Mathematical Aspects of Quantum Information Theory:

Lecture 5

Dario Trevisan

Università di Pisa
dario.trevisan@unipi.it

Plan

1. Distances (conclusion)
   - Quantum optimal transport
2. Entropy
   - Classical entropy
   - Quantum entropy

Plan

1. Distances (conclusion)
   - Quantum optimal transport
2. Entropy
   - Classical entropy
   - Quantum entropy

Errata

- Recall the quantum fidelity between two states ρ, $\sigma \in \mathcal{S}(H)$:

$$F(\rho, \sigma) = \mathrm{tr}[\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}]^2.$$

- Motivated by the analogy with the Bhattacharyya coefficient, the analogue of the Hellinger distance is the Bures metric

$$D_B(\rho, \sigma)^2 = 2\left(1 - \sqrt{F(\rho, \sigma)}\right).$$

- The Bures metric is an actual distance (check the updated notes, reference e.g. in Holevo's book).

OT via Lipschitz operators

- In the classical case, we can use Kantorovich duality to define W^d:

$$W^d(p,q) = \sup_{f \text{ is 1-Lip}} \left\{ \sum_{x \in \mathcal{X}} f(x)\,(p(x) - q(x)) \right\}.$$

- A similar strategy in the quantum setting dates back to Connes: define first what are Lipschitz observables and obtain the cost via duality.
- We proposed to consider the case of product systems

$$H = \bigotimes_{i \in I} H_i,$$

providing a quantum analogue of OT with respect to the Hamming distance

- Recall that on sets $\Pi_{i \in I} \mathcal{X}_i$,

$$d_{\mathrm{Ham}}((x_i)_{i \in I}, (y_i)_{i \in I}) = \sum_{i \in I} 1_{\{x_i \neq y_i\}}.$$

- $f : \Pi_{i \in I} \mathcal{X}_i \to \mathbb{R}$ is (Hamming) 1-Lipschitz if and only if, for every $i \in I$,

$$|f(x) - f(y)| \leq 1$$

whenever x, y differ only at the coordinate i (write $x \sim_i y$).

- Equivalently, define the oscillation at $i \in I$ as

$$\partial_i f = \sup_{x \sim_i y} |f(x) - f(y)| = 2 \inf_{g_i} \sup_{x} |f(x) - g_i(x)|$$

where g_i does not depend upon the coordinate i. Then,

$$\|f\|_{\mathrm{Lip}} = \max_{i \in I} \partial_i f.$$

- On a product system $H = \bigotimes_{i \in I} H_i$, for every $i \in I$ and observable $A \in \mathcal{O}(H)$, define

$$\partial_i A = \inf \left\{ 2 \left\| A - G_i \otimes \mathbb{1}_{H_i} \right\|_\infty \ : \ G_i \in \mathcal{O}(\bigotimes_{j \neq i} H_j) \right\},$$

- The quantum Lipschitz constant of $A \in \mathcal{O}(H)$ is

$$\|A\|_L := \max_{i \in I} \partial_i A.$$

- The quantum Wasserstein distance of order 1 between $\rho, \sigma \in \mathcal{S}(H)$ is

$$\begin{aligned} \|\rho - \sigma\|_{W_1} &= \sup \left\{ \mathrm{tr}[A(\rho - \sigma)] \ : \ \|A\|_L \le 1 \right\} \\ &= \sup \left\{ (A)_\rho - (A)_\sigma \ : \ \|A\|_L \le 1 \right\} \end{aligned}$$

- Back to the classical case, forget about the product structure (i.e., consider the set $\mathcal{X}$ a single factor): then the Hamming distance is the trivial distance and
$$W^{d_{trivial}}(p,q) = \|p-q\|_{TV}.$$
- Since
$$1_{\{x\neq y\}} \leq \sum_{i\in I} 1_{\{x_i\neq y_i\}} \leq |I| 1_{\{x\neq y\}},$$
this leads to a comparison between OT distances.
- Also in the quantum case, we can compare
$$D_{\mathrm{tr}}(\rho,\sigma) \leq \|\rho-\sigma\|_{W_1} \leq |I| D_{\mathrm{tr}}(\rho,\sigma).$$

- For product states $\rho = \otimes_{i \in I} \rho_i$, $\sigma = \otimes_{i \in I} \sigma_i$, then

$$\|\rho - \sigma\|_{W_1} = \sum_{i \in I} D_{\mathrm{tr}}(\rho_i, \sigma_i).$$

- Exercise: Compute the Wasserstein distance of order 1 between any two Bell states on the composite system $H = \mathbb{C}^2 \otimes \mathbb{C}^2$, e.g.

$$\rho = \frac{1}{2} (|00\rangle + |11\rangle) (\langle 00| + \langle 11|),$$

$$\sigma = \frac{1}{2} (|01\rangle + |10\rangle) (\langle 01| + \langle 10|).$$

Plan

1. Distances (conclusion)
   - Quantum optimal transport
2. Entropy
   - Classical entropy
   - Quantum entropy

- Given a probability p over a set Ω, its Shannon entropy is

$$S(p) = -\sum_{\omega \in \Omega} p(\omega) \log p(\omega),$$

- We assume $0 \log 0 = 0$ and that $\log = \log_2$ (S is measured in bits)
- $S(p) \geq 0$, and $p \mapsto S(p)$ is concave.
- Examples:
  1. If p is uniform over n values,

$$S((1/n)_{i=1}^{n}) = -n \cdot \frac{1}{n} \log(1/n) = \log n.$$

  2. For a probability distribution over two values (a Bernoulli law),

$$S((\alpha, 1-\alpha)) = -\alpha \log \alpha - (1-\alpha) \log(1-\alpha) = h_2(\alpha).$$

- $\alpha \mapsto h_2(\alpha)$ is called binary entropy function.

Entropy as information content

- The entropy of a random variable $X : \Omega \mapsto \mathcal{X}$ is

$$S(X) = S((\mathbb{P}(X = x))_{x\in\mathcal{X}}) = -\sum_{x\in\mathcal{X}} \mathbb{P}(X = x)\log(\mathbb{P}(X = x)).$$

- To avoid(!) ambiguities, $S(X) = S_p(X)$ (p is the low X or the probability on Ω)
- $S(X)$ measures the information content of a random variable X:

- It holds $0 \leq S(X) \leq \log|\mathcal{X}|$.

Conditional entropy

- If Bob observes another random variable Y (possibly correlated with X), how should he update the entropy of X?
- After Bob observes $Y = y$, he updates the law of X, hence

$$S(X)_{\mathbb{P}|Y=y} = -\sum_{x\in\mathcal{X}} \mathbb{P}(X = x|Y = y) \log \mathbb{P}(X = x|Y = y).$$

- From the engineer's viewpoint, we are more interested in the average over the values Bob may observe. Hence the conditional entropy of X given Y is

$$S(X|Y) = \sum_{y\in\mathcal{Y}} S(X)_{\mathbb{P}|Y=y}\mathbb{P}(Y = y).$$

- Notice that $S(X|Y) \geq 0$
- Moreover, $S(X|Y) = S((X, Y)) - S(Y)$, or equivalently the chain rule:

$$S(X, Y) = S(Y) + S(X|Y).$$

Lemma (Fano's inequality)

Let X, X' be random variables and set $p = \mathbb{P}(X \neq X')$. Then,

$$S(X|X') \leq h_2(p) + p\log(|\mathcal{X}| - 1).$$

Sketch of proof:

- Write $S(X|X') = \sum_{x \in \mathcal{X}} S_{\mathbb{P}|X'=x}(X)\mathbb{P}(X' = x)$.
- For each x, use the chain rule

$$\begin{aligned} S(X)_{\mathbb{P}|X'=x} &= S(X, I_{X=x})_{\mathbb{P}|X'=x} \\ &= S(I_{X=x})_{\mathbb{P}|X'=x} + S(X|I_{X=x})_{\mathbb{P}|X'=x} \\ &\leq h_2(\mathbb{P}(X = x|X' = x)) + \log(|\mathcal{X}| - 1)\mathbb{P}(X \neq x|X' = x). \end{aligned}$$

- Summation over x and concavity of h_2 yields the thesis.

Mutual information

- How to quantify the average gain of information of Bob about X, after receiving Y?
- Shannon proposed the mutual information:

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

- Intuitively, $I(X;Y) \geq 0$ (proof later). By definition,

$$\begin{aligned} I(X;Y) &= S(X) - (S(X,Y) - S(Y)) = S(X) + S(Y) - S(X,Y) \\ &= I(Y;X). \end{aligned}$$

- More explicit expression:

$$I(X;Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y) \log \left(\frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)\mathbb{P}(Y = y)} \right).$$

Relative entropy

- The last formula suggests replace the denominator with a general probability density.
- We define the relative entropy (or Kullback-Leibler divergence) of p with respect to q (both defined on a set $\mathcal{X}$) as

$$\begin{aligned} D_{KL}(p\|q) &= \sum_{x\in\mathcal{X}} p(x)\log(p(x)/q(x)) \\ &= \sum_{x\in\mathcal{X}} p(x)\left(\log p(x) - \log q(x)\right) \\ &= -S(p) + \sum_{x\in\mathcal{X}} p(x)\log q(x), \end{aligned}$$

- The above holds $p << q$, otherwise $D_{KL}(p\|q) = \infty$.
- The relative entropy can be conveniently thought as a “distance” between p, however it is not symmetric,

$$D_{KL}(p\|q) \neq D_{KL}(q\|p) \quad \text{(in general)}.$$

- D_{KL} enjoys natural monotonicity and convexity properties.
- Given Markov kernel $N(x,y)_{x\in\mathcal{X},y\in\mathcal{Y}}$, from $\mathcal{X}$ to $\mathcal{Y}$, the relative entropy decreases:

$$D_{KL}(N^\dagger p||N^\dagger q) \leq D_{KL}(p||q),$$

- By taking any kernel such that $N^\dagger p = N^\dagger q$, we obtain

$$0 = D_{KL}(N^\dagger p||N^\dagger q) \leq D_{KL}(p||q).$$

- Monotonicity implies also that

$$(p,q) \mapsto D_{KL}(p||q) \quad \text{is jointly convex.}$$

Maximum entropy distributions

- Given $E : \mathcal{X} \to \mathbb{R}$ and for $m \in \mathbb{R}$, what is the probability p on $\mathcal{X}$ which maximizes Shannon's entropy $S(p)$, with the constraint
$$\sum_{x \in \mathcal{X}} E(x)p(x) = m?$$
- For $\min E < m < \max E$, (the) answer is given by Gibbs distribution
$$p_\beta(x) = e^{-\beta E}/z,$$
where $\beta \in \mathbb{R}$ is a parameter, and
$$z = z(\beta) = \sum_{x \in \mathcal{X}} e^{-\beta E(x)}$$
is a normalization constant.
- Why? for every p,
$$D_{KL}(p||q_\beta) = -S(p) + \beta m + \log z(\beta) \geq 0.$$
- Example: The uniform distribution maximizes the entropy (put $E = 0$):
$$S(p) \leq \log |\mathcal{X}|.$$

- The mutual information $I(X;Y)$ is a special case of relative entropy:

$$I(X;Y) = D_{KL}(\mathbb{P}_{XY}||\mathbb{P}_X \otimes \mathbb{P}_Y) \geq 0$$

- This can be rewritten as subadditivity

$$S(X,Y) \leq S(X) + S(Y).$$

- Data processing inequality: given a Markov chain (X,Y,Z), i.e., X and Z are conditionally independent given Y, it holds

$$I(X;Z) \leq I(X;Y).$$

- Interpretation: by further transforming Y, Bob cannot increase the information received about X!

Proof of the data processing inequality

- By assumption, the joint law factorizes

$$\mathbb{P}_{XYZ}(x,y,z) = \mathbb{P}_{XY}(x,y)N(y,z),$$

where N is a Markov kernel from $\mathcal{Y}$ to $\mathcal{Z}$.

- Extend N to a kernel from $\mathcal{X} \times \mathcal{Y}$ to $\mathcal{X} \times \mathcal{Z}$ by acting trivially on $\mathcal{X}$,

$$\tilde{N}((x,y),(x',z)) = \delta_x(x')N(y,z),$$

- Check that

$$\tilde{N}^\dagger(\mathbb{P}_{XY}) = \mathbb{P}_{XZ}, \quad \tilde{N}^\dagger(\mathbb{P}_X \otimes \mathbb{P}_Y) = \mathbb{P}_X \otimes \mathbb{P}_Z.$$

Strong subadditivity

- Consider the case $Z = f(Y)$. Then,

$$I(X; f(Y)) \leq I(X; Y).$$

- Replacing Y with a joint variable (Y, Z) and letting $f(y, z) = y$, we obtain

$$I(X; Y) \leq I(X; (Y, Z)).$$

- The above is equivalent to

$$S(X|(Y, Z)) \leq S(X|Y),$$

or to the strong subadditivity property of the Shannon entropy

$$S(X, Y, Z) \leq S(X, Y) + S(Y, Z) - S(Y),$$

Plan

1. Distances (conclusion)
   - Quantum optimal transport
2. Entropy
   - Classical entropy
   - Quantum entropy

von Neumann entropy

- Consider a finite-dimensional quantum system H and a state $\rho \in \mathcal{S}(H)$. von Neumann defined its entropy as

$$S(\rho) = -\mathrm{tr}[\rho \log \rho],$$

  where $\rho \log \rho$ is obtained via functional calculus.

- $S(\rho)$ is Shannon entropy of the probability distribution associated to the spectrum of ρ (with multiplicities)
- Hence, $S(\rho) \geq 0$ with equality if and only if $\sigma(\rho) \subseteq \{0, 1\}$ is pure.
- Notation: $S(H)_\rho$ or simply $S(H)$ if the state ρ is understood.

Quantum relative entropy

- We introduce quantum relative entropy of ρ with respect to another state $\sigma \in \mathcal{S}(H)$ as
$$S(\rho||\sigma) = \mathrm{tr}[\rho(\log \rho - \log \sigma)],$$
where the operators $\rho \log \rho$ and $\log \sigma$ are defined via functional calculus.

- The formula above requires that the kernel of σ is contained in the kernel of ρ (recall that in the classical case we require $p << q$), we interpret
$$\rho(\log \rho - \log \sigma) = 0$$
on the kernel of ρ. Otherwise, $S(\rho||\sigma) = \infty$.

- If ρ and σ commute, then
$$S(\rho||\sigma) = D_{KL}(p||q),$$
where p, q are probability distribution associated to the spectra of ρ, σ.

Monotonicity of relative entropy

Theorem (data processing inequality, DPI)

Let

- $H, \tilde{H}$ *be quantum systems*
- $\Phi^\dagger$ *be a quantum channel from* H *to* $\tilde{H}$,
- $\rho, \sigma \in \mathcal{S}(H)$.

Then, it holds

$$S(\Phi^\dagger(\rho)||\Phi^\dagger(\sigma)) \leq S(\rho||\sigma).$$

Proof

- We use general differentiation trick (much employed in entropic inequalities).
- Let $f, g : [a, b] \to \mathbb{R}$ be such that, for $t \in [a, b]$

$$f(t) \leq g(t) \quad \text{and} \quad f(a) = g(a).$$

- If both f and g are (right-)differentiable at $t = a$, then

$$f'(a) \leq g'(a).$$

- By Lieb's concavity theorem, for $K = \mathbb{1}_{\tilde{H}}$, and $X = \rho$, $Y = \sigma$, $t \in [0, 1]$,

$$\mathrm{tr}[\rho^{1-t}\sigma^t] \leq \mathrm{tr}[\Phi^\dagger(\rho)^{1-t}\Phi^\dagger(\sigma)^t].$$

- For $t = 0$, we have equality (Φ is trace preserving).
- Assume for simplicity that ρ, σ, $\Phi^\dagger(\rho)$, $\Phi^\dagger(\sigma)$ are all invertible, then both sides in the inequality are smooth functions of t.
- We have

$$\frac{d}{dt}\Big|_{t=0^+} \mathrm{tr}[\rho^{1-t}\sigma^t] \leq \frac{d}{dt}\Big|_{t=0^+} \mathrm{tr}[\Phi^\dagger(\rho)^{1-t}\Phi^\dagger(\sigma)^t].$$

- We compute

$$\frac{d}{dt}\Big|_{t=0^+} \mathrm{tr}[\rho^{1-t}\sigma^t] = \mathrm{tr}[-\rho\log\rho + \rho\log\sigma] = -S(\rho||\sigma),$$

and similarly for the right hand side.

1. Consider any trivial channel that maps any state into the same state, e.g. $\Phi^\dagger(\rho) = \mathbb{1}_H/\dim(H)$: then

$$S(\rho||\sigma) \geq S(\mathbb{1}_H/\dim(H); \mathbb{1}_H/\dim(H)) = 0.$$

2. The quantum relative entropy is jointly convex, i.e.,

$$(\rho, \sigma) \mapsto S(\rho||\sigma) \quad \text{is convex.}$$

   Apply the DPI to the partial trace channel $\Phi^\dagger(M) = \mathrm{tr}_2[M]$ to

$$\rho = \begin{pmatrix} \rho_0 & 0 \\ 0 & \rho_1 \end{pmatrix}, \quad \sigma = \begin{pmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{pmatrix}.$$

3. For $E \in \mathcal{O}(H)$, Gibbs states $\rho_\beta = e^{-\beta H}/z$ for $\beta \in \mathbb{R}$, $z = \mathrm{tr}[e^{-\beta H}] > 0$ are a maximizer of von Neumann entropy (keeping fixed $(H)_\rho = \mathrm{tr}[E\rho_\beta]$).
4. In particular, von Neumann entropy always satisfies the inequalities

$$0 \leq S(H)_\rho \leq \dim(H).$$

Quantum conditional entropy

- The analogue of $S(X|Y)$ is a delicate quantity, since a "quantum conditional density" is not available.
- We impose the validity of the chain rule: given $\rho \in \mathcal{S}(H \otimes K)$ with educed density operator $\rho_H = \mathrm{tr}_K[\rho] \in \mathcal{S}(H)$, its *quantum conditional entropy* is

$$S(K|H)_\rho = S(\rho) - S(\rho_H) = S(H \otimes K)_\rho - S(H)_{\rho_H}.$$

- Notation $S(HK)_\rho = S(H \otimes K)_\rho$.
- Now the chain rule holds, but $S(H|K)$ may be strictly negative, because of entangled states!

Proposition (purification of a state)

Given $\rho \in \mathcal{S}(H)$, there exists an auxiliary quantum system K and a pure state $|\Psi\rangle\langle\Psi| \in \mathcal{S}(H \otimes K)$ such that

$$\mathrm{tr}_K[|\Psi\rangle\langle\Psi|] = \rho.$$

- The chain rule implies

$$0 = S(H \otimes K)_{|\Psi\rangle\langle\Psi|} = S(H)_\rho + S(K|H)_{|\Psi\rangle\langle\Psi|},$$

  hence the relative entropy must be negative in this case!

- This observation is turned into an indicator of entanglement (entanglement entropy).

Proof of purification

- Let $K = H^*$ be the dual of H, and consider the isomorphism

$$H \otimes H^* \ni |\psi\rangle \otimes \langle\varphi| \quad \mapsto \quad |\psi\rangle \langle\varphi| \in \mathcal{L}(H).$$

- The $|\Psi\rangle \in H \otimes H^*$ corresponding to $\sqrt{\rho} \in \mathcal{L}(H)$ is a purification of ρ.
- Pick orthonormal basis $(|i\rangle)_{i\in I}$ of eigenvectors of ρ and write

$$\sqrt{\rho} = \sum_{i\in I} \sqrt{p_i}\, |i\rangle \langle i| \,,$$

hence

$$|\Psi\rangle = \sum_{i\in I} \sqrt{p_i}\, |i\rangle \otimes \langle i| \,.$$

- Since $|\Psi\rangle \langle\Psi| = \sum_{i,j\in I} \sqrt{p_i p_j}(|i\rangle \otimes \langle i|)(\langle j| \otimes |j\rangle)$, taking the partial trace

$$\mathrm{tr}_K[|\Psi\rangle \langle\Psi|] = \sum_{i\in I} p_i\, |i\rangle \langle i| = \rho.$$

Quantum mutual information

- To define the quantum mutual information, we mimic the classical case: given $\rho \in \mathcal{S}(H \otimes K)$ with reduced density operators $\rho_H \in \mathcal{S}(H)$, $\rho_K \in \mathcal{S}(K)$,

$$\begin{aligned} I(H;K)_\rho &= S(\rho || \rho_H \otimes \rho_K) \\ &= S(H)_{\rho_H} - S(H|K)_\rho \\ &= S(H)_{\rho_H} + S(K)_{\rho_K} - S(H \otimes K)_\rho. \end{aligned}$$

- From the DPI: given $\rho \in \mathcal{S}(H \otimes K)$ and a quantum channel $\Phi^\dagger$ from K to $\tilde{K}$, then

$$I(H;\tilde{K})_{\mathbb{1}_{\mathcal{L}(H)} \otimes \Phi^\dagger(\rho)} \leq I(H;K)_\rho$$

- Replace K with $K \otimes L$ and let $\Phi^\dagger = \mathrm{tr}_L$ be the partial trace channel: for every $\rho \in \mathcal{S}(H \otimes K \otimes L)$,

$$I(H;K)_{\rho_{HK}} \leq I(H;K \otimes L)_\rho,$$

  which is equivalent to the strong subadditivity of von Neumann entropy

$$S(H \otimes K \otimes L) \leq S(H \otimes K) + S(K \otimes L) - S(K).$$