Mathematical Aspects of Quantum Information Theory:

Lecture 6

Dario Trevisan

Università di Pisa
dario.trevisan@unipi.it

Plan

Plan

1. A quantum coding theorem
   - The classical case
   - The quantum case

- Shannon's limit which quantifies the maximum rate at which information can be transferred via a noisy communication channel.
- Alice sends a message X (with values in a set $\mathcal{X}$) through a noisy communication channel. Bob receives a distorted Y (with values $\mathcal{Y}$).
- The channel is modelled as Markov kernel N from $\mathcal{X}$ to $\mathcal{Y}$, so that, if p denotes the law of X,

$$\mathbb{P}(X = x, Y = y) = p(x)N(x, y).$$

- The channel is memoryless, i.e., n applications give $N^{\otimes n}$ from $\mathcal{X}^n$ to $\mathcal{Y}^n$,

$$N^{\otimes n}((x_i)_{i=1}^n, (y_i)_{i=1}^n) = \prod_{i=1}^n N(x_i, y_i).$$

- Alice and Bob agree to use iterated applications of the channel and transmit the message via a coding procedure.

Classical codes

A code (W, V) consists of

1. a codebook
$$W : \{1, \ldots, m\} \to \mathcal{X}^n$$
with m codewords (size of the code) of a fixed length n, to be transmitted by Alice via the composite channel $N^{\otimes n}$,
2. a decision rule,
$$V : \mathcal{Y}^n \to \{0, 1, \ldots, m\}$$
which represents Bob's estimate:
   - if $V(y) = i$, $i \neq 0$, Bob decodes y as $W(i)$,
   - if $V(y) = 0$, Bob makes no decision.

The transmission rate, i.e. the number of bits of information per application of the channel is therefore $\log m/n$.

Error probabilities

- For each $i \in \{1, \ldots, m\}$, the probability that Bob decodes correctly the word, given that Alice sent word i, is

$$\mathbb{P}(V(y) = i | W(i)) = \sum_{y \in \{V = i\}} N^{\otimes n}(W(i), y) = \sum_{y \in \{V = i\}} \prod_{j=1}^{m} N(y_j | W(i)_j).$$

- Two indicators:
  1. the maximal error probability

$$p_e(W, V) = \max_{i=1,\ldots,m} (1 - \mathbb{P}(V(y) = i | W(i)),$$

  2. the mean error probability

$$\bar{p}_e(W, V) = \frac{1}{m} \sum_{i=1}^{m} (1 - \mathbb{P}(V(y) = i | W(i)).$$

- By Markov inequality, from any code (W, V) with size $2m$ one can extract a sub-code $(\tilde{W}, \tilde{V})$ of size at least m such that

$$p_e(W, V) \leq 2\bar{p}_e(\tilde{W}, \tilde{V}).$$

Operational channel capacity

- For given length n and size m, let

$$p_e(n,m) = \min_{(W,V)} p_e(W,V), \quad \bar{p}_e(n,m) = \min_{(W,V)} \bar{p}_e(W,V),$$

- $r > 0$ is an achievable transmission rate for the channel N if

$$\lim_{n\to\infty} p_e(n, 2^{nr}) = 0.$$

- The (operational) channel capacity $\mathcal{C}(N)$ is the largest achievable transmission rates r:
  1. (direct statement) for every $r < \mathcal{C}(N)$,

$$\lim_{n\to\infty} \bar{p}_e(n, 2^{nr}) = 0,$$

  2. (weak converse) for every $r > \mathcal{C}(N)$,

$$\limsup_{n\to\infty} \bar{p}_e(n, 2^{nr}) > 0,$$

- The channel capacity is additive

$$\mathcal{C}(N^{\otimes k}) = k\mathcal{C}(N), \quad \text{for every } k \geq 1.$$

- Any code with respect to $N^{\otimes k}$ of length n is also a code with respect to N, with length kn.
- Viceversa, any code with respect to N can be turned into a code with respect to $N^{\otimes k}$ by k repeated applications.

Information channel capacity

- Recall that we introduced the mutual information

$$I(X;Y) = S(X) - S(X|Y) = S(Y) - S(Y|X)$$

- Using that $\mathbb{P}(X = x, Y = y) = p(x)N(x,y)$, we have

$$I(X;Y) = S\left(\sum_{x\in\mathcal{X}} p(x)N(x,\cdot)\right) - \sum_{x\in\mathcal{X}} p(x)S(N(x,\cdot)),$$

- $p \mapsto I(X;Y)$ is concave.
- The information channel capacity is

$$\mathcal{C}_I(N) = \max_p I(X;Y) = \max_p \left\{ S\left(\sum_{x\in\mathcal{X}} p(x)N(x,\cdot)\right) - \sum_{x\in\mathcal{X}} p(x)S(N(x,\cdot))\right\}.$$

Theorem (Shannon's limit)

It holds

$$\mathcal{C}(N) = \mathcal{C}_I(N).$$

Structure of the proof:

- the weak converse statement (inequality $\leq$) via Fano's inequality
- the direct statement (inequality $\geq$) via a random coding argument.

Weak converse, $\mathcal{C}(N) \leq \mathcal{C}_I(N)$

- Let (W, V) be any code of length n and size $m = 2^{nr}$.
- We turn W a random variable $X^n = W$ with uniform distribution on the m codewords.
- Applying $N^{\otimes n}$, V becomes a random variable with values in $\{0, 1, \ldots, m\}$, yielding an estimator W' of W, with

$$\mathbb{P}(W' \neq W) \leq \frac{1}{m} \sum_{i=1}^{m} (1 - \mathbb{P}(V = i | W(i)) = \bar{p}_e(W, V).$$

- Fano's inequality yields

$$\begin{aligned} S(W|V) \leq S(W|W') &\leq h_2(\bar{p}_e(W, V)) + \bar{p}_e(W, V) \log(m-1) \\ &\leq 1 + \bar{p}_e(W, V) \log m. \end{aligned}$$

- Since $S(W) = \log m$,

$$I(W; V) = S(W) - S(W|V) \geq \log m - \bar{p}_e(W, V)) \log m - 1.$$

$$I(W;V) \geq \log m - \bar{p}_e(W,V)) \log m - 1.$$

- Since $I(W;V) \leq \mathcal{C}_I(N^{\otimes n})$, we deduce

$$\bar{p}_e(n, 2^{nr}) \geq 1 - \frac{\mathcal{C}_I(N^{\otimes n}) + 1}{nr}.$$

- As $n \to \infty$, we obtain

$$\limsup_{n\to\infty} \bar{p}_e(n, 2^{nr}) \geq 1 - \frac{1}{r} \liminf_{n\to\infty} \frac{\mathcal{C}_I(N^{\otimes n})}{n}.$$

- Any

$$r > \liminf_{n\to\infty} \frac{\mathcal{C}_I(N^{\otimes n})}{n}$$

  is not an achievable transmission rate, hence

$$\mathcal{C}(N) \leq \liminf_{n\to\infty} \frac{\mathcal{C}_I(N^{\otimes n})}{n}.$$

- To conclude, we argue that

$$\liminf_{n\to\infty} \frac{\mathcal{C}_I(N^{\otimes n})}{n} = \mathcal{C}_I(N).$$

- $\mathcal{C}_I$ is super-additive, i.e.,

$$\mathcal{C}_I(N^{\otimes(k+h)}) \geq \mathcal{C}_I(N^{\otimes k}) + \mathcal{C}_I(N^{\otimes h}).$$

- Tensorization property of relative entropy

$$S(\rho \otimes \rho' || \sigma \otimes \sigma') = S(\rho||\sigma) + S(\rho'||\sigma').$$

Hence

$$I((X^k, X^h); (Y^k, Y^h))_{p^k \otimes p^h} = I(X^k; Y^k)_{p_k} + I(X^h; Y^h)_{p_h}.$$

- For optimal p^k, p^h, we have stationarity conditions $\Rightarrow$ optimality (because of concavity).

Direct statement, $\mathcal{C}(N) \geq \mathcal{C}_I(N)$

- The strategy is to sample a code randomly among all the possible codes.
- Law of large numbers:

$$S(p) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -\mathbb{E}\left[\log p(X)\right]$$

is close to the empirical average

$$-\frac{1}{n}\sum_{i=1}^{n} \log p(X_i) = -\frac{1}{n} \log \left(\prod_{i=1}^{n} p(X_i) \right).$$

- A typical random word $W = (X_i)_{i=1}^{n}$ will have probability of occurrence

$$\mathbb{P}(W = w) = \prod_{i=1}^{n} p(X_i) \approx 2^{-nS(p)},$$

- for large n, W behaves as a uniformly distributed variable over $2^{nS(p)}$ values (asymptotic equipartition property)

- We build codes (W, V) attaining any rate $r < \mathcal{C}_I(N)$.
- We sample $m = 2^{nr}$ independent words $(W(i))_{i=1}^m$, each of length n, according to a single letter distribution p.
- Bob's decision rule V. After receiving y,
  1. first, he checks if y is a typical word for Y, otherwise he sets $V(y) = 0$.
  2. for every $i \in \{1, \ldots, m\}$, he checks (in sequence) if y is conditionally typical for the word $W(i)$ in the codebook, i.e.,

     $$\mathbb{P}(y|W(i)) \approx 2^{-nS(Y|X)}.$$

     He stops at the first affirmative case and sets $V(y) = i$.
  3. If no word is conditionally typical, he sets $V(y) = 0$.

- The output of the channel is effectively is uniformly distributed over $2^{nS(Y)}$ values.
- On average, for each codeword in W, we have $2^{nS(Y|X)}$ conditionally typical outputs.
- By independence, the conditionally typical outputs are well-separated.
- We are able to build a code of size m,

$$m \approx \frac{2^{nS(Y)}}{2^{nS(Y|X)}} = 2^{n(S(Y)-S(Y|X))} = 2^{nI(X;Y)_p},$$

  with asymptotically small error.
- Choosing p in order to maximize $I(X;Y)_p$ leads to $\mathcal{C}(N) \geq \mathcal{C}_I(N)$.

Plan

1. A quantum coding theorem
   - The classical case
   - The quantum case

- We focus on the case of a classical to quantum channel,

$$\Phi : \mathcal{X} \ni x \mapsto \Phi_x \in \mathcal{S}(H).$$

- We can extend Φ to a quantum channel from $\mathbb{C}^{\mathcal{X}}$ into H, with Kraus representation

$$\Phi(\rho) = \sum_{x \in \mathcal{X}} \sqrt{\Phi_x} \langle x | \rho x \rangle \sqrt{\Phi_x}.$$

- Bob measures the state to extract information on Alice's message.

- We consider general non-sharp measurements $M = (M_y)_{y \in \mathcal{Y}}$ given by POVM's, i.e., $M_y \in \mathcal{O}_{\geq 0}(H)$ and such that

$$\sum_{y \in \mathcal{Y}} M_y = \mathbb{1}_H.$$

- Such measurements strictly includes the sharp case $V = (\mathbb{1}_{V_y})_{y\in\mathcal{Y}}$, but greatly simplifies the mathematical derivation.
- The analogy with classical case is to allow for probabilistic decision rules (i.e., given by Markov kernels)
- We associate to M a quantum to classical channel Φ_M, from H to $\mathbb{C}^{\mathcal{Y}}$,

$$\mathcal{S}(H) \ni \rho \mapsto \sum_{y\in\mathcal{Y}} \mathrm{tr}[M_y \rho] \left| y \right\rangle \left\langle y \right| ,$$

- $\mathrm{tr}[M_y\rho] = \mathbb{P}_\rho(M = y)$ is the probability that, measuring M, we observe y.

- We assume that repeated applications of the channel are memoryless: a word $w = (x_i)_{i=1}^n \in \mathcal{X}^n$ sent by Alice arrives to Bob as the product state

$$\Phi_w = \Phi_{x_1} \otimes \Phi_{x_2} \otimes \ldots \otimes \Phi_{x_n} \in \mathcal{S}(H^{\otimes n}).$$

- n applications of Φ correspond to a single application of the channel $\Phi^{\otimes n}$.

- A code (W, M), consists of
  1. a (classical) codebook $W : \{1, \ldots, m\} \to \mathcal{X}^n$ with size m,
  2. a quantum decision rule $M = (M_i)_{i=0,\ldots m} \subseteq \mathcal{O}_{\geq 0}(H^{\otimes n})$, such that

$$\sum_{i=0}^{m} M_i = \mathbb{1}_{H^{\otimes n}}.$$

- A code of size m, and length n, has transmission rate $\log m/n$.

- We have

$$\mathbb{P}(\text{“measures } M \text{ and observes } j\text{”} \,|\, \text{“Alice sent the word } w\text{”}) = \mathrm{tr}[M_j \Phi_w].$$

- We define
  1. the maximal error probability

$$p_e(W, M) = \max_{j=1,\ldots,m} \left(1 - \mathrm{tr}[M_j \Phi_{W(j)}]\right),$$

  2. the mean error probability

$$\bar{p}_e(W, M) = \frac{1}{m} \sum_{i=1}^{m} \left(1 - \mathrm{tr}[M_j \Phi_{W(j)}]\right).$$

- We set

$$p_e(n, m) = \min_{(W,M)} p_e(W, M), \quad \bar{p}_e(n, m) = \min_{(W,M)} \bar{p}_e(W, M).$$

- $r > 0$ is an achievable transmission rate for the channel Φ if

$$\lim_{n \to \infty} p_e(n, 2^{nr}) = 0.$$

The (operational classical) channel capacity $\mathcal{C}(\Phi)$ is the largest achievable transmission rate:

- (direct statement) for every $r < \mathcal{C}(\Phi)$,

$$\lim_{n\to\infty} \bar{p}_e(n, 2^{nr}) = 0,$$

- (weak converse) for every $r > \mathcal{C}(\Phi)$,

$$\limsup_{n\to\infty} \bar{p}_e(n, 2^{nr}) > 0,$$

As in the classical case, the channel capacity is an additive quantity, i.e.,

$$\mathcal{C}(\Phi^{\otimes k}) = k\mathcal{C}(\Phi), \quad \text{for every } k \geq 1.$$

- For a $\Phi = (\Phi_x)_{x\in\mathcal{X}}$ and a probability distribution p on $\mathcal{X}$, define

$$\chi(\Phi)_p = S\left(\sum_{x\in\mathcal{X}} p(x)\Phi_x\right) - \sum_{x\in\mathcal{X}} p(x)S(\Phi_x),$$

where S denotes von Neumann entropy

- By concavity of von Neumann entropy, $p \mapsto \chi(\Phi)_p$ is concave.
- Define the χ-capacity of Φ as

$$\mathcal{C}_\chi(\Phi) = \max_p \chi(\Phi)_p = \max_p \left\{ S\left(\sum_{x\in\mathcal{X}} p(x)\Phi_x\right) - \sum_{x\in\mathcal{X}} p(x)S(\Phi_x)\right\}.$$

- As in the classical case (same argument) it is additive:

$$\mathcal{C}_\chi(\Phi^{\otimes k}) = k\mathcal{C}_\chi(\Phi).$$

Holevo's bound

Theorem (Schumacher-Westmoreland, Holevo)

It holds

$$\mathcal{C}(\Phi) = \mathcal{C}_\chi(\Phi).$$

Notice that

$$\mathcal{C}_\chi(\Phi) \leq \max_p S\left(\sum_{x \in \mathcal{X}} p(x)\Phi_x\right) \leq \log \dim H,$$

(it was similarly true, but less surprising, in the classical case).

Weak converse, $\mathcal{C}(\Phi) \leq \mathcal{C}_\chi(\Phi)$

- Idea: reduce to the classical case, with Markov kernel $N(x, y) = \mathrm{tr}[M_y \Phi_x]$.
- Notice that (exercise):

$$\chi(\Phi)_p = I(\mathbb{C}^{\mathcal{X}}; H)_\rho = S(\rho || \rho_{\mathbb{C}^{\mathcal{X}}} \otimes \rho_H)$$

where

$$\rho = \sum_{x \in \mathcal{X}} p(x) \left|x\right\rangle \left\langle x\right| \otimes \Phi_x.$$

- Given any measurement M, consider the channel Φ_M from H to $\mathbb{C}^{\mathcal{Y}}$. By the data processing inequality,

$$\mathcal{C}_\chi(\Phi) \geq I(\mathbb{C}^{\mathcal{X}}; H)_\rho \geq I(\mathbb{C}^{\mathcal{X}}; \mathbb{C}^{\mathcal{Y}})_{\mathbb{1}_{\mathcal{L}(\mathbb{C}^{\mathcal{X}})} \otimes \Phi_M(\rho)} = I(X; Y)_p,$$

where the classical variable (X, Y) have density

$$\mathbb{P}(X = x, Y = y) = p(x) \mathrm{tr}[M_y \Phi_x].$$

- Repeating the argument with $\Phi^{\otimes n}$ yields

$$\mathcal{C}_\chi(\Phi^{\otimes n}) \geq \chi(\Phi^{\otimes n})_\rho \geq I(X^n; Y^n)_p.$$

- We follow the argument as in the classical case: given a code (W, M) of size $m = 2^{nr}$ assign uniform probability to W and use Fano's inequality

$$\bar{p}_e(W, M)) \geq 1 - \frac{I(X^n; Y^n)_p + 1}{\log m} \geq 1 - \frac{\mathcal{C}_\chi(\Phi^{\otimes n}) + 1}{\log m}.$$

- Any rate r such that

$$r > \liminf_{n \to \infty} \frac{\mathcal{C}_\chi(\Phi^{\otimes n})}{n}$$

is not admissible, hence

$$\mathcal{C}(\Phi) \leq \liminf_{n \to \infty} \frac{\mathcal{C}_\chi(\Phi^{\otimes n})}{n} = \mathcal{C}_\chi(\Phi).$$

- We have the inequality

$$\mathcal{C}_{\chi}(\Phi) \geq \sup_{M,\rho} \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x)\mathrm{tr}[M_y\Phi_x] \log\left(\frac{\mathrm{tr}[M_y\Phi_x]}{p(x)\sum_{x'\in\mathcal{X}} p(x')\mathrm{tr}[M_y\Phi_{x'}]}\right),$$

- It is known that the inequality can be strict.
- The right hand side coincides with the operational channel capacity of Φ when Bob is restricted to measurements of product type

$$M_j = M_{j_1} \otimes M_{j_2} \otimes \ldots \otimes M_{j_n}.$$

- This is advantage is a manifestation of entanglement: even if Alice's messages are presented to Bob as product states, using general non-product observables can be an advantage for Bob.

Direct statement, $\mathcal{C}(\Phi) \geq \mathcal{C}_\chi(\Phi)$.

- Given a state $\rho \in \mathcal{S}(H)$, $n \geq 1$ and $\delta > 0$, its δ-typical subspace on $H^{\otimes n}$ consists of the span of the eigenvectors of $\rho^{\otimes n}$ with eigenvalues λ such that
$$2^{-nS(p)-n\delta} \leq \lambda \leq 2^{-nS(p)+n\delta}.$$
- The asymptotic equipartition property states that, for every $\delta, \varepsilon > 0$,
  1. For every n, the dimension of the δ-typical subspace of ρ is
  $$\mathrm{tr}[P^{\delta,n}] \leq 2^{n(S(p)+\delta)},$$
  2. For $n \gg 1$, the contribution of vectors not δ-typical is
  $$\mathrm{tr}[(1 - P^{\delta,n})\rho^{\otimes n}] \leq \varepsilon.$$
  3. For $n \gg 1$, the dimension of the δ-typical subspace of ρ is
  $$\mathrm{tr}[P^{\delta,n}] \geq (1 - \varepsilon)2^{n(S(p)-\delta)},$$

- Given a probability distribution p over $\mathcal{X}$, write

$$S(H|\mathbb{C}^{\mathcal{X}})_\rho = \sum_{x \in \mathcal{X}} p(x) S(\Phi_x),$$

for the the quantum conditional entropy of

$$\rho = \sum_{x \in \mathcal{X}} p(x) \, |x\rangle \, \langle x| \otimes \Phi_x .$$

- For $n \geq 1$, $\delta > 0$ and $w = (x_i)_{i=1}^n \in \mathcal{X}^n$, define the conditionally typical subspace of Φ given w (and p) as the linear span of the eigenvectors of $\Phi_w = \otimes_{i=1}^n \Phi_{x_i}$ whose eigenvalues λ satisfy

$$2^{-nS(H|\mathbb{C}^{\mathcal{X}})_\rho - n\delta} \leq \lambda \leq 2^{-nS(H|\mathbb{C}^{\mathcal{X}})_\rho + n\delta},$$

- Properties:
  1. For every n, and w, we have

  $$P_w^{\delta,n} \leq 2^{n(S(H|\mathbb{C}^{\mathcal{X}})_\rho + \delta)} \Phi_w,$$

  2. For n sufficiently large, and $\varepsilon > 0$,

  $$\mathbb{E}\left[\mathrm{tr}[(1 - P_W^{\delta,n})\Phi_W] \right] \leq \varepsilon,$$

  where $W = (X_i)_{i=1}^n$ are i.i.d. with common distribution p.

- The codebook W is obtained via i.i.d. random sampling of m codewords $(w_j)_{j=1}^m$ of length n.

- Fix $\delta > 0$ and n. Write $P = P^{\delta,n}$ for the typical projector associated to

$$\rho_H = \sum_{x \in \mathcal{X}} p_x \Phi_x,$$

and $P_w = P_w^{\delta,n}$ for the conditional typical projectors.

- Intuition: define $M_j = P_{w_j} P$, but not self-adjoint.

- Definition:

$$M_j = A^{-1/2} P P_{w_j} P A^{-1/2} = (P_{w_j} P A^{-1/2})^* P_{w_j} P A^{-1/2},$$

where

$$A = \sum_{j=1}^m P P_{w_j} P.$$

- Working with such definition, one obtains the upper bound

$$\bar{p}_e(W, M) \leq \frac{1}{m} \sum_{j=1}^{m} 4\mathrm{tr}[\Phi_{w_j}(1 - P)] + 4\mathrm{tr}[\Phi_{w_j}(1 - P_{w_j})] + \sum_{i \neq j} \mathrm{tr}[P\Phi_{w_j} P P_{w_i}].$$

- Taking expectation (w.r.t the sampling generating the codebook),

$$\mathbb{E}\left[\Phi_X\right] = \sum_{x \in \mathcal{X}} p(x)\Phi_x = \rho_H.$$

- Using independence,

$$\begin{aligned}&\mathbb{E}\left[\bar{p}_e(W, M)\right] \\ &\quad \leq 4\mathrm{tr}[\rho_H^{\otimes n}(1 - P)] + 4\mathbb{E}\left[\mathrm{tr}[\Phi_w(1 - P_w)]\right] + (m-1)\mathrm{tr}[P\rho_H^{\otimes n} P\mathbb{E}\left[P_w\right]].\end{aligned}$$

$$\mathbb{E}\left[\bar{p}_e(W,M)\right] \leq 4\mathrm{tr}[\rho_H^{\otimes n}(1-P)] + 4\mathbb{E}\left[\mathrm{tr}[\Phi_w(1-P_w)]\right] + (m-1)\mathrm{tr}[P\rho_H^{\otimes n}P\mathbb{E}\left[P_w\right]].$$

- For $n \gg 1$,

$$\mathrm{tr}[\rho_H^{\otimes n}(1-P)] + \mathbb{E}\left[\mathrm{tr}[\Phi_w(1-P_w)]\right] \leq 2\varepsilon,$$

- By definition of typical subspace,

$$P\rho_H^{\otimes n}P \leq 2^{-nS(H)_{\rho_H}+n\delta}\mathbb{1}_{H^{\otimes n}},$$

- Hence,

$$\begin{aligned}\mathrm{tr}[P\rho_H^{\otimes n}P\mathbb{E}\left[P_w\right]] &\leq 2^{-nS(H)_{\rho_H}+n\delta}\mathbb{E}\left[\mathrm{tr}[P_w]\right] \\ &\leq 2^{-nS(H)_{\rho_H}+nS(H|\mathbb{C}^{\mathcal{X}})_\rho+2\delta n} \\ &= 2^{-nI(H;\mathbb{C}^{\mathcal{X}})_\rho+2\delta n}.\end{aligned}$$

- Choosing p such that $I(H;\mathbb{C}^{\mathcal{X}})_\rho = \mathcal{C}_\chi(\Phi)$ we obtain that any $r < \mathcal{C}_\chi(\Phi)$ is an achievable transmission rate.