Entropy and Ergodic Theory
Lecture 5: Joint typicality and conditional AEP

1 Notation: from RVs back to distributions

Let \((\Omega, \mathcal{F}, P)\) be a probability space, and let \(X\) and \(Y\) be \(A\)- and \(B\)-valued discrete RVs, respectively. Then \(p_{X,Y} \in \text{Prob}(A \times B)\), and we may write it using the multiplication rule for conditional probabilities:

\[ p_{X,Y}(a, b) = p_X(a) \cdot p_{Y|X}(b|a). \]

In various applications of the theory, we need to fix one of the two factors on the right of this equation, and consider all the possible joint distributions that can be obtained by varying the other factor.

**Definition 1.1.** A **kernel** from \(A\) to \(B\) is an \(A \times B\) matrix \(\theta = (\theta(b|a))_{a,b}\) such that the row vector \((\theta(b|a))_{b \in B}\) is a probability vector on \(B\) for each \(a \in A\).

Later in this course we extend this definition to more general measurable spaces \(A\) and \(B\).

The \(n^{th}\) **extension** of a kernel \(\theta\) is the kernel \(\theta^n\) from \(A^n\) to \(B^n\) defined by

\[
\theta^n(y_1 \cdots y_n \mid x_1 \cdots x_n) = \prod_{i=1}^{n} \theta(y_i \mid x_i).
\]

Each of the functions \(\theta^n(\cdot \mid x)\) for \(x \in A^n\) is a product probability distribution on \(B^n\).

A kernel \(\theta\) may be seen as specifying a collection of ‘conditional probabilities’ \(\theta(\cdot \mid a), a \in A\), but not a probability distribution on \(A\) itself. If \((X, Y)\) are as above, then we say that \(p_{Y|X}\) **agrees with** a kernel \(\theta\) if

\[ p_{Y|X}(b|a) = \theta(b|a) \quad \text{whenever } p_X(a) > 0. \]
If \( p \in \text{Prob}(A) \) and \( \theta \) is a kernel from \( A \) to \( B \), then they can be combined to give a probability distribution on \( A \times B \):

\[
q(a, b) := p(a) \cdot \theta(b | a).
\]

This is sometimes called the **input-output distribution** of \( p \) and \( \theta \), or the **hookup** of \( p \) with \( \theta \). We sometimes denote if by the (non-standard) notation \( p \upharpoonleft \theta \), which is meant to suggest a kind of generalization of a product measure. If \( (X,Y) \) are an \((A \times B)\)-valued RV with this distribution, then \( p_X = p \) and \( p_{Y|X} \) agrees with \( \theta \). Any pair of RVs \((X,Y)\) with this distribution is called a **randomization** of \( p \) and \( \theta \).

If \((X,Y)\) is a randomization of \( p \) and \( \theta \), then the conditional entropy \( H(Y | X) \) clearly depends only on the joint distribution \( p \upharpoonleft \theta \). We sometimes use the alternative notation

\[
H(\theta | p) := H(Y | X) = \sum_{a \in A} p(a) \cdot H(\theta(\cdot | a)).
\]

Since this equals

\[
H(X,Y) - H(X) = H(p \upharpoonleft \theta) - H(p),
\]

it is clearly a continuous function of the joint distribution \( p \upharpoonleft \theta \).

## 2 Joint typicality and the conditional AEP

Let \( q \in \text{Prob}(A \times B) \), and let \( p_A \) and \( p_B \) its marginals on \( A \) and \( B \). In order to emphasize that \( q \) is a distribution on a product of two alphabets, we sometimes refer to elements of \( T_{n,\delta}(q) \) as **jointly \( \delta \)-typical** for \( q \). This nomenclature also emphasizes the following important point:

If \( x \in T_{n,\delta}(p_A) \) and \( y \in T_{n,\delta}(p_B) \), it does not follow that \((x,y)\) is jointly \( \varepsilon \)-typical for \( q \) for some small \( \varepsilon \).

**Example.** Let \( x \in \{0,1\}^n \) be \( \delta \)-typical for the uniform distribution on \( \{0,1\} \), and let \( y = x \). Then \((x,y)\) is very far from typical for the uniform distribution on \( \{0,1\} \times \{0,1\} \), since

\[
N(01 | (x,y)) = N(10 | (x,y)) = 0 \neq n/4.
\]

Instead, \((x,y)\) is typical for the distribution \( p = \frac{1}{2}(\delta_{00} + \delta_{11}) \) on \( \{0,1\} \times \{0,1\} \).
Our next goal is a counting-problem interpretation of conditional entropy, analogous to the meaning established for unconditional entropy in Lecture 1. In proving it, we will also meet conditional versions of the LLN for types and the AEP. We build up all these results following roughly the same sequence as in the unconditional setting of Lecture 1.

Let us now write

\[ q = p_A \times \theta \]

for some kernel \( \theta \) from \( A \) to \( B \). We fix all of these notations for the rest of the lecture.

### 2.1 Conditional typicality

**Definition 2.1.** Let \( x \in A^n \) and \( y \in B^n \). Then the conditional type of \( y \) given \( x \) is the collection of values

\[ p_{y|x}(b|a) := \frac{N((a,b) \mid (x,y))}{N(a \mid x)}, \]

defined for all \( a \in A \) such that \( N(a \mid x) > 0 \).

Henceforth, we usually abbreviate \( N(ab \mid xy) := N((a,b) \mid (x,y)) \).

Intuitively, the quantity \( p_{y|x}(b|a) \) does the following: among all \( i \in \{1, 2, \ldots, n\} \), we consider only those for which \( x_i = a \), and among those we record the fraction which also satisfy \( y_i = b \).

Sometimes we refer to the conditional type as being indexed by all \( a \in A \), without worrying about whether \( N(a \mid x) > 0 \). We do this only in cases where it causes no real problems.

The first part of our interpretation of conditional entropy is the following. It is the analog of the upper bound in Theorem 3.1 from Lecture 1.

**Lemma 2.2.** Fix a kernel \( \theta \) and a string \( x \in A^n \). The number of strings \( y \in B^n \) for which \( p_{y|x} \) agrees with \( \theta \) is at most \( 2^{H(\theta \mid p_x)n} \).

**Proof.** Exercise: mimic the proof of the upper bound in Lecture 1, Theorem 3.1. The key fact is that, if \( p_{y|x} \) agrees with \( \theta \), then

\[ N(ab \mid xy) = \theta(b|a)p_x(a)n \quad \text{for } a \in A, \ b \in B. \]
Like Theorem 3.1 in Lecture 1, there is a matching lower bound for this lemma, but we leave it aside here. Instead let us introduce an approximate form of the problem, and prove upper and lower bounds for that. The approximate form is more important for later applications.

**Definition 2.3.** Let $\delta > 0$ and $x \in A^n$. Then $y \in B^n$ is **conditionally $\delta$-typical for $\theta$ given $x$** if

$$\|p_{x,y} - p_x \times \theta\| < \delta.$$ 

The set of all such $y$ is denoted by $T_{n,\delta}(\theta, x)$.

More expansively, $y \in T_{n,\delta}(\theta, x)$ if

$$\sum_{a,b} \left| \frac{N(ab|x)}{n} - \frac{N(a|x)}{n} \frac{\theta(b|a)}{\theta(a)} \right| < \delta.$$ 

Beware that some authors use slightly different definitions, as in the case of unconditional typicality.

This definition requires just a little care for the following reason: if $\theta$ and $\lambda$ are two kernels satisfying $\theta(b|a) = \lambda(b|a)$ whenever $N(a|x) > 0$, then $y$ is conditionally $\delta$-typical for $\theta$ given $x$ if and only if the same holds for $\lambda$.

As in the unconditional setting, the upper bound of Lemma 2.2 is easily turned into an upper bound for the problem of counting approximately conditionally typical strings.

**Corollary 2.4.** For $x \in A^n$ and $\delta > 0$ we have

$$|T_{n,\delta}(\theta, x)| \leq 2^{H(\theta | p_x)n + \Delta(\delta)n + o(n)},$$

where the estimates in the two error terms of the exponent are independent of $x$.

**Proof.** The quantity $H(\theta | p)$ is continuous as a function of the joint distribution $p \times \theta$. From this and Lemma 2.2 it follows that

$$|\{y : p_{y|x} \text{ agrees with } \lambda\}| \leq 2^{H(\lambda | p_x)n} \leq 2^{H(\theta | p_x)n + \Delta(\delta)n}$$

whenever $\|p_x \times \lambda - p_x \times \theta\| < \delta$.

The values $p_{y|x}(b|a)$ taken by the conditional type of $y$ given $x$ are all ratios of integers between 0 and $n$, and there are at most $|A \times B|$ of these values. Therefore
we may choose a set $\mathcal{C}$ of at most $(n + 1)^{2|A \times B|}$ rational stochastic matrices such that every conditional type between strings of length $n$ agrees with some $\lambda \in \mathcal{C}$. Having done so, we obtain

$$|T_{n,\delta}(\theta, x)| \leq \sum_{\lambda \in \mathcal{C}, \|p_x \times \lambda - p_x \times \theta\| < \delta} 2^{H(\lambda | p_x)n} \leq \sum_{\lambda \in \mathcal{C}, \|p_x \times \lambda - p_x \times \theta\| < \delta} 2^{H(\theta | p_x)n+\Delta(\delta)n} \leq 2^{H(\theta | p_x)n+\Delta(\delta)n+o(n)}.$$

\hfill $\square$

In the remainder of this section, we introduce an LLN and AEP for conditional types, and obtain a lower bound which complements the previous corollary.

### 2.2 A LLN for conditional types

We need a version of the weak law of large numbers for independent RVs of possibly different distributions. We give a very simple version which assumes a uniform bound on the RVs. It can be proved by the same application of Chebyshev’s inequality as in the i.i.d. case (which in fact requires only a uniform bound on the variances).

**Proposition 2.5** (Weak LLN for RVs with differing distributions). Fix $M > 0$. For any $\varepsilon > 0$ there is an $n_0 \in \mathbb{N}$, depending on $M$ and $\varepsilon$, with the following property: whenever $n \geq n_0$ and $X_1, \ldots, X_n$ are independent $[-M, M]$-valued RVs, we have

$$\mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}\left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] \right| > \varepsilon \right\} < \varepsilon.$$

\hfill $\square$

**Proposition 2.6** (LLN for conditional types). For any $\delta > 0$, we have

$$\min_{x \in A^n} \theta^n(T_{n,\delta}(\theta, x) | x) \longrightarrow 1$$

as $n \rightarrow \infty$. 

5
Proof. Fix $x \in A^n$, and let $Y \in B^n$ be drawn from the distribution $\theta^n(\cdot | x)$. This means that $Y_1, \ldots, Y_n$ are independent and that $Y_i$ has distribution $\theta(\cdot | x_i)$. For any $(a, b) \in A \times B$ we have

$$p_{x,Y}(a, b) = \frac{N(ab | xY)}{n} = \frac{1}{n} \sum_{i=1}^n 1\{x_i = a, Y_i = b\}.$$ 

The RVs $1\{x_i = a, Y_i = b\}$ are independent, but in general not identically distributed: for those $i$ which satisfy $x_i = a$ the distribution is Bernoulli($\theta(b|a)$), and for the other $i$ this RV is identically zero. However, it does follow that they are all bounded by 1, and so Proposition 2.5 gives

$$\mathbb{P}\left\{ \left| p_{x,Y}(a, b) - \mathbb{E}[p_{x,Y}(a, b)] \right| < \delta \right\} \to 1 \quad \forall \delta > 0,$$

where the rate of convergence does not depend on the specific choice of $x$. This gives the correct limiting behaviour for $p_{x,Y}(a, b)$ because

$$\mathbb{E}[p_{x,Y}(a, b)] = \frac{1}{n} \sum_{i: x_i = a} \mathbb{E}[1\{Y_i = b\}] = \frac{1}{n} \sum_{i: x_i = a} N(a | x) \theta(b | a) = (p_x \bowtie \theta)(a, b).$$

Combining this convergence for the finitely many choices of $(a, b)$ completes the proof. \qed

2.3 Conditional entropy typicality and AEP

There is also a conditional version of entropy typicality.

Definition 2.7. A string $y \in B^n$ is **conditionally $\varepsilon$-entropy typical for $\theta$ given $x$** if

$$2^{-H(\theta | p_x) n - \varepsilon n} < \theta^n(y | x) < 2^{-H(\theta | p_x) n + \varepsilon n}.$$ 

The set of all such $y$ is denoted by $T_{\text{ent}}^{n,\varepsilon}(\theta, x)$.

The choice of the constant $H(\theta | p_x)$ here is justified by the following.

Theorem 2.8 (Conditional AEP). For any $\varepsilon > 0$ we have

$$\min_{x \in A^n} \theta^n(T_{\text{ent}}^{n,\varepsilon}(\theta, x) | x) \to 1$$

as $n \to \infty$.  

6
Proof. Let \( Y \sim \theta^n(\cdot | x) \) as before. Taking logarithms gives

\[
\log_2 \theta^n(Y \mid x) = \sum_{i=1}^{n} Z_i,
\]

where \( Z_i = \log_2 \theta(Y_i \mid x_i) \). These are independent, finite-valued RVs. Among them, there are at most \(|A|\)-many different possible distributions, one for each of the possible values of \( x_i \). This list of possible distributions depends on \( \theta \), but the specific choice of \( x \) determines only which distribution is chosen for each \( Z_i \). Therefore Proposition 2.5 gives

\[
P \left\{ \left| \frac{1}{n} \log_2 \theta^n(Y \mid x) - E \left[ \frac{1}{n} \log_2 \theta^n(Y \mid x) \right] \right| < \delta \right\} \longrightarrow 1 \quad \forall \delta > 0,
\]

where the rate of convergence again does not depend on the specific choice of \( x \). This completes the proof, because

\[
E[\log_2 \theta^n(Y \mid x)] = \sum_{i=1}^{m} E[\log_2 \theta(Y_i \mid x_i)] = \sum_{i=1}^{n} \sum_{b} \theta(b \mid x_i) \log_2 \theta(b \mid x_i)
\]

\[
= -\sum_{i=1}^{n} H(\theta(\cdot \mid x_i)) = -n \sum_{a} p_x(a) \cdot H(\theta(\cdot \mid a)) = -n H(\theta \mid P_x).
\]

\( \square \)

From this we obtain a covering-number corollary exactly as in the unconditional setting.

**Corollary 2.9.** For any \( \varepsilon > 0 \), we have

\[
\text{cov}_\varepsilon(\theta^n(\cdot \mid x)) = 2^{H(\theta \mid P_x)n + o_\varepsilon(n)},
\]

where the estimate in ‘\( o_\varepsilon(n) \)’ also depends on \( \theta \) but not on \( x \). \( \square \)

This now implies the lower bound on the number of conditionally typical strings which accompanies Corollary 2.4

**Corollary 2.10.** For any \( x \in A^n \) we have

\[
|T_{n,\delta}(\theta, x)| \geq 2^{H(\theta \mid P_x)n - o_\delta(n)},
\]

where the estimate in ‘\( o_\delta(n) \)’ depends on \( \theta \) but not on \( x \).

**Proof.** Combine Proposition 2.6 with the previous corollary. \( \square \)
2.4 Intuitive picture of the above results

Let \((X, Y)\) be a randomization of \(q\). The original AEP lets us roughly visualize \(q^n\) as a uniform distribution on typical strings. If we wish to think about \(q^n\) as a joint distribution for pairs of strings \((x, y)\), then the results of this section give us an enhanced version of that picture. To explain this, we start with the following.

The following lemma gives some simple but useful relations between conditional and unconditional typicality.

**Lemma 2.11.**

1. If \(x \in T_{n,\delta}(p)\) and \(y \in T_{n,\delta}(\theta, x)\) then \((x, y) \in T_{n,2\delta}(p \times \theta)\).

2. On the other hand, if \((x, y) \in T_{n,\delta}(p \times \theta)\), then \(y \in T_{n,2\delta}(\theta, x)\).

**Proof.** The key is that for any \(p' \in \text{Prob}(A)\) we have

\[
\|p' \times \theta - p \times \theta\| = \sum_{a,b} |p'(a)\theta(b|a) - p(a)\theta(b|a)|
\]

\[
= \sum_a |p'(a) - p(a)| \left(\sum_b \theta(b|a)\right) = \|p' - p\|.
\]

**Part 1.** Using the calculation above, our first pair of assumptions gives

\[
\|p_{x,y} - p \times \theta\| \leq \|p_{x,y} - p_x \times \theta\| + \|p_x \times \theta - p \times \theta\|
\]

\[
= \|p_{x,y} - p_x \times \theta\| + \|p_x - p\| < 2\delta.
\]

**Part 2.** Clearly if \((x, y) \in T_{n,\delta}(p \times \theta)\), then \(x \in T_{n,\delta}(p)\). Given this, we obtain

\[
\|p_{x,y} - p_x \times \theta\| \leq \|p_{x,y} - p \times \theta\| + \|p_x \times \theta - p \times \theta\| < 2\delta.
\]

\[\square\]

More succinctly, this lemma asserts that

\[
T_{n,\delta/2}(p \times \theta) \subseteq \bigcup_{x \in T_{n,\delta}(p)} (\{x\} \times T_{n,\delta}(\theta, x)) \subseteq T_{n,2\delta}(p \times \theta) \quad \forall \delta > 0. \quad (2)
\]

Using (2), we see that \(q^n\) is something like the uniform distribution on the subset of the Cartesian product

\[
T_{n,\delta}(q) \subseteq T_{n,\delta}(p_A) \times T_{n,\delta}(p_B),
\]

where moreover
1. most vertical slices $T_{n,\delta}(q) \cap (\{x\} \times B^n)$ have size roughly $2^{\mathcal{H}(Y|X)n}$ (where we restrict attention to $x$ which is itself typical for $p_A$), and similarly

2. most horizontal slices $T_{n,\delta}(q) \cap (A^n \times \{y\})$ have size roughly $2^{\mathcal{H}(X|Y)n}$.

This intuitive picture gives us a nice way to visualize the chain rule. There are some messy $\varepsilon$s and $\delta$s to keep track of, but the idea is this:

$$2^{\mathcal{H}(X,Y)n} \approx |T_{n,\delta}(q)| = \sum_{x \in T_{n,\delta}(p_A)} |T_{n,\delta}(q) \cap (\{x\} \times B^n)|$$

$$\approx |T_{n,\delta}(p_A)| \times \text{(typical size} |T_{n,\delta}(q) \cap (\{x\} \times B^n)|\text{)}$$

$$\approx 2^{\mathcal{H}(X)n} \cdot 2^{\mathcal{H}(Y|X)n},$$

where the approximation between the first and the second lines uses property 1 above.

3 Notes and remarks

Basic sources for this lecture: [CT91, Chapter 10, Exercise 16].


References


Tim Austin
Courant Institute of Mathematical Sciences, New York University, 251 Mercer St, New York NY 10012, USA

Email: tim@cims.nyu.edu
URL: cims.nyu.edu/~tim