Friedrich-Alexander-Univ
Erlangen-Nürnberg

FAU

Relative Entropy and the (Quantum) Method of Types

Mathematical Physics Seminar

Ian Koot

Friedrich-Alexander Universität Erlangen-Nürnberg, Department Mathematik June 1

2023

FAU

1. The Empirical Distribution and Types

2. The Method of Types

3. A Noncommutative Method of Types?

Probability distributions

We start with a finite space $X = \{1, 2, \ldots, d\}$ and a probability distribution $P : X \to \mathbb{R}$. We are interested in understanding P^n on X^n, where

$$P^n(\vec{x}) = P(x_1)P(x_2)\cdots P(x_n).$$

It is very natural to associate the following probability distribution to $\vec{x}$:

$$ED_n[\vec{x}](i) := \frac{1}{n}\sum_{k=1}^{n} \delta_{i,x_k} = \frac{|\{k : x_k = i\}|}{n}.$$

This is the **empirical distribution** associated to $\vec{x}$. Intuitively, we would expect to see an outcome whose empirical distribution is 'close to' the true probability distribution. But why exactly? And what does 'close to' mean?

Repeated measurement & ED

The crucial observation is the following:

$$P^n(\vec{x}) = P(x_1)P(x_2)\cdots P(x_n) = \prod_{i=1}^{d} P(i)^{nED_n[\vec{x}](i)} = \left(\prod_{i=1}^{d} P(i)^{ED_n[\vec{x}](i)}\right)^n$$

Note:

- The probability of $\vec{x}$ occurring depends only on its empirical distribution.
- If we have a sequence of possible outcomes $\vec{x}_n \in X^n$ with 'similar' empirical distributions, then the likelihood $P^n(\vec{x}_n)$ will decay exponentially, with its exponential factor given by a function of the empirical distribution.

Entropy! And Relative Entropy!

We note:

$$P^n(\vec{x}) = \prod_{i=1}^{d} P(i)^{nED_n[\vec{x}](i)} = \prod_{i=1}^{d} \left(\frac{P(i)}{ED_n[\vec{x}](i)} \right)^{nED_n[\vec{x}](i)} ED_n[\vec{x}](i)^{nED_n[\vec{x}](i)}$$

We can write this as follows:

Proposition

Let X be a finite set and $P : X \to \mathbb{R}$ a probability distribution. Then we have

$$P^n(\vec{x}) := \exp\left(-n(S(ED_n[\vec{x}], P) + S(ED_n[\vec{x}]))\right)$$

What happens if $P^n(\vec{x}) = 0$?

Types and Type classes

It is at this point very natural to group all possible outcomes by their empirical distributions.

Definition

An probability distribution $P : X \to \mathbb{R}$ is called an n-**type** if there is a $\vec{x} \in X^n$ so that $P = ED_n[\vec{x}]$. It is called a **type** if it is an n-type for some $n \in \mathbb{N}$.
The **type class** associated to the n-type P is the set

$$T_n(P) := \{\vec{x} \in X^n \mid P = ED_n[\vec{x}]\}$$

Note that for any type $P = ED_n[\vec{x}]$ and probability distribution Q we have

$$Q^n(T_n(P)) := \sum_{\vec{y} \in T_n(P)} Q^n(\vec{y}) = |T_n(P)| Q^n(\vec{x})$$

The Method of Types

- The element in X^n that has the highest likelihood of occurring with respect to P^n is obviously $\vec{x} = (i, i, i, \ldots, i)$, where $i \in X$ is an element for which P is maximal. However, unless P is a delta distribution, we never see such a result. Why?
- Answer: *Because the probability of that outcome occurring drops exponentially, and there is only 1 element with that empirical distribution!*
- So the more interesting question is: which *types* are very likely to occur? And how likely are they? The answer to this question (and applications of this answer) is called the **Method of Types**.

To summarize:

- The **empirical distribution** of an outcome determines its likelihood of occurring;
- The **entropy** and **relative entropy** naturally show up as decay rates for the probability;
- The **method of types** is the method of using the knowledge of which **types** are likely to occur to prove results.
- Specifically, we have

$$P^n(\vec{x}) = \prod_i^d P(i)^{nED_n[\vec{x}](i)} = e^{-n(S(ED_n[\vec{x}])+S(ED_n[\vec{x}],P))}$$

for

$$S(P) = -\sum_{i=1}^{d} P(i)\ln P(i),$$

$$S(P,Q) = \sum_{i=1}^{d} P(i)\ln P(i) - P(i)\ln Q(i).$$

Necessary relations

The method of types relies on the following observations:

- $Q^n(T_n(P)) = e^{-nS(P,Q)} P^n(T_n(P))$ for all probability distributions Q and n-types P;
- $P^n(T_n(Q)) \leq P^n(T_n(P))$ for all types n-types P and Q;
- $|ED_n[X^n]| \leq (n+1)^d$;
- $\bigcup_{Q \in ED_n[X^n]} T_n(Q) = X^n$;

Probability asymptotics

Proposition

Let P be an n-type and Q a probability distribution on X. Then

$$\frac{1}{(n+1)^d} e^{-nS(P,Q)} \leq Q^n(T_n(P)) \leq e^{-nS(P,Q)}.$$

Proof.

Note that

$$1 = \sum_{Q \in ED_n[X^n]} P^n(T_n(Q)) \leq (n+1)^d P(T_n(P)).$$

So

$$(n+1)^{-d} \leq P^n(T_n(P)) \leq 1$$

and so the result follows. □

So the decay rate of the probability of the type P occurring under the distribution Q is equal to $S(P, Q)$.

Set Size Asymptotics

Consequently, we see that the *size* of the type class of P grows with exponential rate $S(P)$:

Corollary

Let P be an n-type. Then

$$\frac{1}{(n+1)^d} e^{nS(P)} \leq |T_n(P)| \leq e^{nS(P)}.$$

Proof.

If $\vec{x} \in |T_n(P)|$, then $P^n(T_n(P)) = |T_n(P)|P(\vec{x}) = |T_n(P)|e^{-nS(P)}$. So the result follows from

$$(n+1)^{-d} \leq P^n(T_n(P)) \leq 1.$$

□

Note that this is for example also a really easy way to show that $S(P) \leq \ln d$ for types P, because $|T_n(P)| \leq |X^n| = d^n$ and

$$S(P) \leq \frac{d \ln(n+1)}{n} + \ln(d) \to \ln(d)$$

Sanov's Theorem

As an immediate application, we can prove **Sanov's theorem**:

Theorem

Let $E \subseteq \mathrm{Pr}(X)$, and $P \in \mathrm{Pr}(X)$. Then

$$P^n(E) := \sum_{Q \in E \cap ED_n[X^n]} P^n(T_n(Q)) \leq (n+1)^d \sup_{Q \in E} \left(e^{-nS(Q,P)} \right)$$

It is proven by simply bounding all the terms.
The Theorem shows us that if E is at least some $\varepsilon > 0$ removed from P in terms of relative entropy, then the probability of E occurring decays exponentially, and the exponent is given by the entropy distance. We therefore define for all $P \in \mathrm{Pr}(X)$ the 'entropy typical subsets'

$$A_{n,\varepsilon}(P) := \{\vec{x} \in X^n \mid S(ED_n[\vec{x}], P) < \varepsilon\}$$

Chernoff-Stein Lemma

In order to better compare to the noncommutative setting, where no analogue of the empirical distribution exists (yet), we prove a strengthening of Sanov's theorem:

Theorem (Chernoff-Stein Lemma)

Let $P, Q \in \mathsf{Pr}(X)$ and let $B_n \subseteq X^n$ be a sequence of subsets such that $\lim_{n\to\infty} P(B_n) = 1$. Then we have

$$\liminf_{n\to\infty} -\frac{1}{n} \ln Q(B_n) \geq S(P, Q)$$

Furthermore, there is a sequence that achieves this rate (independent of Q).

The idea is that typical sequences of sets for P (i.e. sequences $B_n \subseteq X^n$ such that $P^n(B_n) \to 1$) must have increasingly large intersections with the entropy typical subsets $A_{n,\varepsilon}(P)$, and those only decay as $e^{-nS(P,Q)}$.

To summarize:

- For P an n-type and Q a probability distribution, we now know that

$$\begin{aligned} \frac{1}{(n+1)^d} e^{-nS(P,Q)} &\leq Q^n(T_n(P)) \leq e^{-nS(P,Q)} \\ \frac{1}{(n+1)^d} e^{nS(P)} &\leq |T_n(P)| \leq e^{nS(P)} \end{aligned}$$

  Compare this also to the expression $Q_n(\vec{x}) = \exp(-n(S(P) + S(P,Q))$ (for $P = ED_n[\vec{x}]$); a part of the probability is compensated by the size of the type class, a part is not.
- By Sanov's theorem, we see that the sets

$$A_{n,\varepsilon}(P) := \{\vec{x} \in X^n \mid S(ED_n[\vec{x}], P) < \varepsilon\}$$

  with vectors of types we are likely to see under P become exponentially likely.
- The Chernoff-Stein Lemma tells us that for a P-typical sequence of sets $B_n \subset X^n$ (i.e. such that $\lim_{n\to\infty} P^n(B_n) = 1$), the Q-probability cannot fall off faster than exponentially with rate $S(P,Q)$.

Noncommutative Probability Theory

In our research we are interested in the noncommutative analogue of Probability Theory and Information Theory; this translation is made as follows.

- To each finite space (or measure space) X we can associate the von Neumann Algebra $L^\infty(X)$.
- To each probability distribution $P \in \mathrm{Pr}(X)$ (or probability measure) we can associate the state (= positive and unital linear functional) $\mathbb{E}_P$ given by the expectation value.
- Events A (= measurable subsets of X) correspond to the projections in $L^\infty(X)$ given by the characteristic functions χ_A. In particular, in the finite case, we can recover $P(i) = \mathbb{E}_P[\chi_{\{i\}}]$.

So we consider von Neumann algebras $\mathcal{A}$ with states $\omega = \mathrm{tr}[D_\omega \cdot] \in \mathcal{S}(\mathcal{A})$. Furthermore, we have the noncommutative analogues

$$S(\omega) = -\mathrm{tr}[D_\omega \ln D_\omega]$$
$$S(\omega, \psi) - \mathrm{tr}[D_\omega \ln D_\omega - D_\omega \ln D_\psi]$$

What doesn't work

There is no generally accepted noncommutative equivalent to the method of types. This starts with the lack of a definition of an empirical distribution and of a type class. The empirical distribution associates a state to a measurement outcome. This means that in this noncommutative case, we are looking for a map

$$ED_n : \mathcal{P}(\mathcal{A}^{\otimes n}) \supset \mathcal{P}_n \to \mathcal{S}(\mathcal{A}).$$

And a set of projections $T_n(\omega)$ for $\omega \in ED_n[\mathcal{P}_n]$.
We can use our 'dictionary' to translate properties that the classical concepts satisfy:

- $(ED_n[p])^{\otimes n}(p) = e^{-nS(ED_n[p])}$.
- $\omega^{\otimes n}(p) = (ED_n[p])^{\otimes n}(p)e^{-nS(ED_n[p],\omega)}$.
- $ED_n[p]$ maximizes the expression $\omega \mapsto \omega^{\otimes n}(p)$.
- ...

However, many of these properties cannot be realized, are ambiguous or contradict each other.

What does work

As mentioned before, we have a good definition for relative entropy. Even though the empirical distribution and type classes do not have an equivalent, the Chernoff-Stein Lemma *does* hold:

Theorem

Let $\mathcal{A}$ be finite dimensional, and $\phi, \psi \in \mathcal{S}(\mathcal{A})$. Then every sequence of projections $p_n \in \mathcal{P}(\mathcal{A}^{\otimes n})$ that satisfies $\lim_{n\to\infty} \psi^{\otimes n}(p_n) = 1$ also satisfies

$$\lim_{n\to\infty} -\frac{1}{n} \ln(\phi(p_n)) \leq S(\psi, \phi).$$

Furthermore, there is a sequence p_n that achieves this rate (this depends on ϕ and ψ).

This was proven in [**Bjelakovic2005**] based on results from [**Hiai1991**].

How not to prove the Noncommutative C.-S. Lemma

We cannot prove this by using empirical distributions and the method of types, because those don't exist in this setting. The following example highlights some difficulties:
Let $\mathcal{A} = B(\mathcal{H})$ (with $\dim(\mathcal{H}) < \infty$). Let $\omega_v(A) := \langle v, Av\rangle$. Then $S(\omega_w, \omega_v) = \infty$ if v and w are linearly independent.
Clearly, the projection $|w^{\otimes n}\rangle\langle w^{\otimes n}|$ is typical for ω_w: in a technical sense (since $(\omega_w)^{\otimes n}(|w^{\otimes n}\rangle\langle w^{\otimes n}|) = 1$ for all n), but also conceptually. However,

$$\omega_v^{\otimes n}(|w^{\otimes n}\rangle\langle w^{\otimes n}|) = |\langle v, w\rangle|^{2n}$$

does not reach the decay rate of ∞ if $\langle v, w\rangle \neq 0$. What *does* reach the decay rate is the projection onto

$$\mathcal{W}_n := \frac{1}{\sqrt{C_n}}\left(v^\perp \otimes w \otimes \ldots \otimes w + w \otimes v^\perp \otimes \ldots \otimes w + \ldots + w \otimes w \otimes \ldots \otimes v^\perp\right)$$

with the normalization $C_n := n(1 + (n-1)|\langle w, v^\perp\rangle|^2)$. It is typical:

$$\langle w^{\otimes n}, |\mathcal{W}_n\rangle\langle\mathcal{W}_n| w^{\otimes n}\rangle = \frac{n^2|\langle w, v^\perp\rangle|^2}{C_n} \to \begin{cases} 0 & \text{if } \langle w, v^\perp\rangle = 0 \\ 1 & \text{else} \end{cases}$$

and also $\langle v^{\otimes n}, |\mathcal{W}_n\rangle\langle\mathcal{W}_n| v^{\otimes n}\rangle = 0$.

How to prove the Noncommutative C.-S. Lemma

Instead, it turns out we can actually reduce to the commutative case. This is because there exists a commutative subalgebra $\mathcal{D}_l$ such that

$$S(\psi^{\otimes l}, \varphi^{\otimes l}) - S(\psi^{\otimes l}|_{\mathcal{D}_l}, \varphi^{\otimes l}|_{\mathcal{D}_l}) \leq |\mathcal{H}| \ln(l+1)$$

The algebra $\mathcal{D}_l$ is the one generated by the spectral projections of $D_{\varphi^{\otimes l}} = D_{\varphi}^{\otimes l}$; these are precisely the projections

$$T_l^{\varphi}(Q) := \bigvee_{\vec{x} \in T_l(Q)} (p^{\varphi})^{\otimes \vec{x}}$$

where $D_\phi = \sum_k \lambda_k p_k^{\varphi}$ und $(p^{\varphi})^{\otimes \vec{x}} = p_{x_1}^{\varphi} \otimes p_{x_2}^{\varphi} \otimes \cdots \otimes p_{x_l}^{\varphi}$.
One can then use the commutative results on $\mathcal{D}_l$, together with the fact that $S(\psi^{\otimes l}, \varphi^{\otimes l})$ to arrive at the desired result.

Summary

To summarize:

- Even though there exist notions of noncommutative entropy and relative entropy, there exist no generally accepted notion of the method of types in the noncommutative setting. One of the more fundamental problems seems to be that projections can be ‘typical’ in a lot of different ways.
- However, there does exist a noncommutative version of the Chernoff-Stein Lemma, which can be used to give an operational meaning to the relative entropy. Specifically: the relative entropy $S(\psi, \varphi)$ is the most extreme fall off in φ-probability that we can get for projections that are still typical for ψ.
- The crucial observation is that for large enough tensor powers, the noncommutative relative entropy can be approximated by a commutative relative entropy.

Outlook/Questions

Questions we are interested in, are for example:

- How should we interpret the reduction to the commutative subalgebra?
- Could one use the Chernoff-Stein characterization to give more intuitive proofs of the known properties of relative entropy?
- For infinite dimensional algebras, a notion of relative entropy exists. Does the Chernoff-Stein characterization still hold for that setting?
- If so, can we get a better understanding of for example mutual information and entanglement entropy in the QFT setting?
- Modular theory plays a vital role in the definition of relative entropy in infinite dimensions. Can we see this from such a Chernoff-Stein characterization? Can we maybe even learn more about modular theory from this perspective?

Friedrich-Alexander-Univ
Erlangen-Nürnberg

FAU

Relative Entropy and the (Quantum) Method of Types

Mathematical Physics Seminar

Ian Koot

Friedrich-Alexander Universität Erlangen-Nürnberg, Department Mathematik June 1

2023

Weakly Typical vs. Entropy Typical

For the proof of the Chernoff-Stein Lemma, we need some more control over what probability the individual elements in our typical sets have. We therefore define the **weakly typical sets**

$$A_{n,\varepsilon}^{\text{weak}}(P,Q) := \{\vec{x} \in X^n | P^n(\vec{x})e^{-n(S(P,Q)+\varepsilon)} \leq Q^n(\vec{x}) \leq P^n(\vec{x})e^{-n(S(P,Q)-\varepsilon)}.\}$$

Why this expression? We recall that if $P = ED_n[\vec{x}]$ we have the equality

$$Q^n(\vec{x}) = P^n(\vec{x})e^{-nS(P,Q)}.$$

So one might expect that outcomes that have empirical distributions that are *similar* to P, will satisfy $Q^n(\vec{x}) \approx P^n(\vec{x})e^{-nS(P,Q)}$.

Entropy Typical $\Rightarrow$ Weakly Typical

FAU

Proposition

Let $P, Q \in \mathsf{Pr}(X)$. For every $\varepsilon > 0$ there is a $\varepsilon' > 0$ such that $A_{n,\varepsilon}(P) \subseteq A^{\mathsf{weak}}_{n,\varepsilon'}(P, Q)$ for all $n \in \mathbb{N}$. Conversely, every small enough $\varepsilon' > 0$, there is a $\varepsilon > 0$ such that $A_{n,\varepsilon}(P) \subseteq A^{\mathsf{weak}}_{n,\varepsilon'}(P, Q)$ for every $n \in \mathbb{N}$.

The proof relies on the following observation:

$$P^n(\vec{x})e^{-n(S(P,Q)-\varepsilon)} < Q^n(\vec{x}) \Rightarrow e^{n\varepsilon} < \frac{Q^n(\vec{x})}{P^n(\vec{x})e^{-nS(P,Q)}} = \prod_{i=1}^{d} \frac{Q(i)^{nED_n[\vec{x}](i)}}{P(i)^{nED_n[\vec{x}](i)}} \cdot \frac{P(i)^{nP(i)}}{Q(i)^{nP(i)}}$$

$$= \prod_{i=1}^{d} \left(\frac{Q(i)}{P(i)}\right)^{n(ED_n[\vec{x}](i)-P(i))}$$

and the fact that $\frac{1}{2}(\|ED_n[\vec{x}] - P\|_1)^2 \leq S(ED_n[\vec{x}], P)$.

Proof of CS-Lemma

Let $\varepsilon > 0$. There is a $\delta > 0$ such that $A_{n,\delta}(P_1) \subseteq A^{\text{weak}}_{n,\varepsilon}(P_1, P_2)$. There is also an $n \in \mathbb{N}$ such that $P_1^n(B_n) > 1 - \varepsilon$ and $P_1^n(A_{n,\delta}(P_1)) > 1 - \varepsilon$.
Then $P_1^n(A^{\text{weak}}_{n,\varepsilon}(P_1, P_2)) > 1 - \varepsilon$, and therefore

$$P_1^n((B_n \cap A^{\text{weak}}_{n,\varepsilon}(P_1, P_2))^c) = P_1^n((B_n)^c \cup (A^{\text{weak}}_{n,\varepsilon}(P_1, P_2))^c) < 2\varepsilon.$$

But

$$\begin{aligned}(1 - 2\varepsilon)e^{-n(S(P_1,P_2)+\varepsilon)} &\leq P_1^n(B_n \cap A^{\text{weak}}_{n,\varepsilon}(P_1, P_2))e^{-n(S(P_1,P_2)+\varepsilon)} \\ &\leq P_2^n(B_n \cap A^{\text{weak}}_{n,\varepsilon}(P_1, P_2)) \leq P_2^n(B_n)\end{aligned}$$

Summary of Commutative Part

- The method of types tells us that

$$\frac{1}{(n+1)^d}\, e^{-nS(P,Q)} \leq Q^n(T_n(P)) \leq e^{-nS(P,Q)}$$
$$\frac{1}{(n+1)^d}\, e^{nS(P)} \leq |T_n(P)| \leq e^{nS(P)}$$

  i.e. the entropy of P reflects the size of the type class of P, and the relative entropy reflects the probability of the type class of P with respect to Q.
- This leads directly to Sanov's theorem, which says that for large enough n the only types that will likely occur are those with low relative entropy with respect to the probability distribution under consideration.
- We can then translate that fact to the Chernoff-Stein Lemma: any sequence of sets that becomes arbitrarily likely for P, cannot decay faster than $e^{-nS(P,Q)}$ in Q-probability (and there is a sequence of sets that reaches this rate).