the Euclidean distance, i.e., ℓ^2 , norm; but the same tools and analysis can be applied to determine the leading term in Taylor series expansion when distance is measure using another norm on the affine space including the convex constraint set \mathcal{A} , e.g., when we work with $\|\cdot\|_1$ instead of $\|\cdot\|$.

Under appropriate technical assumptions, one can obtain (1) for Augustin information [22]–[25] using the same framework, as well.

ACKNOWLEDGMENT

Authors would like to thank Michael X. Cao and Marco Tomamichel for bringing to their attention the gap in Strassen's proof [1] and for the subsequent discussion on the topic.

APPENDIX A

GAP IN STRASSEN'S ARGUMENT

The first two terms of the Taylor expansion characterizing the change of the mutual information around any capacity-achieving input distribution is determined in [1] to be

$$f(v) = \langle v, D(W \| q_W) \rangle + \frac{1}{2} \|v\|_{A_W}^2 \quad \forall v \in \mathbb{R}^{\mathcal{X}}.$$

[1, (4.41)] asserts that for small enough δ there exists a $\gamma > 0$ satisfying

$$f(p - p_*) \leq -\gamma \|p - p_*\|^2 \quad \forall p \in \Pi^{\delta}, \quad (4.41)$$

where p_* is the projection of p to Π . To establish (4.41) Strassen asserts that if (4.41) does not hold then there must exist a sequence $\{p_j\}_{j \in \mathbb{Z}_+} \subset \Pi^{\delta}$ satisfying

$$\liminf_j f(p_j - p_{j,*}) \geq 0. \quad (104)$$

Furthermore, Strassen asserts since $\langle p - p_*, D(W \| q_W) \rangle \leq 0$ for all $p \in \mathcal{P}(\mathcal{X})$, one can assume

$$\|p_j - p_{j,*}\| = \delta \quad \forall j \in \mathbb{Z}_+. \quad (105)$$

We agree with this assertion because of the following reasoning: If Π is the in the relative interior of the probability simplex, i.e., $\Pi \cap \partial \mathcal{P}(\mathcal{X}) = \emptyset$, then for small enough δ any point p on the boundary Π^{δ} will satisfy $\|p - p_*\| = \delta$ and the identity $\langle p - p_*, D(W \| q_W) \rangle \leq 0$ for all $p \in \mathcal{P}(\mathcal{X})$ implies

$$f(p - p_*) \leq \frac{\|p - p_*\|^2}{\delta^2} f\left(\frac{p - p_*}{\|p - p_*\|} \delta\right) \quad \forall p \in \Pi^{\delta}. \quad (106)$$

Thus if the sequence satisfying (104) does not satisfy (105), then we can replace each p_j with $\tilde{p}_j = p_{j,*} + \frac{p_j - p_{j,*}}{\|p_j - p_{j,*}\|} \delta$ to get a sequence satisfying both (104) and (105). Note that $\tilde{p}_{j,*} = p_{j,*}$ by construction.

However, for certain channels Π might have points outside the relative interior of the probability simplex, i.e., $\Pi \cap \partial\mathcal{P}(\mathcal{X}) \neq \emptyset$. The unconstrained version of the channel considered in Example 1 is such a channel. The argument we have for $\Pi \cap \partial\mathcal{P}(\mathcal{X}) = \emptyset$ case will not work as is in this case because there might not be a positive δ for which infinitely many \tilde{p}_j 's are guaranteed to be in the probability simplex $\mathcal{P}(\mathcal{X})$, and hence in Π^δ . Nevertheless, a sequence satisfying both (104) and (105) exists as claimed by Strassen. To see why first recall that the projection of a $p \in \mathcal{P}(\mathcal{X})$ to Π is p_* iff $p - p_* \in \mathcal{N}_\Pi^A(p_*)$; see (23) and (24). Furthermore, both $\{\mathcal{N}_\Pi^A(p_*) : p_* \in \Pi\}$ and $\{\mathcal{T}_{\mathcal{P}(\mathcal{X})}(p_*) : p_* \in \Pi\}$ are finite sets as a result of the polyhedral convexity of Π and $\mathcal{P}(\mathcal{X})$. Thus the set $\mathcal{S} = \{\mathcal{N}_\Pi^A(p_*) : p_* \in \Pi\}$ is finite and for each $s \in \mathcal{S}$ there exists at least one (often uncountably many) $p_* \in \Pi$ satisfying $s = \mathcal{N}_\Pi^A(p_*)$. For each $s \in \mathcal{S}$ we choose a $p_s \in \Pi$ satisfying $s = \mathcal{N}_\Pi^A(p_s)$. Among $\{\mathcal{N}_\Pi^A(p_{j,*})\}_{j \in \mathbb{Z}_+}$ at least one $s_* \in \mathcal{S}$ will be repeated infinitely often. Let $\{p_{i_j}\}_{j \in \mathbb{Z}_+}$ be a subsequence satisfying $\mathcal{N}_\Pi^A(p_{i_j,*}) = s_*$ for all $j \in \mathbb{Z}_+$. If $\hat{p}_j = p_{s_*} + \frac{p_{i_j} - p_{i_j,*}}{\|p_{i_j} - p_{i_j,*}\|} \delta$, then the projection of \hat{p}_j onto Π is p_{s_*} for all $j \in \mathbb{Z}_+$, because $\hat{p}_j - p_{s_*} \in \mathcal{N}_\Pi^A(p_{s_*})$. Furthermore, as a result of the polyhedral convexity of $\mathcal{P}(\mathcal{X})$ for each $p_* \in \Pi$, there exists a $\delta(p_*) > 0$ such that

$$\left\{ p_* + \delta v : v \in \mathcal{N}_\Pi^A(p_*) \text{ and } \delta \leq \delta(p_*) \right\} \subset \mathcal{P}(\mathcal{X}).$$

Thus (105) holds for \hat{p}_j by construction for $\delta = \min_{s_* \in \mathcal{S}} \delta(p_{s_*})$ and (104) holds for \hat{p}_j by (106).

APPENDIX B

A COUNTER-EXAMPLE FOR [2, (500)]

Example 4. Let W be a channel with 9 input letters and 8 output letters given in the following

$$W = \begin{bmatrix} \varepsilon/3 \mathbf{1}_{5 \times 1} & \varepsilon/3 \mathbf{1}_{5 \times 1} & \varepsilon/3 \mathbf{1}_{5 \times 1} & (1 - \varepsilon) \mathbf{I}_5 \\ 1/2 & 1/3 & 1/6 & \mathbf{0}_{1 \times 5} \\ 1/6 & 1/2 & 1/3 & \mathbf{0}_{1 \times 5} \\ 1/3 & 1/6 & 1/2 & \mathbf{0}_{1 \times 5} \\ 1/3 & 1/2 & 1/6 & \mathbf{0}_{1 \times 5} \end{bmatrix},$$

where $\mathbf{1}_{5 \times 1}$ is a column vector of ones, \mathbf{I}_5 is 5-by-5 identity matrix, $\mathbf{0}_{1 \times 5}$ is a row vector of zeros, and ε is the unique solution of the equation $\sqrt[3]{3} \sqrt[3]{0.002} = \varepsilon 5^{-\varepsilon}$ on $\varepsilon \in (0, 1/2)$.

With a slight abuse of notation when $\mathcal{A} = \mathcal{P}(\mathcal{X})$, we denote the Shannon capacity by C_W and the Shannon center by q_W . Let us assume $\mathcal{A} = \mathcal{P}(\mathcal{X})$. Then the capacity-achieving input distribution is unique and it is the uniform distribution on the first 5 input letters. Furthermore,

$$C_W = (1 - \varepsilon) \ln 5 \quad \text{and} \quad q_W = \left[\frac{\varepsilon}{3} \quad \frac{\varepsilon}{3} \quad \frac{\varepsilon}{3} \quad \frac{1 - \varepsilon}{5} \mathbf{1}_{1 \times 5} \right]^T.$$

Note that $D(W(x) \| q_W) = C_W$ for all input letters x .

On the other hand $\mathcal{K}_W = \{\tau s : \tau \in \mathbb{R}\}$ where the vector s is given by

$$s = [0_{1 \times 5} \quad 2 \quad 2 \quad -1 \quad -3]^T.$$

Note that $\langle s, D(W \| q_W) \rangle = 0$. Thus $\langle v_0, D(W \| q_W) \rangle = 0$ for any p , where v_0 is the projection of $p - p_*$ onto \mathcal{K}_W considered in [2]. On the other hand if p puts non-zero probability only on

one of the last four input letters then $\|v_0\| \neq 0$. Consequently, $\langle v_0, D(W \| q_W) \rangle \leq -\Gamma \|v_0\|$, i.e., [2, (500)], cannot be true for any positive Γ .

APPENDIX C PROOF OF (6)

As a result of the Taylor series expansion of the function $z \ln z$ around $z = 1$ we know that for each $z \in (0, \infty)$ there exists a number \bar{z} between z and 1 such that

$$\begin{aligned} z \ln z &= z - 1 + \frac{1}{2}(z - 1)^2 - \frac{1}{6} \frac{(z - 1)^3}{\bar{z}^2}, \\ &\geq z - 1 + \frac{1}{2}(z - 1)^2 - \frac{1}{6} |(z - 1)^3|^+. \end{aligned} \quad (107)$$

where $|z|^+ = z \vee 0$.

$$\begin{aligned} D(w \| q) &\geq \int \left(\frac{1}{2} \left(\frac{dw}{dq} - 1 \right)^2 - \frac{1}{6} \left| \left(\frac{dw}{dq} - 1 \right)^3 \right|^+ \right) dq \\ &\geq \int \left(\frac{1}{2} \left(\frac{dw}{dq} - 1 \right)^2 - \frac{1}{6} \left| \left(\frac{dw}{dq} - 1 \right)^3 \right| \right) dq \\ &= \frac{1}{2} \chi^2(w \| q) - \frac{1}{6} \chi^3(w \| q) \end{aligned}$$

Note that by (107) we know that

$$z \ln z \leq z - 1 + \frac{1}{2}(z - 1)^2 \quad z > 1. \quad (108)$$

On the other hand, as a result of Taylor series expansion of the function $\ln z$ around $z = 1$ we know that for each $z \in (0, \infty)$ there exists a number \bar{z} between z and 1 such that

$$\ln z = z - 1 - \frac{1}{2} \frac{(z - 1)^2}{\bar{z}^2}$$

Thus for all $z \in (0, 1)$ we have,

$$\begin{aligned} \ln z &\leq z - 1 - \frac{1}{2}(z - 1)^2 \\ z \ln z &\leq z(z - 1) - \frac{z}{2}(z - 1)^2 \\ &= z - 1 + \frac{1}{2}(z - 1)^2 + \frac{1}{2}(1 - z)^3 \quad z \in (0, 1). \end{aligned} \quad (109)$$

As a result of (108) and (109) we have

$$z \ln z \leq z - 1 + \frac{1}{2}(z - 1)^2 + \frac{1}{2}|1 - z|^{+3}.$$

Thus

$$\begin{aligned} D(w \| q) &\leq \int \left(\frac{1}{2} \left(\frac{dw}{dq} - 1 \right)^2 + \frac{1}{2} \left| \left(1 - \frac{dw}{dq} \right)^3 \right|^+ \right) dq \\ &\leq \int \left(\frac{1}{2} \left(\frac{dw}{dq} - 1 \right)^2 + \frac{1}{2} \left| \left(1 - \frac{dw}{dq} \right)^3 \right| \right) dq \\ &= \frac{1}{2} \chi^2(w \| q) + \frac{1}{2} \chi^3(w \| q). \end{aligned}$$

APPENDIX D PROOF OF (86)

We first prove (86a). Fix an invertible density operator σ and let $\delta := \rho - \sigma$. We consider the Taylor's series expansion of the map $[0, 1] \ni \tau \mapsto f(\tau) := D(\sigma + \tau \delta \| \sigma)$ around 0:

$$f(\tau) = \sum_{n=0}^3 \frac{f^{(n)}(0)}{n!} \cdot \tau^n + \frac{f^{(4)}(\bar{\tau})}{4!} \cdot \tau^4$$

for some $\bar{\tau} \in [0, \tau]$. By standard calculations (see e.g., [21, §3]), we have

$$\begin{aligned} f^{(0)}(\tau) &= f(\tau), \\ f^{(1)}(\tau) &= \text{Tr}[\delta \cdot (I + \ln(\sigma + \tau\delta) - \ln \sigma)], \\ f^{(2)}(\tau) &= \int_0^\infty \text{Tr} \left[\left(\frac{\delta}{\sigma + \tau\delta + uI} \right)^2 \right] du, \\ f^{(3)}(\tau) &= -2 \cdot \int_0^\infty \text{Tr} \left[\left(\frac{\delta}{\sigma + \tau\delta + uI} \right)^3 \right] du, \\ f^{(4)}(\tau) &= 6 \cdot \int_0^\infty \text{Tr} \left[\left(\frac{\delta}{\sigma + \tau\delta + uI} \right)^4 \right] du. \end{aligned}$$

For each $u \geq 0$ and $\tau \in [0, 1]$,

$$\text{Tr} \left[\left(\frac{\delta}{\sigma + \tau\delta + uI} \right)^3 \right] \leq \text{Tr} \left[\left(\frac{|\delta|}{\sigma + \tau\delta + uI} \right)^3 \right]$$

by the operator inequality $\delta \leq |\delta|$, because the map $T(\cdot)T^*$ is a positive map for any bounded operator $T \in \mathcal{L}(\mathcal{H})$ (see e.g., [16, §4]), and the map $\text{Tr}[(\cdot)^3]$ is monotone.

Moreover, the fourth-order term is non-negative for all $\tau, \bar{\tau} \in [0, 1]$, we immediately obtain

$$f(\tau) \geq \sum_{n=0}^3 \frac{f^{(n)}(0)}{n!} \cdot \tau^n.$$

Letting $\tau = 1$, we conclude the lower bound (86a).

Next, we prove the upper bound (86b). Again, we use Taylor's series expansion of the map $[0, 1] \ni \tau \mapsto f(\tau)$ around 0, but now up to the third order:

$$f(\tau) = \sum_{n=0}^2 \frac{f^{(n)}(0)}{n!} \cdot \tau^n + \frac{f^{(3)}(\bar{\tau})}{3!} \cdot \tau^3$$

for some $\bar{\tau} \in [0, \tau]$. It remains to upper bound the third-order term. For each $u \geq 0$, we have,

$$\begin{aligned} -\text{Tr} \left[\left(\frac{\delta}{\sigma + \bar{\tau}\delta + uI} \right)^3 \right] &\stackrel{(a)}{\leq} \text{Tr} \left[\left(\frac{|\delta|}{\sigma + \bar{\tau}\delta + uI} \right)^3 \right] \\ &\stackrel{(b)}{=} \text{Tr} \left[\left(|\delta|^{\frac{1}{2}} (\sigma + \bar{\tau}\delta + uI)^{-1} |\delta|^{\frac{1}{2}} \right)^3 \right] \\ &\stackrel{(c)}{\leq} \text{Tr} \left[\left(|\delta|^{\frac{1}{2}} ((1-\tau)\sigma + uI)^{-1} |\delta|^{\frac{1}{2}} \right)^3 \right] \\ &= \frac{1}{(1-\tau)^3} \text{Tr} \left[\left(\frac{|\delta|}{\sigma + \bar{u}I} \right)^3 \right], \end{aligned}$$

where $\bar{u} := \frac{u}{1-\tau}$ and (a) follows from the operator inequality $-\delta \leq |\delta|$, because the map $T(\cdot)T^*$ is a positive map and the map $\text{Tr}[(\cdot)^3]$ is monotone, (b) follows from the cyclic property of trace, (c) follows from the operator inequality $\sigma + \bar{\tau}\delta = \bar{\tau}\rho + (1-\bar{\tau})\sigma \geq (1-\bar{\tau})\sigma \geq (1-\tau)\sigma$ because the inverse is operator monotone decreasing.

After re-parameterizing and integrating on $u \in (0, \infty)$, we get the upper bound (86b).

REFERENCES

[1] V. Strassen, "Asymptotische abschätzungen in Shannons Informations-theorie," in *Trans. Third Prague Conf. Inf. Theory*, 1962, pp. 689–723, (<https://pi.math.cornell.edu/~pmlut/strassen.pdf>).

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

[3] M. Tomamichel and V. Y. F. Tan, "A tight upper bound for the third-order asymptotics for most discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7041–7051, Nov 2013.

[4] M. X. Cao and M. Tomamichel, "On the quadratic decaying property of the information rate function," *arXiv:2208.12945v1 [cs.IT]*, 2022. [Online]. Available: <https://arxiv.org/abs/2208.12945v1>

[5] —, "Comments on "channel coding rate in the finite blocklength regime": On the quadratic decaying property of the information rate function," *arXiv:2208.12945v2 [cs.IT]*, 2023. [Online]. Available: <https://arxiv.org/abs/2208.12945v2>

[6] I. Csizsár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, no. 3-4, pp. 299–318, 1967.

[7] F. E. Su, "Methods for quantifying rates of convergence for random walks on groups," Ph.D. Thesis, Harvard University, 1995.

[8] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review / Revue Internationale de Statistique*, vol. 70, no. 3, pp. 419–435, 2002. [Online]. Available: <http://www.jstor.org/stable/1403865>

[9] I. Vajda, *Theory of statistical inference and information*. Dordrecht: Kluwer Academic Publishers, 1989.

[10] —, " χ^α -divergence and generalized fisher's informations," in *Proceedings 6th Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes*, 1973, pp. 873–886.

[11] F. Liese and I. Vajda, *Convex Statistical Distances*, ser. Teubner-Texte zur Mathematik. Teubner, 1987, vol. 95.

[12] F. Topsøe, "An information theoretical identity and a problem involving capacity," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 291–292, 1967.

[13] J. H. B. Kemperman, "On the Shannon capacity of an arbitrary channel," *Indagationes Mathematicae (Proceedings)*, vol. 77, no. 2, pp. 101–115, 1974.

[14] B. Nakiboğlu, "The Rényi Capacity and Center," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 841–860, Feb 2019, (arXiv:1608.02424 [cs.IT]).

[15] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, 1st ed., ser. Grundlehren Text Editions. Heidelberg: Springer-Verlag Berlin, 2001.

[16] T. Heinosaari and M. Ziman, *The Mathematical Language of Quantum Theory*. Cambridge University Press, dec 2011.

[17] H. Umegaki, "Conditional expectation in an operator algebra. IV. entropy and information," *Kodai Mathematical Journal*, vol. 14, no. 2, jan 1962.

[18] F. Hiai, M. Ohya, and M. Tsukada, "Sufficiency, KMS condition and relative entropy in von Neumann algebras," *Pacific Journal of Mathematics*, vol. 96, no. 1, pp. 99–109, sep 1981.

[19] K. Temme, M. J. Kastoryano, M. B. Ruskai, M. M. Wolf, and F. Verstraete, "The χ^2 -divergence and mixing times of quantum Markov processes," *Journal of Mathematical Physics*, vol. 51, no. 12, p. 122201, dec 2010.

[20] M. Tomamichel and V. Y. F. Tan, "Second-order asymptotics for the classical capacity of image-additive quantum channels," *Communications in Mathematical Physics*, vol. 338, no. 1, pp. 103–137, May 2015.

[21] F. Hiai and D. Petz, *Introduction to Matrix Analysis and Applications*. Springer International Publishing, 2014.

[22] U. Augustin, "Noisy channels," Habilitation Thesis, Universität Erlangen-Nürnberg, 1978, (<http://bit.ly/3bsWDgG>).

[23] I. Csizsár, "Generalized cutoff rates and Rényi's information measures," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 26–34, Jan 1995.

[24] B. Nakiboğlu, "The Augustin Capacity and Center," *Problems of Information Transmission*, vol. 55, no. 4, pp. 299–342, October 2019, (arXiv:1803.07937 [cs.IT]).

[25] H.-C. Cheng, M. H. Hsieh, and M. Tomamichel, "Quantum sphere-packing bounds with polynomial prefactors," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 2872–2898, May 2019, (arXiv:1704.05703 [quant-ph]).