Channel Additivity Conjectures

Alex Kazachek

Abstract

This note aims to be a self-contained exposition on the use of probabilistic arguments to disprove the additivity of quantum channel capacity. It contains a proof of the Dvoretzky-Milman theorem, as well as a primer on channel additivity problems, before using the theorem to show that minimum output p-entropy does not hold for $p > 1$.

18 December 2023

1 Asymptotic Geometric Analysis

The goal of this section is to prove a single result – the Dvoretzky-Milman theorem. The actual theorem is due to Dvoretzky alone [Dvo64]. However, Milman contributed a novel proof [Mil71] which spawned the field of asymptotic geometric analysis – the name of this section. We will follow exactly the approach of [AS17], where it is given as THEOREM 7.19, and we merely descend the referenced lemmas and propositions from earlier chapters.

Asymptotic geometric analysis concerns itself with the structure of finite-dimensional spaces, such as subspaces and convex bodies, in asymptotic cases. The understanding comes from probabilistic arguments to tame deviations and sizes.

We care because this theorem will help us to prove (or rather, disprove) claims about quantum channel additivity. This is as quantum channels may be viewed as subspaces of some Hilbert space, and we can therefore use probabilistic arguments à la Dvoretzky to speak of them. We will do this in the second section.

1.1 Norms and Notation

We will let a generic Hilbert space be $\mathcal{H}$, with some norm $\|\cdot\|_{\mathcal{H}}$ and inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. The subscripts will be omitted if the context is clear. For $\mathcal{H} = \mathbb{R}^n, \mathbb{C}^n$ with $\|\cdot\|_{\mathcal{H}}$ being the standard Euclidean norm, we will just write $|\cdot|$. Spaces of m-by-n matrices over a field $K = \mathbb{C}, \mathbb{R}$ will be denoted $\mathrm{M}_{mn}(K)$. For square ones, we will only write a single subscript.

Definition 1 (Schatten p-Norm)**.** *Consider $\mathrm{M}_{mn}(K)$. For $p \in [1, \infty)$ we endow it with the Schatten p-norm $\|M\|_p = (\operatorname{tr}|M|^p)^{1/p}$. The absolute value $|M| = \sqrt{M^*M}$ is interpreted in the sense of functional calculus.*

We may extended this to the ∞-norm by taking the limit, and this just corresponds to the usual operator norm:

$$\|M\|_\infty = \sup_{x\neq 0} \frac{|Mx|}{|x|} = \|M\|_{\mathrm{op}}.$$

Also note that due to the functional calculus, Schatten p-norms can be rephrased in terms of ℓ^p-norms on singular values.

There are two other special cases. Namely, $p = 1$, which is simply $\|M\|_1 = \operatorname{tr}|M|$ – the trace norm. There is also the Hilbert-Schmidt norm when $p = 2$, given by $\|M\|_2 = \sqrt{\operatorname{tr} M^*M}$. These both are readily extended to infinite-dimensional Hilbert spaces, provided they take finite values (the set of operators which do are denoted respectively as trace-class or Hilbert-Schmidt).

Lemma 1. *Let T be Hilbert-Schmidt operator. Then, $\|T\|_{op} \leq \|T\|_{HS}$*

Proof. Denote the underlying Hilbert space by $\mathcal{H}$ and find some orthonormal basis e_i. By definition of the norm and trace we have

$$\|T\|_{\mathrm{HS}}^2 = \operatorname{tr} T^*T = \sum_i \langle T^*Te_i, e_i\rangle = \sum_i \|Te_i\|_{\mathcal{H}}^2.$$

Now, let $x \in \mathcal{H}$ and witness

$$\|Tx\|_{\mathcal{H}} = \left\|\sum_i \langle x, e_i\rangle Te_i\right\|_{\mathcal{H}} \leq \sum_i |\langle x, e_i\rangle| \|Te_i\|_{\mathcal{H}}.$$

Applying Hölder's inequality gives

$$\sum_i |\langle x, e_i\rangle| \|Te_i\|_{\mathcal{H}} = \sqrt{\sum_i |\langle x, e_i\rangle|^2}\sqrt{\sum_i \|Te_i\|^2}.$$

We recognize the first radicand is $\|x\|_{\mathcal{H}}^2$ due to Parseval's identity, while the latter is $\|T\|_{\mathcal{HS}}$ due to our previous work. That is, $\|Tx\|_{\mathcal{H}} \leq \|x\|_{\mathcal{H}}\|T\|_{\mathrm{HS}}$. Optimizing over all unit vectors gives the desired inequality with the operator norm. □

We will let S^{n-1} be the unit sphere in $\mathbb{R}^n$. We will denote by $S_{\mathbb{C}}^{n-1}$ the unit sphere in $\mathbb{C}^n$. Since $\mathbb{C}^n \cong \mathbb{R}^{2n}$, there will be lots of doubling and halving as we move between "which" sphere we are considering.

1.2 Groups and Geodesics

We will mostly consider ourselves with the two standard matrix groups – the orthogonal group $\mathrm{O}(n)$ of orthogonal n-by-n real matrices and the unitary group $\mathrm{U}(n)$ of unitary n-by-n complex matrices. They also have the associated special orthogonal and special unitary groups, respectively $\mathrm{SO}(n)$ and $\mathrm{SU}(n)$, which is the restriction to those with unit determinant.

Theorem 1 (Haar Measure)**.** *Let G be a compact group. Given any Borel $S \subseteq G$ and $g \in G$, we define the left-translation gS. Now, there exists a unique probability measure defined on the Borel sets of G which is left-translation invariant, as well as inner- and outer-regular. We call this the Haar measure of G.*

Actually, this measure exists even if G is not compact. But then its total measure is infinite, so it cannot be made into a probability measure. In fact, the Lebesgue measure is precisely this infinite Haar measure for the group $\mathbb{R}$ under addition.

Definition 2 (Grassmannian)**.** *Define the Grassmannian $\mathrm{Gr}(k, n)$ of k-dimensional subspaces of $\mathbb{C}^n$. We similarly define $\mathrm{Gr}_{\mathbb{R}}(k, n)$ for k-dimensional subspaces of $\mathbb{R}^n$.*

The Grassmannian itself is not a group and so a priori it does not make sense to speak of random k-dimensional subspaces. However, $\mathrm{U}(n)$ is a perfectly fine group and has its own Haar measure. Therefore, when we speak of a randomly chosen k-dimensional subspace of $\mathbb{C}^n$, we mean fixing an arbitrary $F \in \mathrm{Gr}(k, n)$ and consider the random subspace $E = UF$, where U is a Haar-distributed random unitary over $U(n)$. We will often take $F = \mathbb{C}^k$, by which we mean the inclusion $\imath(\mathbb{C}^k) \subseteq \mathbb{C}^n$.

Definition 3 (Geodesic). *Let (M,d) be a metric space, and let $x,y \in M$. We say a curve $\gamma\colon [0,1] \to M$ is a geodesic connecting x and y if $\gamma(0) = x$, $\gamma(1) = y$, and for all $t \in [0,1]$, there exists some neighbourhood T about t and some constant v so*

$$d(\gamma(a), \gamma(b)) = v|a-b|$$

for all $a,b \in T$.

Definition 4 (Geodesic Distance). *Let (M,d) be a metric space, and let $x,y \in M$. We define the length of a geodesic γ connecting x and y by*

$$L(\gamma) = \sup_{0=x_0<\cdots<x_n=1} \sum_{i=1}^{n} d(\gamma(x_{i-1}), \gamma(x_i)).$$

Then, we define the geodesic distance between x and y by $g_d(x,y) = \inf_\gamma L(\gamma)$, where the infimum is of course over all relevant geodesics connecting the two points.

Theorem 2. *Let $M = \mathrm{SO}(n), \mathrm{U}(n), \mathrm{SU}(n)$, and let $p \in [1,\infty]$. For $U,V \in M$, denote by $\exp(i\vartheta_j)$ the eigenvalues of $U^{-1}V$, choosing $|\vartheta_j| \leq \pi$. Then,*

$$g_p(U,V) = \|(\vartheta_1, \ldots, \vartheta_n)\|_p$$

is the geodesic distance on M with respect to the Schatten p-norm. In particular, the minimizing geodesic is given by $\gamma(t) = U\exp(itA)$, where A satisfies $\exp(iA) = U^{-1}V$ and $\|A\|_\infty \leq \pi$. Moreover,

$$\frac{2}{\pi} g_p(U,V) = \|U-V\|_p \leq g_p(U,V).$$

We can actually do a little better and push this procedure up to Grassmannians by means of the orthogonal and unitary groups. Specifically, if $M = \mathrm{Gr}_{\mathbb{R}}(k,n)$ let $Q = \mathrm{O}(n)$, and if $M = \mathrm{Gr}(k,n)$ let $Q = \mathrm{U}(n)$. Then, define the p-norm extrinsic distance between two subspaces $E,F \in M$ by

$$d_p(E,F) = \min\{d_p(U,V) = \|U-V\|_p : U,V \in Q, U\mathbb{R}^k = E, V\mathbb{R}^k = F\},$$

and then it turns out the geodesic distance is given by the same procedure:

$$g_p(E,F) = \min\{g_p(U,V) = \|U-V\|_p : U,V \in Q, U\mathbb{R}^k = E, V\mathbb{R}^k = F\}.$$

1.3 Nets and Packing

Within metric spaces we will denote open balls of radius ε about a point x by $B(x,\varepsilon)$.

Definition 5 (ε-Net and ε-Covering). *Let K be a compact subset of some metric space (M,d). If $N \subset K$ is a finite subset with $d(x,N) \leq \varepsilon$ for all $x \in K$, we say N is an ε-net. If $P \subseteq K$ has $d(x,y) > \varepsilon$ for all $x,y \in P$, we say P is ε-separated.*

The cardinality of the smallest ε-net will be denoted $N(K,\varepsilon)$, and the cardinality of the largest ε-separated subset will be denoted $P(K,\varepsilon)$. It follows that

$$P(K,2\varepsilon) \leq N(K,\varepsilon) \leq P(K,\varepsilon).$$

Lemma 2. *For all $n > 1$ and $\varepsilon \leq 1$, there exists a ε-net of S^{n-1} with cardinality less than $(1 + 2/\varepsilon)^n$.*

Proof. Let $\{x_i\}_{i=1}^P$ be an ε-separated set of S^{n-1}. Then, the (open!) balls $B_i = x_i + B(0, \varepsilon/2)$ are disjoint. Let V denote the volume of the unit ball, and notice that each B_i has volume $(\varepsilon/2)^n V$. Also notice that

$$\bigcupdot_{i=1}^{P} B_i \subseteq B(0, 1 + \varepsilon/2),$$

placing the $\varepsilon/2$-balls right on the circumference. Therefore, we have

$$P\left(\frac{\varepsilon}{2}\right)^n V \leq \left(1 + \frac{\varepsilon}{2}\right)^n V,$$

or that $P \leq (1 + 2/\varepsilon)^n$. Since the largest packing never has cardinality lower than the smallest covering, we are done. □

This same technique can actually be applied to more general problems, using a set (whose volume makes sense) to cover another set (whose volume also makes sense). There are indeed tighter bounds (turns out a cardinality of $(2/\varepsilon)^n$ is achievable), but we will not need these.

Theorem 3. *Let $M = \mathrm{SO}(n), \mathrm{U}(n), \mathrm{SU}(n), \mathrm{Gr}_{\mathbb{R}}(k, n), \mathrm{Gr}(k, n)$. Suppose M is endowed with the metric g_p generated by some Schatten p-norm. Define the diametre of M by*

$$\operatorname{diam} M = \sup_{U,V \in M} g_p(U, V).$$

Then, for any $\varepsilon \in (0, \operatorname{diam} M]$ we have

$$\left(\frac{c \operatorname{diam} M}{\varepsilon}\right)^{\dim M} \leq N(M, \varepsilon) \leq \left(\frac{C \operatorname{diam} M}{\varepsilon}\right)^{\dim M}$$

for some constants c and C. Of note is that these constants are independent of n, k, p and ε. That is, they solely depend on the flavour of M.

It turns out the universal constants in the preceding theorem are unimportant, though the diametres and dimensions are, which are provided below (and for the Grassmannians, we only list it for $k \leq n/2$ since the spaces of k and $n - k$ dimensional subspaces are isometric).

M	$\dim M$	$\operatorname{diam} M$
$\mathrm{SO}(n)$	$n(n-1)/2$	$2n^{1/p}$
$\mathrm{U}(n)$	n^2	$2n^{1/p}$
$\mathrm{SU}(n)$	$n^2 - 1$	?
$\mathrm{Gr}_{\mathbb{R}}(n, k)$	$k(n-k)$	$2^{1/2}(2k)^{1/p}$
$\mathrm{Gr}(k, n)$	$2k(n-k)$	$2^{1/2}(2k)^{1/p}$

1.4 Probability Theory

Definition 6 (Subgaussian). *Let (M, d) be a compact metric space and $\{X_m\}_{m \in M}$ some stochastic process. We say the process is subgaussian if there exist some constants $A, \alpha > 0$ such that for all distinct $m, n \in M$ and $\lambda > 0$ we have*

$$\mathbb{P}(X_m - X_n > \lambda) \leq A \exp\left(-\alpha \frac{\lambda^2}{d(m, n)^2}\right).$$

Definition 7 (Centred). *Let (M, d) be a compact metric and $\{X_m\}_{m\in M}$ some stochastic process. We say the process is centred if $\mathbb{E}X_m = 0$ for all $m \in M$.*

We introduce these concepts because our later concentration of measure results can often be interpreted as saying some stochastic process is subgaussian, and the following inequality lets us numerically handle these bounds.

Proposition 1 (Dudley). *Let (M, d) be a compact metric space with radius ρ. Take some stochastic process $\{X_m\}_{m\in M}$, and suppose it is centred and subgaussian with parametres $A \geq 1/2$ and any α. Then,*

$$\mathbb{E}\sup_{m\in M} X_m \leq \frac{6}{\sqrt{\alpha}}\int_0^{\rho/2}\sqrt{1+2\log\Big(N(M,\varepsilon)\sqrt{A}\Big)}\,\mathrm{d}\varepsilon.$$

We will often work on the sphere, where we will speak of its uniform distribution. The area of an $(n-1)$-sphere is $2\pi^{n/2}/\Gamma(n/2)$ and so the uniform distribution on the sphere σ simply maps Borel sets A to $\Gamma(n/2)/2\pi^{n/2}$. There is in fact another way to attain this measure.

Proposition 2. *Let μ be the Haar probability measure on $\mathrm{O}(n)$. For any fixed $x \in S^{n-1}$ define the map*

$$\alpha\colon\ \mathrm{O}(n) \to S^{n-1} \text{ by } U \mapsto Ux.$$

Define the pushforward $\alpha_\sharp\mu = \sigma$. Then, σ is precisely the uniform distribution on the sphere.

Proof. Let A be some Borel set in S^{n-1}. By definition, we have

$$\sigma(A) = \mu(\alpha^{-1}(A)) = \mu\{U \in \mathrm{O}(n) : \alpha(U) \in A\}.$$

Without loss of generality we need only consider A to be of the form $B(x,\varepsilon) \in S^{n-1}$ for some $x \in S^{n-1}$ and $\varepsilon > 0$. Since α is a surjection we may find some $U \in \mathrm{O}(n)$ and $x_0 \in S^{n-1}$ so $Ux_0 = x$, meaning

$$B(x,\varepsilon) \cap S^{n-1} = \{Vx_0 : V \in \mathrm{O}(n) \text{ and } |Vx_0 - Ux_0| < \varepsilon\}$$

and so

$$\alpha^{-1}(B(x,\varepsilon) \cap S^{n-1}) = \{V \in \mathrm{O}(n) : |Vx_0 - Vx| < \varepsilon\}.$$

Consider as well some other $\tilde{x} \in S^{n-1}$, and find some $\tilde{U} \in \mathrm{O}(n)$ so $\tilde{x} = \tilde{U}x_0$. Witness that

$$\begin{aligned}\tilde{U}U^{-1}\alpha^{-1}(B(x,\varepsilon)\cap S^{n-1}) &= \{\tilde{U}U^{-1}V \in \mathrm{O}(n) : |Vx_0 - Vx| < \varepsilon\}\\ &= \{W \in \mathrm{O}(n) : |Wx_0 - \tilde{U}U^{-1}Ux_0| < \varepsilon\}\\ &= \{W \in \mathrm{O}(n) : |Wx_0 - \tilde{U}x_0| < \varepsilon\}\\ &= \alpha^{-1}(B(y,\varepsilon)\cap S^{n-1}).\end{aligned}$$

Since μ is translation invariant, this shows us

$$\mu(\alpha^{-1}(B(x,\varepsilon)\cap S^{n-1})) = \mu(\alpha^{-1}(B(y,\varepsilon)\cap S^{n-1})),$$

which forces σ to be the uniform distribution. □

Therefore, the Haar measure on a matrix group (the same argument works for other rotation groups, as well as when we wish to consider the complex sphere) induces a uniform distribution on a sphere. Clever readers might notice the function α is actually just an orbit of the orthogonal group acting on the sphere.

1.5 Convex Analysis

Definition 8 (Convex Body and Circled). *Let $K \subseteq \mathbb{C}^n$ or $K \subseteq \mathbb{R}^n$. We say K is a convex body if it is convex, compact, and has non-empty interior.*

If we are in $\mathbb{C}^n$ and for every $x \in K$ we have $\exp(i\vartheta)x \in K$ for all $\vartheta \in \mathbb{R}$, as well as $0 \in K$, then we say K is circled.

If we are in $\mathbb{R}^n$ and for every $x \in K$ we have $-x \in K$, as well as $0 \in K$, then we say K is symmetric.

Proposition 3. *Circled convex bodies in $\mathbb{C}^n$ exactly correspond to norms on $\mathbb{C}^n$, and symmetric convex bodies in $\mathbb{R}^n$ exactly correspond to norms on $\mathbb{R}^n$.*

Proof. We will only consider the more difficult complex case. Suppose we have a norm $\|\cdot\|_K$ over $\mathbb{C}^n$. Let K be the closed unit disk with respect to this norm, which is a circled convex body.

Now, say we have a circled convex body K. For $0 \neq x \in \mathbb{C}^n$ define

$$\|x\|_K = \inf\{t > 0 : x \in tK\},$$

which is clearly positive. At the origin, define the above to vanish. For any scalar $\lambda \in \mathbb{C}$ we have

$$\|\lambda x\|_K = \inf\{t > 0 : \lambda x \in tK\} = \inf\{t > 0 : x \in (t/\lambda)K\}.$$

Find some phase ϑ so $\exp(i\vartheta)\lambda = \eta > 0$. Since K is circled we know $(1/\eta)K = (1/\lambda)K$. Now,

$$\inf\{t > 0 : x \in (t/\lambda)K\} = \inf\{t > 0 : x \in (t/\eta)K\} = \eta\|x\|_K.$$

Since $\eta = |\lambda|$, we have homogeneity. For a second $0 \neq y \in \mathbb{C}^n$, let $s > 0$ be such that $x/s \in K$, with t defined similarly for y. By convexity

$$\frac{s}{s+t}\frac{x}{s} + \frac{t}{s+t}\frac{y}{t} = \frac{x+y}{s+t} \in K.$$

We thus have

$$\left\|\frac{x+y}{s+t}\right\|_K \leq 1 \implies \|x+y\|_K \leq s+t \leq \|x\|_K + \|y\|_K,$$

the desired triangle inequality. Altogether we see $\|\cdot\|_K$ is a norm. □

This norm is closely related to the Minkowski functional. Actually, it is precisely the Minkowski functional, and it turns out being a circled convex body is the necessary and sufficient condition for the functional to be a norm.

Definition 9 (Inradius). *Define the inradius of a circled convex body K by*

$$\operatorname{inrad} K = \sup\{r > 0 : B(0, r) \subseteq K\},$$

where $B(0, r)$ is a ball of radius r centred at the origin.

Definition 10 (Mean Width). *Let $K \subseteq \mathbb{C}^n$ be a circled convex body and σ be the uniform probability distribution on S^{n-1}. Define the mean width of K by*

$$w(K) = \int_{S^{n-1}} \|x\|_K \, \mathrm{d}\sigma(x).$$

Proposition 4. *For all circled convex bodies K, $w(K)\,\mathrm{inrad}(K) \le 1$.*

Proof. We will first provide an alternative definition of the inradius. First, let $r > 0$ be such that $B(0,r) \subseteq K$. This means that for all $x \in \mathbb{C}^n$ such that $|x| < 1$ we have $\|x\|_K \le 1$. Now, let $0 \ne x \in \mathbb{C}^n$ be arbitrary, and $\varepsilon > 0$. Then,

$$\left|\frac{xr}{(1+\varepsilon)|x|}\right| < r \text{ and so } \left\|\frac{xr}{(1+\varepsilon)|x|}\right\|_K \le 1.$$

Rearranging this we get $r/(1+\varepsilon) \le |x|/\|x\|_K$. Since ε was arbitrary, this means $|x|/\|x\|_K \ge r$.

On the other hand, say we have some r such that $|x|/\|x\|_K \ge r$ for all $0 \ne x \in \mathbb{C}^n$. Then, for all $0 \ne x \in B(0,r)$ we have

$$r \le \frac{|x|}{\|x\|_K} < \frac{r}{\|x\|_K},$$

and thus $\|x\|_K \le r$. In particular, $x \in K$. So,

$$\mathrm{inrad}\, K = \sup\left\{r > 0 : \frac{|x|}{\|x\|_K} \ge r \text{ for all } x \ne 0\right\}.$$

From here, we observe that if r is such that $|x|/\|x\|_K \ge r$ for all $0 \ne x \in \mathbb{C}^n$, this will in particular hold for $x \in S^{n-1}$ where $1/\|x\|_K \ge r$. Therefore,

$$\frac{1}{\sup\{\|y\|_K : y \in S^{n-1}\}} \ge \mathrm{inrad}\, K.$$

Denote the supremum above by L. Going the other way, for all $0 \ne x \in \mathbb{C}^{n-1}$ we have

$$\frac{|x|}{\|x\|_K} = \frac{1}{\left\|\frac{x}{|x|}\right\|_K} \ge \frac{1}{L}$$

Optimizing over all x, we see that the inradius and $1/L$ are actually equal. Equivalently,

$$\left(\sup\{\|y\|_K : y \in S^{n-1}\}\right) \mathrm{inrad}\, K = 1.$$

However, it is clear to see that $L \ge w(K)$, giving the desired inequality. □

Lemma 3. *Let $K \subseteq \mathbb{C}^n$ be a circled convex body, and define*

$$L = \sup\{\|y\|_K : y \in S^{n-1}\}.$$

Then, $\|\cdot\|_K$ is L-Lipschitz.

Proof. Consider the function

$$f : S^{n-1}_{\mathbb{C}} \to \mathbb{R} \text{ by } x \mapsto \|x\|_K$$

and define

$$b = \sup\{f(x) : x \in S^{n-1}_{\mathbb{C}}\}.$$

Expanding the definition of the norm, we get

$$L = \sup_{x \in S_{\mathbb{C}}^{n-1}} \inf\{t > 0 : x \in tK\} = \inf\{t > 0 : B(0,1) \subseteq tK\}.$$

Now, take $x, y \in S_{\mathbb{C}}^{n-1}$ and use the reverse triangle inequality to get

$$|f(x) - f(y)| = |\|x\|_K - \|y\|_K| \leq \|x - y\|_K = \inf\{t > 0 : x - y \in tK\}.$$

We divide through both sides and normalize this to unit vector, finding

$$\inf\{t > 0 : x - y \in tK\} = \inf\left\{t > 0 : \frac{x - y}{|x - y|} \in \frac{t}{|x - y|}K\right\} = L|x - y|.$$

That is, f is L-Lipschitz. □

1.6 Concentration of Lipschitz Functions

Definition 11 (Circled). *A function $f : S^{n-1} \to \mathbb{R}$ is circled if $f(x) = f(e^{i\vartheta}x)$ for all $t \in \mathbb{R}$ and $x \in S^{n-1}$.*

Definition 12 (Central Value). *For a random variable X we say m is a central value if $m = \mathbb{E}X$ or $1/4 \leq \mathbb{P}(X \leq m) \leq 3/4$.*

The specific choice of quantiles above, m being between the $1/4$- and $3/4$-quantiles, is nothing special. Raising or lowering the quantiles merely affects various constants throughout the following theorems. The choice is arbitrary.

Lemma 4 (Lévy). *Let $f : S^{n-1} \to \mathbb{R}$ be an L-Lipschitz function and σ the uniform probability measure on S^{n-1}. With $\mu = \mathbb{E}_{x\sim\sigma} f(x)$, for any $\varepsilon > 0$ we have*

$$\mathbb{P}(f \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{4L^2}\right)$$

and $|m - \mu| \leq n^{-1/2}\sqrt{2\log 2}$ for any central value m of f.

We are not going to prove the above theorem. It is actually not too difficult to show for if the central value is the median, but is in general more involved outside of that. We already doing enough work to prove Dvoretzky's theorem alone, and so will concede on this point.

Lemma 5. *Let $f : S_{\mathbb{C}}^{n-1} \to \mathbb{R}$ be a 1-Lipschitz circled function and $U \in \mathrm{SU}(n)$ be a Haar-distributed unitary. Then, we have absolute constants A, C so for any distinct $x, y \in S^{n-1}$ and $\lambda > 0$ we have*

$$\mathbb{P}(f(Ux) - f(Uy) > \lambda) \leq \exp\left(-\frac{(n-1)\lambda^2}{2|x - y|^2}\right).$$

Proof. Start by finding ϑ so $\langle x, e^{i\vartheta}y\rangle \geq 0$, and without loss of generality set $y \leftarrow e^{i\vartheta}y$, which we may due as U is linear and f is circled. We will also assume $y \neq x$.

Note that we now have

$$\langle x - y, x + y\rangle = \langle x, x\rangle - \langle y, y\rangle = 0,$$

since $|x| = 1 = |y|$ and $\langle x, y\rangle \in \mathbb{R}$. Define now $z = (x + y)/2$ and $w = (x - y)/2$. Set $\beta = |w|$ and define $w' = w/\beta$. Take the outcome $u = Uz$, and note that due to PROPOSITION 2 Uw' has the uniform distribution σ on $S_{u^\perp} = S_{\mathbb{C}}^{n-1} \cap u^\perp$ when conditioned on u (since U is an isometry).

We have

$$Ux = u + \beta U w' \text{ and } Uy = u - \beta U w'$$

by construction. So, define

$$f_u \colon S_{u^\perp} \longrightarrow \mathbb{R} \text{ by } v \mapsto f(u + \beta v) - f(u - \beta v),$$

and witness f_u has the same (conditional!) distribution as $f(Ux) - f(Uy)$.

Recalling f is 1-Lipschitz, we see

$$\begin{aligned} |f_u(a) - f_u(b)| &= |f(u + \beta a) - f(u - \beta a) - f(u + \beta b) + f(u - \beta b)| \\ &\leq |f(u + \beta a) - f(u + \beta b)| + |f(u - \beta a) - f(u - \beta b)| \\ &\leq |u + \beta a - u - \beta b| + |u - \beta a - u + \beta b| \\ &= 2\beta|a - b|, \end{aligned}$$

or that f_u is 2β-Lipschitz. Since $S_{u^\perp} = -S_{u^\perp}$ we have

$$\begin{aligned} \mathbb{E}_{v\sim\sigma} f_u(v) &= \int_{S_{u^\perp}} f_u(u + \beta v) - f(u - \beta v)\,\mathrm{d}\sigma(v) \\ &= \int_{S_{u^\perp}} f_u(u + \beta v)\,\mathrm{d}\sigma(v) - \int_{-S_{u^\perp}} f_u(u - \beta v)\,\mathrm{d}\sigma(v) \\ &= \int_{S_{u^\perp}} f_u(u + \beta v)\,\mathrm{d}\sigma(v) - \int_{S_{u^\perp}} f_u(u + \beta v)\,\mathrm{d}\sigma(v) \\ &= 0. \end{aligned}$$

We then apply Lévy's LEMMA 4 (and note that $S^{n-1}_{\mathbb{C}} \cong S^{2n-2}$, so $S_{u^\perp} \cong S^{2n-3}$) and get

$$\mathbb{P}(f(Ux) - f(Uy) > \lambda) \leq \exp\left(-\frac{(2n-2)\lambda^2}{8\beta^2}\right),$$

as a probability conditional on u. We may simplify the fraction to the desired form as $\beta = |x - y|/2$. We also note that although conditioning on u was necessary for computing expectations for f_u, the distribution of $f(Ux) - f(Uy)$ has no material dependence on this outcome. So, the above probability holds unconditionally too. □

The following theorem is a deep result in Riemannian geometry, the Gromov-Bishop comparison theorem. We will not prove it here, though we will make great use of it.

Theorem 4 (Concentration of Measure). *Take $M = \mathrm{SO}(n), \mathrm{U}(n), \mathrm{SU}(n), \mathrm{Gr}_{\mathbb{R}}(k, n), \mathrm{Gr}(k, n)$ as a metric space with respect to g_2 and let $f \colon M \longrightarrow \mathbb{R}$ be 1-Lipschitz and $\mu = \mathbb{E}f$ be the expectation with respect to the Haar measure. Then, for all $t > 0$*

$$\mathbb{P}(f > \mu + t) \leq \exp(-\lambda t^2),$$

where the values of C are as follows:

M	λ
$\mathrm{SO}(n)$	$(n-1)/8$
$\mathrm{U}(n)$	$n/12$
$\mathrm{SU}(n)$	$n/4$
$\mathrm{Gr}_{\mathbb{R}}(n, k)$	$(n-2)/4$
$\mathrm{Gr}(k, n)$	$n/2$

We are now ready to prove a restated version of the Dvoretzky-Milman theorem, from which our desired version (sometimes called the geometric or the tangible version) will be easily derived.

Definition 13 (Oscillation)**.** *Let $f\colon X \longrightarrow \mathbb{R}$ and $A \subseteq X$. For $\mu \in \mathbb{R}$, we define the oscillation of f about μ on A by*

$$\operatorname{osc}(f, A, \mu) = \sup_{x \in A} |f(x) - \mu|.$$

Theorem 5. *There exist universal constants $c, c' > 0$ such that the following always holds. Let $f\colon S_{\mathbb{C}}^{n-1} \longrightarrow \mathbb{R}$ be an L-Lipschitz circled function and μ be any central value of f with respect to the uniform distribution on the sphere σ. Let $0 < \varepsilon < 1$ and $k \leq cn\varepsilon^2/L^2$. Then, for any random k-dimensional subspace $E \subseteq \mathbb{C}^n$ we have*

$$\mathbb{P}(\operatorname{osc}(f, S_E, \mu) \leq \varepsilon) > 1 - \exp(-c'n\varepsilon^2),$$

where $S_E = S_{\mathbb{C}}^{n-1} \cap E$.

Proof. Without loss of generality we suppose $\mu = 0$ and suppose $L = 1$. Now, take any $U, V \in \mathrm{SU}(n)$ and some $x \in S_{\mathbb{C}^k} := S_{\mathbb{C}}^{n-1} \cap \mathbb{C}^k$. As f is 1-Lipschitz and $|x| = 1$, we have

$$|f(Ux) - f(Vx)| \leq |Ux - Vx| \leq \|U - V\|_{\mathrm{op}} \leq \|U - V\|_{\mathrm{HS}}$$

where the last inequality is due to LEMMA 1. However, as the Hilbert-Schmidt norm is merely a special case of the Schatten 2-norm, we use THEOREM 2 to get

$$\|U - V\|_{\mathrm{HS}} \leq g_2(U, V),$$

where g_2 is the geodesic distance with respect to the 2-norm.

Now, let $U \in \mathrm{SU}(n)$ be a Haar-distributed unitary and set $E = U(\mathbb{C}^k)$. Define

$$F\colon \mathrm{SU}(n) \longrightarrow \mathbb{R} \text{ by } U \mapsto \sup_{S_E} |f| = \sup_{x \in S_{\mathbb{C}^k}} |f(Ux)|.$$

Our previous inequality chain, along with the reverse triangle equality and monotonic increase of the absolute value, shows

$$|F(U) - F(V)| = \left| \sup_{x \in S_{\mathbb{C}^k}} |f(Ux)| - \sup_{y \in S_{\mathbb{C}^k}} |f(Vy)| \right| \leq \sup_{x, y \in S_{\mathbb{C}^k}} |f(Ux) - f(Uy)| \leq g_2(U, V).$$

Importantly, this shows F is 1-Lipschitz with respect to the metric g_2. Applying concentration of measure THEOREM 4 we get

$$\mathbb{P}(F \geq \mathbb{E}F + t) \leq \exp\left(-\frac{nt^2}{4}\right).$$

Now, define the process $\{X_s\}_{s \in S_{\mathbb{C}^k}}$ by $X_s = |f(Us)|$, which by LEMMA 5 we know is subgaussian with $A = 1$ and $\alpha = (n-1)/2$. We then apply Dudley's inequality from PROPOSITION 1, getting

$$\mathbb{E} \sup_{s \in S_{\mathbb{C}^k}} X_s \leq \sup_{s \in S_{\mathbb{C}^k}} \mathbb{E}X_s + \frac{6\sqrt{2}}{\sqrt{n-1}} \int_0^{1/2} \sqrt{1 + 2\log N(S_{\mathbb{C}^k}, \delta)}\, \mathrm{d}\delta.$$

We note the extra term on the right side not present in the original statement of the inequality – that is because we do not know if X_s is centred. For this reason, we must actually apply the inequality to the centred $X_s - \mathbb{E}X_s$. The expectations are taken with respect to the Haar measure.

It follows from PROPOSITION 2 that

$$\mathbb{E}_{x\sim\sigma}|f(x)| = \int_{S_{\mathbb{C}^k}} |f(x)|\,\mathrm{d}\sigma(x) = \int_{\mathrm{U}(n)} |f(Us)|\,\mathrm{d}\,\mathrm{Haar}(U) = \mathbb{E}_{U\sim\mathrm{Haar}}|f(Us)| = \mathbb{E}X_s$$

for any fixed s. Therefore, from LEMMA 4 we know $\mathbb{E}X_s \le \sqrt{2\log 2}/\sqrt{2n}$, where the doubling of n is to adjust to the real dimension. From LEMMA 2 we have $N(S_{\mathbb{C}^k}, \delta) \le (1+2/\delta)^{2k}$, and applying the inequality $\sqrt{1+t} \le 1+\sqrt{t}$ for $t > 0$ on the integrand we get

$$\mathbb{E}\sup_{s\in S_{\mathbb{C}^k}} X_s \le \frac{\sqrt{\log 2}}{\sqrt{n}} + \frac{3\sqrt{2}}{\sqrt{n-1}} + \frac{12\sqrt{2k}}{\sqrt{n-1}}\int_0^{1/2}\sqrt{\log(1+2/\delta)}\,\mathrm{d}\delta.$$

The rest of the proof is uninteresting. We simply use numerics to further bound the above, obtaining something on the order of

$$\mathbb{E}\sup_{s\in S_{\mathbb{C}^k}} X_s \lessapprox \sqrt{\frac{k}{n}},$$

and it turns out there is indeed some number c so $k \le cn\varepsilon^2$ implies $\mathbb{E}\sup X_s < \varepsilon/2$. But, $\mathbb{E}\sup X_s = \mathbb{E}F$, and so to finish we recall our previous work and establish

$$\exp\left(-\frac{nt^2}{4}\right) \ge \mathbb{P}(F \ge \mathbb{E}F + t) \le \mathbb{P}\left(F \ge \frac{\varepsilon}{2} + t\right).$$

Taking $t = \varepsilon/2$, we are done. □

There are many slight variations of this proof, and most of them stem on having some extra conditions or tighter bounds (such as on the covering numbers) to better wrangle the constants involved. But if we only care about the existence of the constants, what we did suffices.

1.7 Dvoretzky-Milman

Definition 14 (Dvoretzky Dimension). *The Dvoretzky dimension of a circled convex body $K \subseteq \mathbb{C}^n$ is*

$$k_*(K) = (w(K)\,\mathrm{inrad}(K))^2 n.$$

Theorem 6 (Dvoretzky-Milman). *There exist universal constants $c, c' > 0$ such that the following always holds. Take $K \subseteq \mathbb{C}^n$ to be a circled convex body. Fix some $\varepsilon \in (0,1]$ and define $k = c\varepsilon^2 k_*(K)$. Then, any random subspace $E \subseteq \mathbb{C}^n$ with $\dim E \le k$ satisfies*

$$(1-\varepsilon)w(K)|x| \le \|x\|_K \le (1+\varepsilon)w(K)|x|$$

for all $x \in E$ with probability at least $1 - \exp(-c'\varepsilon^2 k_*(K))$. Here, E is chosen with respect to the Haar probability measure on the Grassmannian.

Proof. Define $S_E = S_{\mathbb{C}}^{n-1} \cap E$ and consider the function

$$f : S_E \to \mathbb{R} \text{ by } x \mapsto \|x\|_K.$$

The proof for LEMMA 3 goes through identically, telling us that f is L-Lipschitz, where

$$L = \sup\{f(x) : x \in S_{\mathbb{C}}^{n-1}\}.$$

Also, note that f is clearly circled.

Recall as well from the proof of PROPOSITION 4 that $L = 1/\operatorname{inrad} K$, allowing us to rewrite the Dvoretzky dimension as

$$k_*(K) = \left(\frac{w(K)}{L}\right)^2 n.$$

We then defer to THEOREM 5, stealing its magical constants c and c' to probabilistically bound $\operatorname{osc}(f, S_E, w(K)) \leq \varepsilon w(K)$ as being no less likely than $1 - \exp(-c'n(\varepsilon w(K))^2)$ for appropriate ε.

Recalling the definition of oscillation, we have with high probability that

$$\sup_{x \in S_E} |\|x\|_K - w(K)| \leq \varepsilon w(K).$$

We may escape unit sphere by normalizing any $x \in E$, and so we are done. □

One way of interpreting this theorem is by saying that convex bodies are "approximately Euclidean" in high enough dimensions, since the convex body norm of random subspaces is typically within an epsilon of error of the standard Euclidean norm (after scaling by the mean width of the body).

2 Quantum Channels

Quantum channels are lines of communication between two parties – Alice and Bob – in the form of transmission of quantum states. A fundamental question is how much information they may transmit, which is precluded by a precise definition of information in the first.

2.1 Classical Information Theory

The definition of information for non-quantum (i.e. classical) channels was answered by Shannon in [Sha48], his seminal work. We will present this section using CHAPTER 10.1 from [Pre16] as reference. Throughout, we will let X and Y be finitely-supported discrete random variables.

Definition 15 (Shannon Entropy). *Let X have mass p. Its Shannon entropy is*

$$H(X) = -\sum_{x \in \operatorname{supp} X} p(x) \log_2 p(x).$$

The base of the logarithm does not really matter, but in the classical case it is convenient to use base 2 since this corresponds to bits. Entropy was introduced to study how data may be compressed. Also, two facts should be obvious from the definition. First, $H(X) \leq 1$ always, and second, this inequality is saturated if and only if X is uniformly distributed.

Definition 16 (δ-Typical Sequence). *Let $\{X_i\}_{i=1}^n$ i.i.d copies of X. Denote by $p(x_1, \ldots, x_n)$ their joint distribution. We say the sample $(x_1, \ldots, x_n)$ is δ-typical for $\delta > 0$ if*

$$H(X) - \delta \leq -\frac{1}{n} \log_2 p(x_1, \ldots, x_n) \leq H(X) + \delta.$$

Proposition 5. *For any $\varepsilon, \delta > 0$, there exists some sufficiently large n such that any n-symbol sequence is δ-typical with probability no less than $1 - \varepsilon$.*

Proof. We appeal to the (weak) law of large numbers. Let $\{X_i\}_{i=1}^n$ describe our random n-symbol sequence of X. The law tells us that there is some sufficiently large n so

$$\mathbb{P}\left(\left|\mathbb{E}X - \frac{1}{n}\sum_{i=1}^n X_i\right| \leq \delta\right) \geq 1 - \varepsilon.$$

The event may be rewritten as

$$\mathbb{E}X - \delta \leq \frac{1}{n}\sum_{i=1}^n X_i \leq \mathbb{E}X + \delta.$$

Let us now instead consider $Y = -\log_2 p(X)$. Note that $H(X) = \mathbb{E}Y$, and applying the weak law to Y we are done, since

$$\sum_{i=1}^n Y_i = -\sum_{i=1}^n \log_2 p(X_i) = -\log_2 p(X_1, \ldots, X_n)$$

by independence. □

Definition 17 (Rate)**.** *A compression rate R is associated with a protocol compressing some messages comprising n-symbols to nR-symbol codewords.*

We say R is achievable if there is asymptotically vanishing error (i.e. the probability of decompression error goes to 0 as $n \longrightarrow \infty$).

Theorem 7 (Shannon's Source Coding)**.** *All compression rates greater than $H(X)$ are achievable.*

Proof. Let $\delta, \varepsilon > 0$. Due to PROPOSITION 5, we can find some sufficiently large n such that an n-symbol message $(x_1, \ldots, x_n)$ satisfies

$$2^{-n(H(X)+\delta)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\delta)}$$

with probability no less than $1 - \varepsilon$. Respectively denote the left-hand and right-hand sides of this inequality by $p_{\min}$ and $p_{\max}$.

Now, let $\mathcal{N}$ be the set of these δ-typical n-symbol messages. By definition, we have

$$|\mathcal{N}|p_{\min} \leq \sum_{x \in \mathcal{N}} p(x) \leq 1 \text{ and } |\mathcal{N}| \geq \sum_{x \in \mathcal{N}} p(x) \geq 1 - \varepsilon.$$

Returning to our first inequality, we thus have

$$(1 - \varepsilon)2^{n(H(X)-\delta)} \leq |\mathcal{N}| \leq 2^{n(H(X)+\delta)}.$$

Therefore, we need only encode messages using $n(H(X) + \delta)$ bits to decode any message in $\mathcal{N}$ with error probability less than ε. Taking $\delta \longrightarrow 0$ and $\varepsilon \longrightarrow 0$, we see a rate of $H(X)$ is indeed achievable asymptotically.

To show this is the best we can do, let $\gamma > 0$ and suppose we had a rate of $H(X) - \gamma$ for an n-symbol message. Let p_{success} denote the probability of successfully decompressing our original message.

Reusing our work on achievability, we know that a δ-typical message occurs with probability no greater than $2^{-n(H(X)-\delta)}$, and our code can uniquely identify $2^{n(H(X)-\gamma)}$ of these. Thus,

$$p_{\text{success}} \leq 2^{n(H(X)-\gamma)}2^{-n(H(X)-\delta)} + q = 2^{-n(\gamma-\delta)} + \varepsilon,$$

where the error ε is the likelihood of an atypical message. This is the most optimistic scenario, the worst possible bound for such a code.

Since ε and δ are arbitrarily small, we see that as $n \to \infty$ we have $p_{\text{success}} \to 0$ necessarily, and so such a code is not achievable. □

We do not really care about information compression here. Rather, we care about information transmission. However, it turns out more-or-less the same principles can be applied to handle this problem.

Definition 18 (Channel). *A channel allows Alice to receive some message in X and send it to Bob, who receives a message in Y. The channel between Alice and Bob is characterized by the conditional distribution $Y|X$.*

The reason we characterize the channel by the conditional distribution is because Alice, in principle, gets to choose what she sends to Bob. Even though she gets some message from X, she may wish to encode it in some manner and send through a slightly different distribution X'. In this case, the channel functionally is $Y|X'$.

Definition 19 (Jointly δ-Typical). *Let $\delta, \varepsilon > 0$, and let $\{X_iY_i\}_{i=1}^{n}\}$ be i.i.d. according to $XY \sim p_{XY}$. We say an n-symbol message $(x_1y_1, \ldots, x_ny_n)$ is jointly δ-typical if*

$$\begin{aligned} 2^{-n(H(X)+\delta)} &\leq p_X(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\delta)} \\ 2^{-n(H(Y)+\delta)} &\leq p_Y(y_1, \ldots, y_n) \leq 2^{-n(H(Y)-\delta)} \\ 2^{-n(H(XY)+\delta)} &\leq p_{XY}(x_1y_1, \ldots, x_ny_n) \leq 2^{-n(H(XY)-\delta)}. \end{aligned}$$

Note that above we write p_{XY} but really mean $p_{X^nY^n}$. We will also sometimes write $(x^n, y^n) = (x_1y_1, \ldots, x_ny_n)$.

Essentially, joint typicality means typicality with respect to the joint distribution as well as the marginals. Exactly the same argument as for one variable will show most messages are jointly typical in the limit.

Things are interesting when X and Y are not independent, since then knowledge of one ought to inform us somewhat of the other. This is where the notion of compression returns. For a jointly typical message (x^n, y^n), we see

$$p(x^n|y^n) = \frac{p_{XY}(x^n, y^n)}{p_Y(y^n)} \leq \frac{2^{-n(H(XY)-\delta)}}{2^{-n(H(Y)+\delta)}} = 2^{-n(H(XY)-H(Y)-2\delta)}$$

and

$$p(x^n|y^n) = \frac{p_{XY}(x^n, y^n)}{p_Y(y^n)} \geq \frac{2^{-n(H(XY)+\delta)}}{2^{-n(H(Y)-\delta)}} = 2^{-n(H(XY)-H(Y)+2\delta)}.$$

This prompts the following definitions.

Definition 20 (Conditional Entropy). *The conditional entropy of X given Y is*

$$H(X|Y) = H(XY) - H(Y).$$

Definition 21 (Mutual Information). *The mutual information between X and Y is*

$$I(X:Y) = H(X) - H(X|Y).$$

Conditional entropy answers the following question: If I know Y, what is my remaining ignorance about X? Well, in the context of jointly δ-typical sequences, we see that once we know Y we can achieve a compression rate of $H(X|Y)$, asymptotically.

The inverse is the answered by mutual information: If I know Y, by how much has my compression burden been reduced? It is precisely the difference between $H(X)$ and $H(X|Y)$, since these are the best achievable rates.

Definition 22 (Rate and Capacity). *A communication rate R is associated with a protocol encoding n-symbol output messages in Y to nR-symbol codewords in X.*

We say R is achievable if there is asymptotically vanishing error (i.e. the probability of a decoding error goes to 0 as $n \longrightarrow \infty$).

The capacity of a channel C is the supremum over all achievable rates R with respect to all codeword distributions X.

Theorem 8 (Shannon's Noisy Coding). *Channel capacity is given by* $\sup_X I(X:Y)$.

Proof. Let $\delta, \varepsilon > 0$ and find some n from PROPOSITION 5 such that n-symbol messages from XY are jointly δ-typical. Now, take such as sample (x, y), which we interpret as Alice drawing an n-symbol message from X and sending it through the channel for Bob to get an n-symbol message y from Y. Henceforth, y is fixed. Also note that we mean $x = (x_1, \ldots, x_n)$, and likewise for y. With high probability (with respect to ε), x and y are jointly δ-typical – we will suppose that they are.

Using the same logic as in the proof of THEOREM 7, we let $\mathcal{N}_{XY}$ be the set of jointly δ-typical n-symbol messages and have $|\mathcal{N}| \leq 2^{n(H(XY)+\delta)}$. We similarly define $\mathcal{N}_X$ and $\mathcal{N}_Y$, and know that if $x' \in \mathcal{N}_X$ then $p_X(x') \leq 2^{-n(H(X)-\delta)}$. The corresponding bound also holds for our fixed y (since joint typicality implies individual typicality).

Let $\mathcal{N}_{Xy}$ be the set of jointly δ-typical n-symbol messages of the form (x', y) (i.e. with our fixed y). Suppose now we sample an n-symbol message z of X. As y is fixed prior to our sampling, we have $p_{XY}(z, y) = p_X(z)p_Y(y)$. The probability that z and y are jointly δ-typical is thus bounded above by

$$\sum_{(x',y)\in\mathcal{N}_{Xy}} p_{XY}(x', y) = \sum_{(x',y)\in\mathcal{N}_{Xy}} p_X(x')p_Y(y) \leq |\mathcal{N}_{Xy}| 2^{-n(H(X)-\delta)} 2^{-n(H(Y)-\delta)}.$$

Clearly, $|\mathcal{N}_{Xy}| \leq |\mathcal{N}_{XY}|$, and so we further bound the probability by

$$2^{n(H(XY)+\delta)} 2^{-n(H(X)-\delta)} 2^{-n(H(Y)-\delta)} = 2^{-n(I(X:Y)-3\delta)}.$$

With this in mind, we establish our code with rate R. Let us create 2^{nR} codewords, n-symbol messages sampled from X^n. Alice sends through a codeword x and Bob receives some output y. He then searches among all codewords x' which are jointly δ-typical with y. The odds of a decoding error – that he mistakenly finds some $x' \neq x$ which is jointly δ-typical too – is bounded above by

$$2^{nR} 2^{-n(I(X:Y)-3\delta)} = 2^{n(R-I(X:Y)+3\delta)}.$$

There are two ways this may fail. First, that our sampled X^n is not δ-typical, which is controlled by ε and may be made arbitrarily small. Second, that we have a collision and thus a decoding error, which is controlled by δ and R. We see that for any $\gamma > 0$, a rate $R = I(X : Y) - \gamma$ will have vanishing error in the latter case as $\delta \to 0$.

There is a subtly – we have only showed that the average decoding error among our protocol vanishes. We require that the protocol itself has vanishing error. Specifically, label our set of codewords $Q = \{x_i\}_{i=1}^{2^{nR}}$ and denote by q_i the probability of a decoding error when sending the i-th codeword. The above work establishes that for any $\eta > 0$, there is some sufficiently large n such that

$$\frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} q_i \leq \eta.$$

Let $Q_{2\eta}$ denote the codewords x_i for which $q_i \geq 2\varepsilon$. Witness that

$$\eta \geq \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} q_i \geq \frac{1}{2^{nR}} \sum_{i \in Q_{2\eta}} \geq \frac{1}{2^{nR}} |Q_{2\eta}| 2\eta.$$

Rearranging, $|Q_{2\eta}| \leq 2^{nR-1}$. Therefore, the new set of codewords $Q' = Q \setminus Q_{2\eta}$ satisfies $q_i < 2\eta$ for all remaining $x_i \in Q_i$ (and we throw away at most half of the originals). This inequality also shows our protocol with Q' attains a rate $R - 1/n$. Taking $n \to \infty$, we see the rate R is still attainable asymptotically. □

2.2 Quantum Shannon Theory

We do not title this part "quantum information theory," for that would be too general. We are specifically concerned with adapting the concepts of classical information theory, such as channels and capacities, to quantum settings.

The set of all quantum states on $\mathcal{H}$ will be $S(\mathcal{H})$, and a generic one will be ρ. Pure states will be identified with unit vectors x (or $|x\rangle$ in Dirac notation), with us denoting the unit ball by $B(\mathcal{H})$. By $|x\rangle\langle x|$ be mean the orthogonal projection onto the span of x. We will make use of use of the Hilbert spaces $\mathcal{A}, \mathcal{B}, \mathcal{E}$ – Alice, Bob, and the environment. The identity operator is $\mathbb{1}$, with a subscript if necessary (denoting the space or the dimension).

Definition 23 (Trace-Preserving). *A map $U \colon S(\mathcal{A}) \to S(\mathcal{B})$ is trace-preserving if* $\operatorname{tr} U(\rho) = 1$ *for all states ρ.*

Definition 24 (Completely-Positive). *A map $U \colon S(\mathcal{A}) \to S(\mathcal{B})$ is completely-positive if $U \otimes \mathbb{1}_n \geq 0$ for all integers $n \geq 0$. Note $n = 0$ means U alone is positive.*

Definition 25 (Channel). *A quantum channel is a trace-preserving completely-positive (CPTP) map $\mathcal{N} \colon S(\mathcal{A}) \to S(\mathcal{B})$.*

Sometimes these are called CPTP maps. We will just call them channels. These two conditions are the minimal conditions necessary to reasonable describe sending a state from one place to another. We want to still have states, so the trace must be preserved. And we want not only positivity in isolation, but also positivity should we append some extra operators acting on ancillary spaces.

Often, channels are given in terms of their Kraus representation, but for us the Steinspring form is not only more operational, but also more mathematically useful.

Theorem 9 (Steinspring Dilation). *Let $\mathcal{N} \colon S(\mathcal{A}) \to S(\mathcal{B})$ be a channel. Then, there exists some larger Hilbert space $\mathcal{E}$ and an isometry $V \colon \mathcal{A} \to \mathcal{B} \otimes \mathcal{E}$ such that*

$$\mathcal{N}(\rho) = \operatorname{tr}_{\mathcal{E}}(V \rho V^*).$$

What this tells us that a quantum channel is not merely Alice sending a state to Bob. Instead, it is Alice sending a state to Bob, with some noise involved – the state interacts somehow the environment. Of course, we do not have access to the environment, so we must trace it out.

There are multiple reasonable notions of "capacity of a quantum channel". We will discuss only their classical capacity – if Alice and Bob wish to exchange classical information (such as some string), and have access to a quantum channel, how can they best do that?

Definition 26 (Positive Operator-Valued Measurement). *A positive operator-valued measurement (POVM) comprises a set $\{E_i\}$ of positive operators such that $\mathbb{1} = \sum_i E_i$.*

Say Alice wishes to send some message given by X. To each element of its support x she may associate some state $\rho(x)$, which she sends through her quantum channel $\mathcal{N}$ with probability $p_X(x)$. Equivalently, Bob is receiving the ensemble $\rho(X) = \sum_i p(x)\rho(x)$. We will denote the set of all ensembles by $\mathscr{E}$, writing ensembles $\varepsilon \in \mathscr{E}$ as tuples $\varepsilon = \{(p_i, \rho_i)\}$ representing the state $\rho(\varepsilon) = \sum_i p_i \rho_i$.

On Bob's end, to recover the classical information he may construct a POVM E, where each element $E(y)$ is associated with some classical information y. By performing the measurement $E\mathcal{N}(\rho(X))$, we interpret the analogue of Bob's classical distribution Y to be the distribution whose outcomes are y with probabilities $p_{Y_E}(y) = \operatorname{tr}(E(y)\mathcal{N}(\rho(X)))$. The conditional distribution analogous to $Y_E | X = x$ has masses $p(y|x) = \operatorname{tr}(E(y)\mathcal{N}(\rho(x)))$.

The key observation is that Bob's distribution no longer solely depends on Alice's distribution X and the behaviour of the channel $\mathcal{N}$, unlike the classical case. Instead, Bob now gets to decide which POVM he uses, and for that reason we chose to label his distribution Y_E.

Definition 27 (Accessible Information). *With the classical communication protocol using a quantum channel as described, let $\rho(X)$ be the ensemble Alice sends to Bob. Then, the accessible information of $\rho(X)$ is*

$$I(\rho(X)) = \sup_E I(X : Y_E).$$

This quantity is not at all easy to compute. But, there is a much more tractable upper-bound.

Definition 28 (von Neumann Entropy). *The von Neumann entropy of a state ρ is $S(\rho) = -\operatorname{tr} \rho \log \rho$, as given by the functional calculus.*

Definition 29 (Holevo Chi). *For the ensemble $\rho(X) = \sum p_X(x)\rho(x)$, its Holevo chi is*

$$\chi(\rho(X)) = S(\rho(X)) - \sum_x p_X(x) S(\rho(x)).$$

The Holevo capacity of a quantum channel $\mathcal{N}$ is the largest Holevo chi of its outputs:

$$\chi(\mathcal{N}) = \sup_\rho \chi(\mathcal{N}(\rho)).$$

Theorem 10 (Holevo). *For every ensemble $\rho(X)$ and quantum channel, $I(\rho(X)) \leq \chi(\mathcal{N}(\rho(X)))$.*

This theorem was originally given in [Hol73]. The name Holevo capacity may be slightly misleading, not only because we have yet to define capacity, but since the Holevo chi of an ensemble may exceed its accessible information. However, this is just clever foreshadowing.

Definition 30 (Rate and Capacity). *A quantum channel $\mathcal{N}$ communicates classically at a rate R if associated with Bob's n-symbol measurement outcomes in Y are 2^{nR} n-symbol codewords in X (which Alice sends as n-register quantum states).*

We say a rate R is achievable if it may be accomplished with asymptotically vanishing error.

The classical capacity of the channel $C(\mathcal{N})$ is the supremum over achievable rates with respect to all input distributions X.

This is exactly the definition we expect. However, the formula for the capacity of quantum channel is significantly less pleasant than for classical channels.

Theorem 11 (Holevo-Schumacher-Westmoreland). *For a quantum channel $\mathcal{N}$, its classical capacity is the regularized Holevo capacity:*

$$C(\mathcal{N}) = \lim_{n\to\infty} \frac{1}{n}\chi(\mathcal{N}^{\otimes n}).$$

This result was originally given in [SW97]. Though remarkable that the Holevo chi returns, the regularization makes it entirely computationally intractable. It would be wonderful if the Holevo capacity were additive, meaning that all quantum channels $\mathcal{N}, \mathcal{S}$ satisfy

$$\chi(\mathcal{N} \otimes \mathcal{S}) = \chi(\mathcal{N}) + \chi(\mathcal{S}).$$

A partial answer was also given in [SW97], showing the quantity is additive if Alice is restricted to product states – i.e. she can only feed $\mathcal{N}^{\otimes n}$ states of the form $\rho_1 \otimes \cdots \otimes \rho_n$. However, if Alice may input entangled states, the question seems much harder.

In this section we are considering channels of finite dimension. In principle, nothing is wrong with infinite-dimensional channels – completely-positive operators make perfect sense, as do trace-preserving maps (provided trace is carefully defined). Even the classical capacity of such channels has been studied [HS13].

2.3 Minimum Output Entropy

Questions about the additivity of various quantities are abundant in quantum information theory, not restricted to just the Holevo capacity. Dvoretzky's theorem is well-suited towards handling one of these conjectures, and is the reason we looked at it in the first section.

Definition 31 (Rényi p-Entropy). *For $p \in [0, \infty]$ we define the Rényi p-entropy of a state by*

$$S_p(\rho) = \frac{1}{1-p} \log \operatorname{tr}(\rho^p).$$

The cases $p = 0$, $p = 1$, and $p = \infty$ are given by the limits.

Proposition 6. *The limit cases of p-entropy are*

- $S_0(\rho) = \log \operatorname{rank} \rho$
- $S_1(\rho) = S(\rho)$

- $S_\infty(\rho) = -\log \|\rho\|_{op}$

Proof. Assume the diagonalization $\rho = \operatorname{diag}(\lambda_i)$, so that

$$S_p(\rho) = \frac{1}{1-p} \log \sum_i \lambda_i^p.$$

Implicitly, we ignore the terms $\lambda_i = 0$ in the summation. We also assume the spectrum is finite (of cardinality N), though as ρ is a compact operator we may attain it as a limit of finite-rank ones and clean these arguments up by taking $N \to \infty$.

We now simply work through each case individually. Starting with $p = 0$, we see that $\lambda_i^p \to 1$ as $p \to 0$. We conclude by noticing the continuity of the logarithm and that $1 - p \to 1$ concurrently.

For $p = 1$, since $\lambda_i^p \to \lambda_i$ and states have unit trace, we get the indeterminate form

$$\lim_{p\to1} S_p(\rho) = \lim_{p\to1} \frac{\log \sum_i \lambda_i^p}{1-p} = \frac{\log 1}{0}.$$

Applying l'Hôpital's rule, the denominator turns into a leading factor of -1, which combined with the numerator yields

$$\lim_{p\to1} S_p(\rho) = -\lim_{p\to1} \frac{\sum_i \lambda_i^p \log \lambda_i}{\sum_i \lambda_i^p} = -\sum_i \lambda_i \log \lambda_i = S(\rho).$$

Lastly, for $p = \infty$ we recall that since ρ is self-adjoint and compact we have

$$\|\rho\|_{\text{op}} = \max\{\lambda_i\} =: \lambda_{\max}.$$

For all $p \geq 1$ we thus have the bound

$$\lambda_{\max}^p \leq \sum_i \lambda_i^p \leq N\lambda_{\max}^p$$

and sandwich the entropy, getting

$$\frac{\log \lambda_{\max}^p}{1-p} \geq S_p(\rho) \geq \frac{\log N\lambda_{\max}^p}{1-p}.$$

Note the equalities flip as we are taking the logarithm of a quantity less than one. Taking $p \to \infty$ we thus have

$$-\log \lambda_{\max} \leftarrow \frac{p \log \lambda_{\max}}{1-p} \geq S_p(\rho) \geq \frac{\log N + p \log \lambda_{\max}}{1-p} \to -\log \lambda_{\max},$$

as needed. □

We use Rényi p-entropy since it is generally nicer than von Neumann entropy. For finite $p > 1$ it is naturally related to the easily computed Schatten p-norm by $S_p(\rho) = p \log \|\rho\|_p/(1-p)$. We then hope that somehow, in the limit, we can recover statements about the $p = 1$ case. Why do we need to recover $p = 1$? We will see shortly.

Definition 32 (Minimum Output p-Entropy)**.** *The minimum output p-entropy of a channel $\mathcal{N}$ is*

$$S_p^{min}(\mathcal{N}) = \inf_{\varepsilon \in \mathscr{E}} S_p(\mathcal{N}(\rho(\varepsilon))).$$

We also need not optimize over all ensembles – it suffices to use pure states:

$$S_p^{\min}(\mathcal{N}) = \min_{x \in B(\mathcal{A})} S_p(\mathcal{N}(|x\rangle\langle x|)).$$

This is a quantity related to channels, and so it makes sense to speak of its additivity too! For two channels $\mathcal{N}, \mathcal{S}$, is it always true that

$$S_p^{\min}(\mathcal{N} \otimes \mathcal{S}) = S_p^{\min}(\mathcal{N}) + S_p^{\min}(\mathcal{S})?$$

There is a crucial connection between the minimum output entropy and the Holevo capacity.

Theorem 12 (Shor)**.** *Additivity of the Holevo capacity is equivalent to additivity of the minimum output* 1*-entropy.*

This result is from [Sho04], and it in fact established the equivalence of several other additivity conjectures (such as for the entanglement of formation). Also recall that 1-entropy is the von Neumann entropy.

We would also like to note that the following inequalities always hold:

$$S_p^{\min}(\mathcal{N} \otimes \mathcal{S}) \leq S_p^{\min}(\mathcal{N}) + S_p^{\min}(\mathcal{S})$$

and

$$\chi(\mathcal{N} \otimes \mathcal{S}) \geq \chi(\mathcal{N}) + \chi(\mathcal{S}).$$

Therefore, the additivity conjectures ask if there are channels with (strictly!) subadditive minimum output 1-entropy or, equivalently, (strictly!) superadditive Holevo capacity.

3 Existence of Superadditive Quantum Channels

Holevo capacity is not additive. This was proven indirectly, by showing that minimum output entropy is not. The first steps were taken by [WH02], where numerical experiments showed minimum output p-entropy is false for $p > 4.79$. However, it was still long suspect that not only does additivity hold for $p = 1$, but in fact all $p \in [1, 2]$ (due to the convexity of the map $\rho \mapsto \rho^p$). However, two classes of counterexamples in [HW08] showed additivity is false for all $p > 1$. Shortly after, [Cub+08] gave numerical results of non-additivity for $p < 0.11$. In spite of all this, the von Neumann case was still believed to be true – until [Has09] settled the conjecture in the negative.

The difficulty of the problem is that constructing explicit counterexamples is very hard. In fact, they still do not exist for the $p = 1$ case, and are rare for other p. We will cover such explicit constructions later, but for now we will focus on the original promised approach – probabilistic arguments.

3.1 Channels as Subspaces

Consider a channel $\mathcal{N} \colon \mathcal{A} \to \mathcal{B}$ and apply Steinspring's THEOREM 9 to obtain its associated environment $\mathcal{E}$ and isometry $V \colon \mathcal{A} \to \mathcal{B} \otimes \mathcal{E}$. Take some pure state $|x\rangle \in \mathcal{A}$ so that

$$\mathcal{N}(|x\rangle\langle x|) = \mathrm{tr}_{\mathcal{E}}(V|x\rangle\langle x|V^*).$$

As V is an isometry, we know $V|x\rangle = |y\rangle$ is still a pure state, and thus admits some Schmidt decomposition

$$|y\rangle = \sum_i \lambda_i |e_i\rangle \otimes |f_i\rangle,$$

where $|e_i\rangle$ and $|f_i\rangle$ are respectively orthonormal bases for $\mathcal{B}$ and $\mathcal{E}$. Tracing out the environment yields

$$\mathrm{tr}_{\mathcal{E}}(|y\rangle\langle y|) = \sum_i \lambda_i^2 |e_i\rangle\langle e_i|,$$

and so

$$S_p(|x\rangle\langle x|) = \frac{1}{1-p} \log \sum_i \lambda_i^2.$$

Therefore, the Schmidt coefficients of $|x\rangle\langle x|$ entirely determine its p-entropy.

Now, V is an isometry and therefore a bijection onto its range $V(\mathcal{A}) \subseteq \mathcal{B} \otimes \mathcal{E}$. Importantly, we have an isometric isomorphism between subspaces: $\mathcal{A} \cong V(\mathcal{A})$. This allows us to uniquely identify $\mathcal{N}$ with a subspace of $\mathcal{B} \otimes \mathcal{E}$.

Furthermore, V gives a bijection between the units balls $B(\mathcal{A})$ and $B(V(\mathcal{A}))$. Recalling that minimum output p-entropy optimizes over all pure states, we have

$$S_p^{\min}(\mathcal{N}) = \min_{x \in B(\mathcal{A})} S_p(|x\rangle\langle x|) = \min_{y \in \mathcal{B}(V(\mathcal{A}))} S_p(|y\rangle\langle y|).$$

Our conclusion is that the study of the minimum output p-entropy of $\mathcal{N}$ reduces to the study of the Schmidt coefficients of pure states in $V(\mathcal{A})$. This goes the other way too – given any subspace $\mathcal{V} \subseteq \mathcal{B} \otimes \mathcal{E}$, we know that there is some isometric isomorphism $V\colon \mathbb{C}^{\dim \mathcal{V}} \to \mathcal{V}$, and that $\rho \mapsto \mathrm{tr}_{\mathcal{E}}(V\rho V^*)$ defines a channel $\mathcal{N}\colon \mathbb{C}^{\dim \mathcal{V}} \to \mathcal{B} \otimes \mathcal{E}$.

This is the modern approach to tackling channel additivity questions. Dealing with subspaces and the Schmidt coefficients of their unit vectors is generally a much more tractable and computationally concrete problem than dealing with the Holevo capacity itself.

3.2 Proof via Dvoretzky's Theorem for $p > 1$

The original proof of the existence of superadditive quantum channels in [Has09] indeed took the approach of estimating Schmidt coefficients. However, it did not do so by looking at subspaces in isolation, but instead at coefficients of the outputs of random unital channels.

This proof was restated by [BH10] in terms of concentration of measure results, in the flavour of THEOREM 4 and LEMMA 4. The full extension to Dvoretzky's THEOREM 6 was ultimately given in [ASW10].

This is not the proof we will present, however, for it is simply too lengthy. Instead we will handle the $p > 1$ case, filling in the gaps of [Wu23]. The $p = 1$ case is actually morally the same, with even the same counterexample working. It is just that far more details must be checked to ensure the required estimates still hold.

The general approach is to get bounds on individual channel output entropies, with some sort of constants that depend on the channel size. We do something similar for a product of channels (we will use a channel tensored with its conjugate). Then, we simply push the dimensions far enough past the bounds to get superadditivity. In this section note that sometimes we will see $\approx$ used to describe values, and by this we mean equality up to some sort of hidden universal constant.

Lemma 6. *Let $\mathcal{N}$ be a quantum channel with associated subspace $\mathcal{V} \subseteq \mathbb{C}^k \otimes \mathbb{C}^d$. Then, for all $p > 1$ we have*

$$S_p^{min}(\mathcal{N}) = \frac{2p}{1-p} \log \sup_{\substack{M \in \mathrm{M}_d(\mathbb{C}) \\ \|M\|_2 = 1}} \|M\|_{2p}.$$

Proof. Take any pure state $|x\rangle \in \mathcal{V}$ and orthonormal bases $\{|i\rangle\}$ and $\{|j\rangle\}$ of $\mathbb{C}^k$ and $\mathbb{C}^d$, respectively. Write $|x\rangle = \sum_{i,j} \lambda_{ij} |ij\rangle$ and represent it via the matrix $M_x = (\lambda_{ij})$. Carelessly pushing around Dirac notation demonstrates

$$\begin{aligned}
\mathrm{tr}_{\mathbb{C}^d} |x\rangle\langle x| &= \mathrm{tr}_{\mathbb{C}^d} \left[\left(\sum_{i,j} |i\rangle \otimes |j\rangle \right) \left(\sum_{i,j} \bar{\lambda}_{ij} \langle i| \otimes \langle j| \right) \right] \\
&= \mathrm{tr}_{\mathbb{C}^d} \left[\sum_{i,j,k,\ell} \lambda_{ij} \bar{\lambda}_{k\ell} |i\rangle\langle k| \otimes |j\rangle\langle \ell| \right] \\
&= \sum_{i,j,k,\ell} \lambda_{ij} \bar{\lambda}_{k\ell} \, \mathrm{tr}(|j\rangle\langle \ell|) |i\rangle\langle k| \\
&= \sum_{i,j,k} \lambda_{ij} \bar{\lambda}_{kj} |i\rangle\langle k|
\end{aligned}$$

due to orthonormality of the bases, linearity of partial trace, and cyclicity of trace. It is then straightforward to verify from here that $\mathrm{tr}_{\mathbb{C}^d} |x\rangle\langle x| = M_x M_x^*$. Therefore,

$$\begin{aligned}
S_p(\mathrm{tr}_{\mathbb{C}^d} |x\rangle\langle x|) &= \frac{1}{1-p} \log \mathrm{tr}(M_x M_x^*)^p \\
&= \frac{1}{1-p} \log \mathrm{tr}(M_x^* M_x)^p \\
&= \frac{1}{1-p} \log \mathrm{tr}\, |M_x|^{2p}.
\end{aligned}$$

By definition we have $\|M_x\|_{2p}^{2p} = \mathrm{tr}\, |M_x|^{2p}$, and so

$$S_p(|y\rangle\langle y|) = \frac{2p}{1-p} \log \|M_x\|_{2p}.$$

Notice that the p-entropy is decreasing in the logarithm for $p > 1$, and so

$$S_p^{\min}(\mathcal{N}) = \frac{2p}{1-p} \log \sup_{x \in B(\mathcal{V})} \|M_x\|_{2p} = \frac{2p}{1-p} \log \sup_{\substack{M \in \mathcal{V} \\ \|M\|_2 = 1}} \|M\|_{2p}.$$

The last equality follows as x is a pure state, it satisfies $|x| = 1$ and so $\|M_x\|_2 = 1$ (going from the "commutative" Euclidean norm to the "non-commutative" Hilbert-Schmidt norm). We make an implicit identification between the space of tensors $\mathcal{V}$ and space of matrices $\mathcal{V}$ via $x \mapsto M_x$. □

Proposition 7. *Let $B_p^{mn} \subseteq \mathrm{M}_{mn}(\mathbb{C})$ be the unit ball with respect to the Schatten p-norm. Then, the inradius is*

$$\mathrm{inrad}(B_p^{mn})\begin{cases} m^{1/2-1/p} & 1 \leq p \leq 2 \\ 1 & 2 \leq p \leq \infty \end{cases}$$

and the mean width is $w(B_p^{mn}) \approx m^{1/p-1/2}$.

Proof. We will only deal with the inradius. Recall that Hölder's inequality (for vectors!) tells us that for $1 \leq p, q, r \leq \infty$ so $1/p + 1/q = 1/r$ that

$$\|xy\|_r \leq \|x\|_p \|y\|_q.$$

Taking the constant sequence $y = (1, \dots, 1)$ and working in $\mathbb{C}^m$ (since we may only have up to m singular values which determine the Schatten norm!) immediately yields

$$\|x\|_r \leq m^{1/r-1/p}\|x\|_p.$$

Take $r = 2$ and from PROPOSITION 4 we have

$$\mathrm{inrad}(B_p^{mn}) = \sup\left\{r > 0 : \frac{\|x\|_2}{\|x\|_p} \geq r \text{ for all } x \neq 0\right\}.$$

Whenever $p \leq 2$, we clearly see from Hölder's inequality that $\mathrm{inrad}(B_p^{mn}) = m^{1/2-1/p}$.

If $p \geq 2$ then we do not meet the conditions for the inequality (since with $r = 2$ we have $q < 0$). However, for all $1 \leq q \leq p \leq \infty$ we have

$$\|x\|_q^p = \left(\sum_i |x_i|^q\right)^{p/q} \geq \sum_i |x_i|^p = \|x\|_p^p.$$

This again extends to Schatten norms, and taking $q = 2$ then similarly allows us to deduce $\mathrm{inrad}(B_p^{mn}) = 1$. □

Proposition 8. *Fix $p > 1$. Then, there exist sufficiently large k, d and some subspace $\mathcal{V} \subseteq \mathbb{C}^k \otimes \mathbb{C}^d$ such that for the associated channel $\mathcal{N}$ we have*

$$\log k - c_1(p) \leq S_p^{min}(\mathcal{N}) \leq \log k - c_2(p).$$

Here, $c_1(p) > c_2(p) > 0$ are constants depending only on p, and $\dim \mathcal{V} \approx k^{1/p} d$.

Proof. Let $K = B_{2p}^{kd}$ be the Schatten $2p$-ball in $\mathrm{M}_{kd}(\mathbb{C})$. From PROPOSITION 7 we know it has Dvoretzky dimension

$$k_*(K) \approx \begin{cases} kd & 1 \leq 2p \leq 2 \\ k^{1/p}d & 2 \leq p \leq \infty \end{cases}$$

and mean width $w(K) \approx k^{1/2p-1/2}$. For the dimension, we of course are solely in the latter case since $p \geq 1$. Therefore, we can apply Dvoretzky's THEOREM 6 and attain with arbitrarily low failure $\varepsilon > 0$ some subspace $\mathcal{V} \subseteq \mathrm{M}_{kd}(\mathbb{C})$ of dimension $m \approx \varepsilon^2 k_*(K)$ such that

$$(1-\varepsilon)k^{1/2p-1/2}\|M\|_2 \leq \|M\|_{2p} \leq (1+\varepsilon)k^{1/2p-1/2}\|M\|_2$$

for all $M \in \mathcal{V}$. Note it is here that the dimensions k, d may need to be pushed very high so that $m > 1$.

We may naturally associate $\mathcal{V}$ with some subset of $\mathbb{C}^k \otimes \mathbb{C}^d$, and thus some channel $\mathcal{N}$. From LEMMA 6 we then have

$$S_p^{\min}(\mathcal{N}) = \frac{2p}{1-p} \log \sup_{\substack{M \in \mathcal{V} \\ \|M\|_2 = 1}} \|M\|_{2p}$$

and so our bounds tell us

$$\frac{2p}{1-p} \log\Big((1-\varepsilon)k^{1/2p-1/2}\Big) \geq S_p^{\min}(\mathcal{N}) \geq \frac{2p}{1-p} \log\Big((1+\varepsilon)k^{1/2p-1/2}\Big),$$

where the supremum is immaterial since $\|M\|_2 = 1$ always. Notice too the inequality flips as $p > 1$. We may rewrite this as

$$\log(k) + \frac{2p}{1-p} \log(1-\varepsilon) \geq S_p^{\min}(\mathcal{N}) \geq \log(k) + \frac{2p}{1-p} \log((1+\varepsilon)).$$

The $1 \pm \varepsilon$ terms define the constants $c_1(p)$ and $c_2(p)$. Though these technically depend on ε, we do not actually care about the probability of getting such a channel, so any fixed ε will work. Setting this constant is exactly how we get $m \approx k^{1/p}d$. □

Proposition 9. *Let $\mathcal{V} \subseteq \mathbb{C}^k \otimes \mathbb{C}^d$ have corresponding channel $\mathcal{N}$ with isometry $V : \mathbb{C}^m \to \mathbb{C}^k \otimes \mathbb{C}^d$, where $m = \dim \mathcal{V}$. Denote by $\bar{\mathcal{N}}$ the conjugate channel given by $\bar{V}$. Then,*

$$S_p^{min}(\mathcal{N} \otimes \bar{\mathcal{N}}) \leq \frac{p}{1-p} \log \frac{m}{kd}.$$

Proof. Let $\{e_i\}_{i=1}^m$ form an orthonormal basis for $\mathcal{V}$. Define $\bar{\mathcal{V}} = \bar{V}(\mathbb{C}^m)$ and let $\{\bar{e}_i\}_{i=1}^m$ be the conjugate orthonormal basis of $\bar{e}_i$. For an arbitrary $z \in \mathcal{V} \otimes \bar{\mathcal{V}}$, denote by $\mu_i(z)$ its i-th largest Schmidt coefficient. Consider now the state

$$x = \frac{1}{\sqrt{m}} \sum_{i=0}^{m-1} e_i \otimes \bar{e}_i,$$

whose maximal Schmidt coefficient is $\mu_1(x) = \sqrt{m/kd}$. It is also easy to see that $\mu_1(x) \leq \mu_1(y)$ for all $y \in \mathcal{V} \otimes \bar{\mathcal{V}}$ (and this is where the conjugacy plays in). Therefore,

$$S_p^{\min}(\mathcal{N} \otimes \mathcal{N}) \leq S_p((\mathcal{N} \otimes \mathcal{N})|x\rangle\langle x|) = \frac{p}{1-p} \sum_{i=1}^{m} \mu_i^2(x) \leq \frac{p}{1-p} \log \frac{m}{kd},$$

as claimed. □

Theorem 13. *Minimum output p-entropy is not additive for all $p > 1$.*

Proof. Start by fixing $p > 1$, and then applying PROPOSITION 8 to obtain a channel $\mathcal{N} : \mathbb{C}^m \to \mathbb{C}^k$ such that

$$S_p^{\min}(\mathcal{N}) \geq \log k - c(p)$$

for some constant $c(p) > 0$. Without loss of generality assume $\mathcal{N}$ extends into an ancilliary environment also of dimension k. Applying PROPOSITION 9 then tells us

$$S_p^{\min}(\mathcal{N} \otimes \bar{\mathcal{N}}) \leq \frac{p}{1-p} \log \frac{m}{k^2}.$$

However, since $m \approx k^{1+1/p}$, we simply have

$$S_p^{\min}(\mathcal{N} \otimes \bar{\mathcal{N}}) \leq \log k + \frac{p}{1-p} c'(p),$$

where $c'(p) > 0$ is some positive constant. So,

$$\begin{aligned} S_p^{\min}(\mathcal{N}) + S_p^{\min}(\bar{\mathcal{N}}) &\geq 2\log k - 2c(p) \\ &> \log k - 2c(p) + \log k - c'(p) \\ &\geq S_p^{\min}(\mathcal{N} \otimes \bar{\mathcal{N}}) + \log k - 2c(p). \end{aligned}$$

Since p is fixed, we may simply take k as large as necessary to ensure $\log k - 2c(p) > 0$, and then we have the desired violation of additivity. □

3.3 Constructive Counterexamples

To date, no constructive counterexamples exist violating additivity for the von Neumann entropy. However, there is an ongoing search. Constructive counterexamples for all $p > 2$ were found by [GHP10], using the antisymmetric subspace

$$W_{\text{as}} = \left\{ \frac{1}{\sqrt{2}} \left(e_i \otimes e_j - e_j \otimes e_i \right) : i < j \right\} \subseteq \mathbb{C}^k \otimes \mathbb{C}^k.$$

This required $d \to \infty$ as $p \to 2$, and since $\dim W_{\text{as}} = \binom{k}{2}$ the channel size explodes. This subspace was revisited recently by [SS23], who found that with m fixed, there exist n-dimensional subspaces $X_n \subseteq \mathbb{C}^k \otimes \mathbb{C}^k$ for all $n = 1, 2, \ldots, \lfloor d/2 \rfloor$ such that $W_n = W_{\text{as}} \oplus X_n$ violates additivity. Though this tames the size of the channel, it comes at the cost of losing control of p, and so cannot approach the von Neumann entropy at all.

In the same work, however, [SS23] also found the first constructive counterexample for $1 < p < 2$. Making use of the completely entangled subspace

$$W_{\text{ce}} = \operatorname{span}\{u_\lambda \otimes u_\lambda : \lambda \in G\}^\perp \subseteq \mathbb{C}^k \otimes \mathbb{C}^k$$

introduced by [Par04], they found numerical evidence that subspaces $S \subseteq W_{\text{ce}}$ can violate additivity. With $k = 5$, when $\dim W_{\text{ce}} = (5-1)^2 = 16$, there exists some 14-dimensional S which violates additivity for $p \approx 1.844$.

These types of "pathological" subspaces, which heavily constrain the Schmidt coefficients of their vectors, are a promising area to find counterexamples. The difficulty in moving from numerical to analytic expressions is the optimization problem inherent in minimum output entropy. Until a precise expression for the Schmidt coefficients is found, the problem can become numerically intractable even for low dimensions – let alone solvable by hand.

References

[Sha48] Claude Shannon. “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27 (3 1948), pp. 379–423.

[Dvo64] Aryeh Dvoretzky. “Some Results on Convex Bodies and Banach Spaces”. In: *Matematika* 8 (1 1964), pp. 73–102.

[Mil71] Vitali Milman. “A New Proof of A. Dvoretzky’s Theorem on Intersections of Convex Bodies”. In: *Funktsionalnyi Analiz i Ego Prilozheniya* 5 (4 1971), pp. 28–37.

[Hol73] Alexander Holevo. “Bounds for the Quantity of Information Transmitted by a Quantum Communication Channel”. In: *Problems of Information Transmission* 9 (1973), pp. 177–183.

[SW97] Benjamin Schumacher and Michael Westmoreland. “Sending Classical Information via Noisy Quantum Channels”. In: *Physical Review A* 56 (1997), pp. 131–138.

[WH02] Reinhard Werner and Alexander Holevo. “Counterexample to an Additivity Conjecture for Output Purity of Quantum Channels”. In: *Journal of Mathematical Physics* 43 (9 2002), pp. 4353–4357.

[Par04] K. R. Parthasarathy. “On the Maximal Dimension of a Completely Entangled Subspace for Finite Level Quantum Systems”. In: *Proceedings Mathematical Sciences* 114 (2004), pp. 365–374.

[Sho04] Peter Shor. “Equivalence of Additivity Questions in Quantum Information Theory”. In: *Communications in Mathematical Physics* 246 (2004), pp. 453–472.

[Cub+08] Toby Cubitt et al. “Counterexamples to Additivity of Minimum Output p-Rényi Entropy for p Close to 0”. In: *Communications in Mathematical Physics* 284 (2008), pp. 281–290.

[HW08] Patrick Hayden and Andreas Winter. “Counterexamples to the Maximal p-Norm Multiplicativity Conjecture for all $p > 1$”. In: *Communications in Mathematical Physics* 284 (2008), pp. 263–280.

[Has09] Matthew Hastings. “Superadditivity of Communication Capacity Using Entangled Inputs”. In: *Nature Physics* 5 (2009), pp. 255–257.

[ASW10] Guillaume Aubrun, Stanisław Szarek, and Elisabeth Werner. “Hasting’s Additivity Counterexample via Dvoretzky’s Theorem”. In: *Communications in Mathematical Physics* 305 (2010), pp. 85–97.

[BH10] Fernando Brandão and Michał Horodecki. “On Hastings’ Counterexamples to the Minimum Output Entropy Additivity Conjecture”. In: *Open Systems & Information Dynamics* 17 (1 2010), pp. 31–52.

[GHP10] Andrzej Grudka, Michał Horodecki, and Łukasz Pankowski. “Constructive Counterexamples to the Additivity of the Minimum Output Rényi Entropy of Quantum Channels for all $p > 2$”. In: *Journal of Physics A* 43 (2010).

[HS13] Alexander Holevo and Maksim Shirokov. “On Classical Capacities of Infinite-Dimensional Quantum Channels”. In: *Problems of Information Transmission* 49 (2013), pp. 15–31.

[Pre16] John Preskill. “Quantum Shannon Theory”. In: *arXiv:1604.07450* (2016).

[AS17] Guillaume Aubrun and Stanisław Szarek. *Alice and Bob Meet Banach. The Interface of Asymptotic Geometric Analysis and Quantum Information Theory*. American Mathematical Society, 2017.

[SS23] K. Szczygielski and M. Studziński. “New Constructive Counterexamples to Additivity of Minimum Output Rényi Entropy of Quantum Channels”. In: *arXiv:2301.07428* (2023).

[Wu23] Peixue Wu. *Generic Nonadditivity of Minimum Output Entropy of Quantum Channels*. Personal Communication. 2023.