Introduction to Smooth Entropies & the i.i.d. limit

(classical)

* <u>Random Variable</u>. X takes value $x \in \mathcal{X}$ with probability $P_X(x)$

* <u>Shannon Entropy</u>. Quantifies the "uncertainty" in X

$$H(X) = \langle -\log P_X(x) \rangle_{P_X} = -\sum_x P_X(x) \log P_X(x)$$

The Shannon entropy characterizes operational tasks:

→ data compression: $H(X)$ = # of bits needed on average to store the content of X.

→ communication (same)

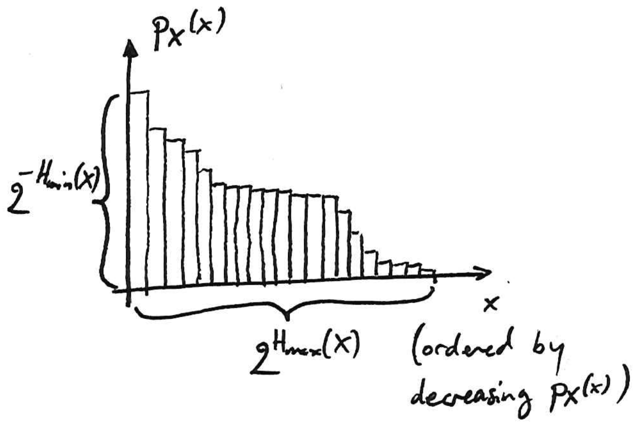
→ randomness extraction: $H(X)$ = # of uniformly random bits one can extract on average from X

* <u>"Single-Shot Regime"</u>: How should we characterize a single instance of these tasks, and not just "on average"?

→ introduce H_{min} & H_{max}:

$$H_{min}(X) = -\log P_X^{max}$$

$$H_{max}(X) = \log |\operatorname{supp} P_X|$$



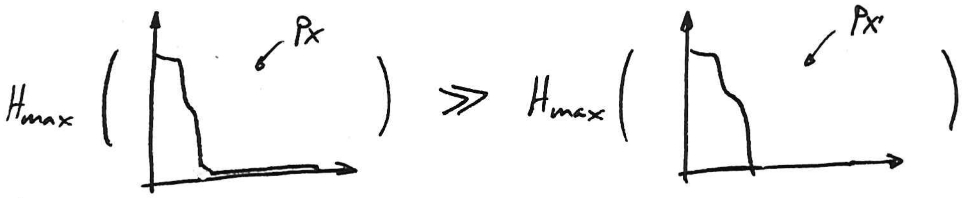
These entropies characterize worst case situations.

→ data compression. To store the content of X with certainty, you need $H_{max}(X)$ bits.
(possibility to encode each symbol of $\mathcal{X}$ which may appear)

→ randomness extraction. $H_{min}(X) = \#$ of uniformly random bits one can extract on a single instance from X

* Problem. These entropies can be very discontinuous.

e.g.:



Why is this a problem?

- H_{max} should correspond to an operational task, that is, something we can observe.
- P_X can't be distinguished from $P_{X'}$ (except with probability ε)

We don't want operational quantities to depend on unobservable features of P_X.

Data compression example: $H_{max}(X)$ (for this P_X) tells you to "reserve space" for symbols which you'll effectively never see!

* Trace Distance. (or Variational Distance)

$$\delta(P_X, Q_X) = \frac{1}{2}\sum_x |P_X(x) - Q_X(x)|$$

characterizes the distinguishability of P_X & Q_X.

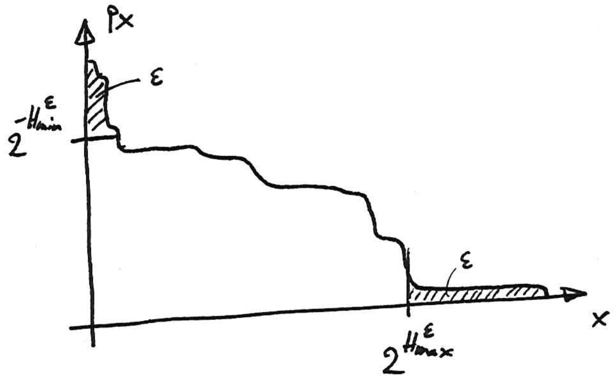
$\rightarrow \delta(P_X, Q_X) \leq \varepsilon \Rightarrow P_X$ can't be distinguished from Q_X, except with probability ε.

* Smooth Entropies.

Idea: "smooth" H_{min} & H_{max} so that they are no longer sensitive to unobservable features of P_X.

$$H^{\varepsilon}_{max}(X)_{P_X} = \min_{Q_X \overset{\varepsilon}{\approx} P_X} H_{max}(X)_{Q_X}$$

$$H^{\varepsilon}_{min}(X)_{P_X} = \max_{Q_X \approx_{\varepsilon} P_X} H_{min}(X)_{Q_X}$$



$\rightarrow$ potentially "save" lots of resources by allowing a small probability of failure. For example, by allowing a failure probability ε in data compression, we can use only $H^{\varepsilon}_{max}(X)$ bits.

* I.I.D. Limit & Typicality. ("Independent & Identically Distributed")

Example: Toss fair coin n times $\rightarrow$ sequence (x^n), $x^i = 0, 1$.

$\rightarrow$ Each sequence x^n has (here) same probability $= 2^{-n}$

$\rightarrow$ but some events are much more probable:

$$\text{Prob}[\text{all } 0\text{'s}] = \text{Prob}[\text{all } 1\text{'s}] = 2^{-n} \quad \text{but}$$

$$\text{Prob}[\#\{i : x^i = 0\} \approx \tfrac{1}{2} n] = 2^{-n} \times (\text{many such sequences})$$

$$\text{actually:} \quad \xrightarrow[n \to \infty]{} 1$$

Most observed sequences are <u>typical</u>, i.e. belong to the <u>typical set</u>.

Typical Set. [there are different definitions. Here: "weak typicality" see Mark Wilde, "Quantum Information Theory", arXiv: 1106.1445]

$$T_\delta^{X^n} = \left\{ (x^n) : \left| \frac{1}{n} \sum h_{p_X}(x^i) - H(X) \right| < \delta \right\}$$

Properties of the typical set:

→ $\mathrm{Prob}\left[(x^n) \in T_\delta^{X^n} \right] \xrightarrow{n \to \infty} 1$ unit probability

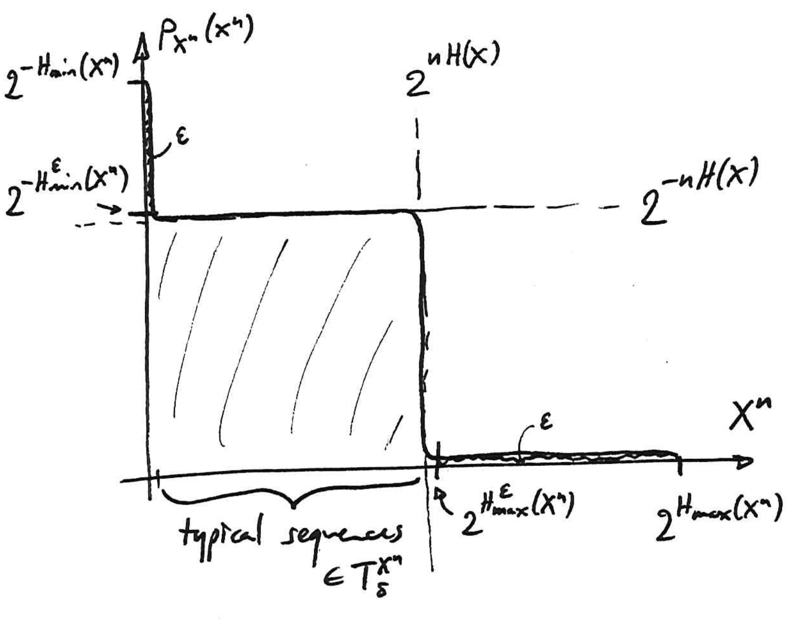
($\Leftarrow$ law of large numbers)

$\Rightarrow$ Protocols or implementations of information-theoretical tasks only have to worry about typical sequences (and can ignore non-typical sequences because they'll never be observed)

→ $|T_\delta^{X^n}| \approx 2^{nH(X)}$ size (given by Shannon entropy)

→ $P_{X^n}(x^n) \approx 2^{-nH(X)}$ for all $x^n \in T_\delta^{X^n}$ equipartition (all typical sequences have $\approx$ same probability)

Picture in the i.i.d. regime.



→ Asymptotic Equipartition Property:

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} H_{max}^{\varepsilon}(X^n) = H(X)$$

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} H_{min}^{\varepsilon}(X^n) = H(X)$$

→ Protocols or implementations of information-theoretic tasks which are characterized by $H_{min}^{\varepsilon}(X)$ or $H_{max}^{\varepsilon}(X)$ in the "single-shot" regime are characterized by $H(X)$, the Shannon entropy, on average in the i.i.d. limit.

* Remark.

The Asymptotic Equipartition Property also holds for the conditional H_{min}/H_{max} and for the quantum (conditional) H_{min}/H_{max}.