Lecture 6: Joint convexity and relative entropy CSE 599I, Spring 2021
Instructor: James R. Lee

1 Relative entropy

In the classical setting, if $p, q \in \mathbb{R}_+^{\mathcal{X}}$ are two nonnegative vectors on a finite set $\mathcal{X}$ one defines the *relative entropy*

$$\mathsf{D}(p \,\|\, q) := \sum_{x \in \mathcal{X}} p_x \log \frac{p_x}{q_x} + \sum_{x \in \mathcal{X}} (q_x - p_x).$$

We take the value to be $+\infty$ if there is some $x \in \mathcal{X}$ with $p_x > 0$ but $q_x = 0$. Restricted to probability densities, i.e., $\sum_{x \in \mathcal{X}} p_x = \sum_{x \in \mathcal{X}} q_x = 1$, the latter sum vanishes.

One can think of this as a measure of the average log-surprise (measured in bits, say, or nats—I suppose—since here log refers to the natural log) when receiving a sample from p when you expected a sample from q. Stated differently, if we decide on a Shannon-optimal code for sending messages whose symbols are sampled i.i.d. from q, then $\mathsf{D}(p \,\|\, q)$ is the average communication per symbol that we'll use if we instead have to send messages whose symbols are sampled i.i.d. from p.

Stein's Lemma in hypothesis testing tells us that we can also think of $\mathsf{D}(p \,\|\, q)$ in the following way. Suppose we are given $X_1, X_2, \ldots, X_n$ all sampled i.i.d. from p or i.i.d. from q. We need to design a test to decide, given the samples, which distribution they came from.

In this setting, q is the null hypothesis, so our method should accept $X_1, \ldots, X_n \sim q$ with probability 1 as $n \to \infty$, and our goal is to minimize the probability of accidentally classifying $X_1, \ldots, X_n \sim p$ as coming from q. In other words, failure is when we don't correctly reject the null hypothesis. The (asymptotically) optimal test with this property has a failure rate of $e^{-n(\mathsf{D}(p \,\|\, q)+o(1))}$ as $n \to \infty$.

Joint convexity. It always holds that $\mathsf{D}(p \,\|\, q) \geqslant 0$, and $\mathsf{D}(p \,\|\, q) = 0 \iff p = q$. One way to see the first fact is as follows: The function $\varphi(p) = \sum_{x \in \mathcal{X}} (p_x \log p_x - p_x)$ is convex on nonnegative p, as one can verify from the second derivative test, and

$$(\nabla \varphi(p))_x = \log p_x.$$

Therefore we have

$$\mathsf{D}(p \,\|\, q) = \varphi(p) - \varphi(q) - \langle \nabla \varphi(q), p - q \rangle \geqslant 0.$$

(This holding for all $p, q \in \mathbb{R}_+^{\mathcal{X}}$ is equivalent to φ being convex on $\mathbb{R}_+^{\mathcal{X}}$.) Moreover, since φ is actually *strictly convex* on the strictly positive orthant, we have $\mathsf{D}(p \,\|\, q) = 0 \iff p = q$. (It takes a bit of care to verify that this holds even for $p, q \in \mathbb{R}_+^{\mathcal{X}}$ which can possibly be 0 in some coordinates.)

It turns out the function $(p, q) \mapsto \mathsf{D}(p \,\|\, q)$ is *jointly convex* on $\mathbb{R}_+^{\mathcal{X}} \times R_+^{\mathcal{X}}$ in the following sense: For all $p_1, p_2, q_1, q_2 \in \mathbb{R}_+^{\mathcal{X}}$ and $t \in [0, 1]$, it holds that

$$\mathsf{D}\left((1-t)p_1 + tp_2 \,\|\, (1-t)q_1 + tq_2\right) \leqslant (1-t)\mathsf{D}(p_1 \,\|\, p_2) + t\mathsf{D}(q_1 \,\|\, q_2).$$

This has a nice interpretation in terms of hypothesis testing: Suppose with probability t, I choose $i = 1$ and with probability $1 - t$, I choose $i = 2$, and then I sample $X_1, \ldots, X_n$ either from p_i or q_i, with your goal being to decide whether the samples came from from the P side or the Q side. The RHS represents the optimal error exponent when I *tell you whether $i = 1$ or $i = 2$*, wile the LHS represents the optimal error exponent if I don't (so you just see samples from the mixture). It should be intuitively obvious that it's harder without that information, as the inequality asserts.

To prove it, simply note that $\mathsf{D}(p \,\|\, q)$ is a sum of terms of the form $p_x \log \frac{p_x}{q_x} + (q_x - p_x)$, so it suffices to prove that the function

$$g(a, b) = a \log \frac{a}{b} + (b - a)$$

is convex on $\mathbb{R}^2_+$. We can do this by evaluating the Hessian:

$$\nabla^2 g(a, b) = \begin{bmatrix} 1/a & -1/b \\ -1/b & a/b^2 \end{bmatrix}.$$

If λ_1, λ_2 are the eigenvalues of $\nabla^2 g(a, b)$, then evaluating the trace gives $\lambda_1 + \lambda_2 \geqslant 0$, and since the determinant is zero, we have $\lambda_1 \lambda_2 = 0$, thus $\nabla^2 g(a, b) \succeq 0$, and g is convex.

1.1 The quantum relative entropy

For $A \succeq 0$, define the (negative) entropy

$$\Phi(A) = \mathrm{Tr}(A \log A - A),$$

with the convention that $0 \log 0 = 0$. Noting that $\nabla \Phi(A) = \log A$ for $A \succ 0$, the corresponding relative entropy can then be defined analogously as

$$\mathsf{S}(A \,\|\, B) := \Phi(A) - \Phi(B) - \mathrm{Tr}(\nabla \Phi(B)(A - B)) = \mathrm{Tr}\left(A(\log A - \log B) + (B - A)\right). \qquad (1.1)$$

Lemma 1.1. *For $A, B \in \mathbf{H}^n_+$, it holds that $\mathsf{S}(A \,\|\, B) \geqslant 0$ and $\mathsf{S}(A \,\|\, B) = 0 \iff A = B$.*

As we argued in the scalar case, convexity of Φ on $\mathbf{H}^n_+$ gives $\mathsf{S}(A \,\|\, B) \geqslant 0$ for all $A, B \in \mathbf{H}^n_+$. We have already seen that $A \mapsto A \log A$ is operator convex on PSD matrices, which is stronger than Φ being convex, but the fact that Φ is convex follows from a much more general statement. In fact, since φ is strictly convex, the same holds for Φ.

Lemma 1.2 (Trace convexity). *Suppose that $f : I \to \mathbb{R}$ is continuous and convex. Then the map $A \mapsto \mathrm{Tr}(f(A))$ is convex on Hermitian matrices with $\mathrm{spec}(A) \subseteq I$. If f is strictly convex, then so is $A \mapsto \mathrm{Tr}(f(A))$.*

Proof. Consider a Hermitian matrix X and write $X = \sum_{k=1}^n \lambda_k v_k v_k^*$, where $\{v_k\}$ is an orthonormal basis. Let $\{u_1, \dots, u_n\}$ be any orthonormal basis of $\mathbb{C}^n$. Then we have

$$\begin{aligned}
\mathrm{Tr}(f(X)) = \sum_{j=1}^n \langle u_j, f(X) u_j \rangle &= \sum_{j=1}^n \left\langle u_j, \sum_{k=1}^n f(\lambda_k) v_k v_k^* u_j \right\rangle \\
&= \sum_{j=1}^n \left(\sum_{k=1}^n |\langle u_j, v_k \rangle|^2 \sum_{k=1}^n f(\lambda_k) \right) \\
&\geqslant \sum_{j=1}^n f\left(\sum_{k=1}^n |\langle u_j, v_k \rangle|^2 \sum_{k=1}^n \lambda_k \right) \\
&= \sum_{j=1}^n f\left(\langle u_j, (\lambda_1 v_1 v_1^* + \cdots + \lambda_n v_n v_n^*) u_j \rangle \right) = \sum_{j=1}^n f\left(\langle u_j, X u_j \rangle \right), \qquad (1.2)
\end{aligned}$$

where the inequality uses convexity of f and the fact that $\sum_{k=1}^n |\langle u_j, v_k\rangle|^2 = 1$ for every j. Note that if f is strictly convex, then the inequality is only tight when $\{u_1, \ldots, u_n\} = \{v_1, \ldots, v_n\}$.

Consider now $A, B \in \mathbf{H}^n$ with $\operatorname{spec}(A), \operatorname{spec}(B) \subseteq I$, and let $\{u_j\}$ be an orthornormal basis of eigenvectors of $(A+B)/2$. Then,

$$\begin{aligned}
\operatorname{Tr}\left(f\left(\tfrac{A+B}{2}\right)\right) &= \sum_{j=1}^n \langle u_j, f(\tfrac{A+B}{2}), u_j\rangle \\
&= \sum_{j=1}^n f\left(\langle u_j, \tfrac{A+B}{2} u_j\rangle\right) \\
&= \sum_{j=1}^n f\left(\tfrac{1}{2}\langle u_j, Au_j\rangle + \tfrac{1}{2}\langle u_j, Bu_j\rangle\right) \\
&\leqslant \sum_{j=1}^n \left[\tfrac{1}{2} f(\langle u_j, Au_j\rangle) + \tfrac{1}{2} f(\langle u_j, Bu_j\rangle)\right] \\
&\leqslant \tfrac{1}{2}\operatorname{Tr}(f(A)) + \tfrac{1}{2}\operatorname{Tr}(f(B)),
\end{aligned}$$

where the first inequality again uses (midpoint) convexity of f, and the second uses (1.2) applied with $X = A$ and $X = B$. This implies that $A \mapsto \operatorname{Tr}(f(A))$ is midpoint convex, and since f is continuous, an approximation argument shows that the map is genuinely convex.

Note that if f is strictly convex and the first inequality holds with equality, then $\langle u_j, Au_j\rangle = \langle u_j, Bu_j\rangle$. Moreover, if the second inequality holds with equality, then by our previous observation, it must also be that $\{u_j\}$ is a basis of eigenvectors for both A and B, implying that $A = B$. Thus $A \mapsto \operatorname{Tr}(f(A))$ is strictly convex as well. □

It turns out that $\mathsf{S}(A \,\|\, B)$ describes the optimal asymptotic error exponent for *quantum hypothesis testing.* When $A, B \succeq 0$ and $\operatorname{Tr}(A) = \operatorname{Tr}(B) = 1$, A and B are density matrices that describe the states of quantum systems. The corresponding fact about joint convexity is deeper than in the classical case, but still true.

Theorem 1.3. *The map $(A, B) \mapsto \mathsf{S}(A \,\|\, B)$ is jointly convex on $\mathbf{H}_+^n \times \mathbf{H}_+^n$.*

Note that the classical proof of joint convexity breaks down for matrices. That's because we used the fact that the relative entropy $\mathsf{D}(p \,\|\, q)$ is *separable* in the scalar setting; it is a sum of two-variate functions. Thus we could establish convexity in two dimensions, where the Hessian calculation was straightforward. For A, B that are not simultaneously diagonalizable, we cannot write $\mathsf{S}(A \,\|\, B)$ as a sum over lower-dimensional terms.

Before addressing the proof, let us argue that it establishes our goal from the last lecture.

Theorem 1.4 (Lieb's concavity theorem)**.** *For every Hermitian H, the map $X \mapsto \operatorname{Tr}(e^{H+\log X})$ is concave on $\mathbf{H}_+^n$.*

To argue that Theorem 1.3 implies Theorem 1.4, we need a standard observation.

Lemma 1.5. *If $\mathcal{A}$ and $\mathcal{B}$ are convex sets and the mapping $(A, B) \to F(A, B)$ is jointly concave on $\mathcal{A} \times \mathcal{B}$, then*

$$f(A) := \sup\{F(A, B) : B \in \mathcal{B}\}$$

is a concave function on $\mathcal{A}$.

Proof. Consider $A_1, A_2 \in \mathcal{A}$ and let $B_1, B_2 \in \mathcal{B}$ be such that $f(A_1) \leqslant F(A_1, B_1) + \varepsilon$ and $f(A_2) \leqslant F(A_2, B_2) + \varepsilon$. Then,

$$\begin{aligned} tf(A_1) + (1-t)f(A_2) &\leqslant tF(A_1, B_1) + (1-t)F(A_2, B_2) + \varepsilon \\ &\leqslant F(tA_1 + (1-t)A_2, tB_1 + (1-t)B_2) + \varepsilon \\ &\leqslant f(tA_1 + (1-t)A_2) + \varepsilon, \end{aligned}$$

and sending $\varepsilon \to 0$ completes the proof. ◻

Recalling (1.1) and $\mathsf{S}(A \,\|\, B) \geqslant 0$ on $\mathbf{H}_+^n \times \mathbf{H}_+^n$, for $B \succeq 0$, we have

$$0 = \min \{\mathsf{S}(A \,\|\, B) : A \succeq 0\} = \min \left\{\operatorname{Tr}\left(A(\log A - \log B) + (B - A)\right) : A \succeq 0\right\},$$

which we can rewrite as

$$\operatorname{Tr}(B) = \max \left\{\operatorname{Tr}(A \log B - A \log A + A) : A \succeq 0\right\}$$

Now substitute $B := e^{H + \log X}$, yielding

$$\begin{aligned} \operatorname{Tr}\left(e^{H+\log X}\right) &= \max \left\{\operatorname{Tr}\left(A(H + \log X) - A\log A + A\right) : A \succeq 0\right\} \\ &= \max \{\operatorname{Tr}(AH) - \mathsf{S}(A \,\|\, X) + \operatorname{Tr}(A) : A \succeq 0\} \end{aligned}$$

Since $\mathsf{S}(\cdot \,\|\, \cdot)$ is jointly convex by Theorem 1.3, it holds that $F(X, A) := \operatorname{Tr}(AH) - \mathsf{S}(A \,\|\, X) + \operatorname{Tr}(X)$ is jointly concave, and thus by Lemma 1.5, the function $X \mapsto \max\{F(X, A) : A \succeq 0\}$ is concave, proving Theorem 1.4 (from Theorem 1.3).

1.2 The parallel sum is jointly concave

In the last lecture, we saw that the map $B \mapsto B^{-1}$ is operator convex on positive matrices. We will need the following generalization.

Lemma 1.6. *The map $(X, B) \mapsto XB^{-1}X^*$ is jointly convex on $\mathbb{M}_n(\mathbb{C}) \times \mathbf{H}_{++}^n$.*

Note that this is a generalization of the fact that $(x, y) \mapsto x^2/y$ is convex for $y > 0$, and captures convexity of both $B \mapsto B^{-1}$ and $X \mapsto X^2$. The following tool will be useful.

Lemma 1.7. *Consider $A, B \succ 0$. Then the block matrix $\left[\begin{smallmatrix} A & X \\ X^* & B \end{smallmatrix}\right]$ is PSD if and only if $A \succeq XB^{-1}X^*$.*

Proof. We have

$$\begin{bmatrix} I & -XB^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A & X \\ X^* & B \end{bmatrix} \begin{bmatrix} I & 0 \\ -B^{-1}X^* & I \end{bmatrix} = \begin{bmatrix} A - XB^{-1}X^* & 0 \\ 0 & B \end{bmatrix}.$$

Note that if $Y \in \mathbf{H}^n$ and U is invertible, then $Y \succ 0 \iff U^*YU \succ 0$, hence $\left[\begin{smallmatrix} A & X \\ X^* & B \end{smallmatrix}\right]$ is positive if and only if the RHS is positive, which is clearly equivalent to $A \succeq XB^{-1}X^*$. ◻

Now we can prove Lemma 1.6 by applying Lemma 1.7 with the proper choice of matrices. For $B_1, B_2 \succ 0$, take

$$\begin{aligned} B &:= (B_1 + B_2)/2 \\ X &:= (X_1 + X_2)/2 \end{aligned}$$

$$A := \frac{X_1 B_1^{-1} X_1^* + X_2 B_2^{-1} X_2^*}{2}.$$

If the corresponding block matrix $\left[\begin{smallmatrix} A & X \\ X^* & B \end{smallmatrix}\right]$ is positive, the resulting inequality $A \geq XB^{-1}X^*$ is precisely midpoint joint concavity of $XB^{-1}X^*$, from which joint concavity follows by continuity.

But observe that $\left[\begin{smallmatrix} X_i B_i^{-1} X_i & X_i \\ X_i^* & B_i \end{smallmatrix}\right]$ is positive for $i \in \{1, 2\}$ by Lemma 1.7, hence their average $\left[\begin{smallmatrix} A & X \\ X^* & B \end{smallmatrix}\right]$ is positive, completing the proof of Lemma 1.6.

Parallel sums. Define the *parallel sum* of two positive matrices by

$$A : B \ := \ (A^{-1} + B^{-1})^{-1}.$$

Lemma 1.8. *It holds that $(A, B) \mapsto A : B$ is jointly operator concave on $\mathbf{H}^n_{++}$.*

Proof. Let us establish that

$$A : B = A - A(A + B)^{-1} A. \tag{1.3}$$

Then joint concavity follows from Lemma 1.6 (with the substitution $X \leftarrow A, B \leftarrow A + B$). Note that (1.3) is true for positive numbers, i.e., $(a^{-1} + b^{-1})^{-1} = \frac{ab}{a+b} = a - a^2/(a + b)$. Thus as was pointed out in class[1], to prove (1.3), it suffices to prove it equivalent to an expression involving commuting matrices.

Multiplying both sides of (1.3) by $A^{-1/2}$ gives

$$A^{-1/2}(A^{-1} + B^{-1})^{-1} A^{-1/2} = I - A^{1/2}(A + B)^{-1} A^{1/2},$$

which is equivalent to

$$(I + A^{1/2} B^{-1} A^{1/2})^{-1} = I - (I + A^{-1/2} B A^{-1/2})^{1/2},$$

as desired. □

[1]Thanks Farzam.