Information Theory II

Signal and Information Processing Laboratory
Institut für Signal- und Informationsverarbeitung

Spring Semester 2017

Prof. Dr. A. Lapidoth / Prof. Dr. S. Moser

Lecture Notes
Guessing and Rényi Entropy

http://www.isi.ee.ethz.ch/teaching/courses/it2.html

---

1 Guessing

How many guesses of the form

"X is equal to fill-in-the-blank"

does it take until one guesses correctly the outcome of a chance variable X that is drawn from a finite set $\mathcal{X}$ according to some PMF P_X?

In the best case it takes one guess. And, unless we repeat guesses (which would be pointless), in the worst case the number of guesses is equal to the cardinality $|\mathcal{X}|$ of the set $\mathcal{X}$: we might have to guess every element of $\mathcal{X}$ other than X before guessing X.

But what about the average case, i.e., the expected number of guesses? The answer would, of course, depend on our guessing order. Here we will derive the guessing order that minimizes this expectation and analyze the expected number of guesses it requires.

More generally, rather than considering the expected number of guesses, we shall study the ρ-th moment of this number for any fixed positive $\rho > 0$. When ρ equals 1, this moment is the expectation.

As we shall shortly see, finding an optimal strategy is straightforward, but its analysis requires some work. But first we introduce some notation. We say that $\mathsf{G}(\cdot)$ is a guessing function for X if G is a bijective function from $\mathcal{X}$ to the set $\{1, \ldots, |\mathcal{X}|\}$

$$\mathsf{G}\colon \mathcal{X} \to \{1, \ldots, |\mathcal{X}|\}. \tag{1}$$

By "guessing according to $\mathsf{G}(\cdot)$" we mean that we guess "X is equal to x" in guess number $\mathsf{G}(x)$. Our first guess is thus the element of $\mathcal{X}$ that is mapped by $\mathsf{G}(\cdot)$ to 1; the second is the element of $\mathcal{X}$ that is mapped by $\mathsf{G}(\cdot)$ to 2; etc. If we guess according to $\mathsf{G}(\cdot)$, and if X is equal to x, then the number of guesses we make until we guess correctly is $\mathsf{G}(x)$. The expected number of guesses to guess X is thus

$$\mathsf{E}[\mathsf{G}(X)],$$

i.e.,

$$\sum_{x \in \mathcal{X}} P_X(x)\, \mathsf{G}(x).$$

More generally, for any

$$\rho > 0$$

the ρ-th moment of the number of guesses is

$$\mathsf{E}[\mathsf{G}(X)^\rho].$$

We say that $\mathsf{G}(\cdot)$ is optimal (for P_X and ρ) if—among all guessing functions—guessing according to $\mathsf{G}(\cdot)$ minimizes the ρ-th moment of the number of guesses needed to guess X.

Example 1 (Guessing a Uniform Chance Variable)**.** Let X be uniformly distributed over a nonempty finite set $\mathcal{X}$, and let $\rho > 0$ be fixed. Then all guessing functions are optimal,

$$\frac{1}{\rho+1}\,|\mathcal{X}|^\rho \le \mathsf{E}[\mathsf{G}^*(X)^\rho] \le \frac{|\mathcal{X}|+1}{|\mathcal{X}|}\cdot\frac{(|\mathcal{X}|+1)^\rho}{\rho+1}, \tag{2}$$

and

$$\frac{1}{\rho+1}\,|\mathcal{X}|^\rho \le \mathsf{E}[\mathsf{G}^*(X)^\rho] \le |\mathcal{X}|^\rho. \tag{3}$$

Consequently,

$$\log|\mathcal{X}|^\rho - \epsilon_{\mathrm{L}}(\rho) \le \log \mathsf{E}[\mathsf{G}^*(X)^\rho] \le \log|\mathcal{X}|^\rho \tag{4}$$

where $\epsilon_{\mathrm{L}}(\rho)$ is nonnegative and does not depend on $|\mathcal{X}|$.
Proof: To establish the upper bound in (2), we proceed as follows:

$$\begin{aligned}
\mathsf{E}[\mathsf{G}^*(X)^\rho] &= \sum_{x\in\mathcal{X}} P_X(x)\,\mathsf{G}^*(x)^\rho && (5)\\
&= \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \mathsf{G}^*(x)^\rho && (6)\\
&= \frac{1}{|\mathcal{X}|}\sum_{m=1}^{|\mathcal{X}|} m^\rho && (7)\\
&= \frac{1}{|\mathcal{X}|}\sum_{m=1}^{|\mathcal{X}|}\int_m^{m+1} m^\rho\,\mathrm{d}\zeta && (8)\\
&\le \frac{1}{|\mathcal{X}|}\sum_{m=1}^{|\mathcal{X}|}\int_m^{m+1} \zeta^\rho\,\mathrm{d}\zeta && (9)\\
&= \frac{1}{|\mathcal{X}|}\int_1^{|\mathcal{X}|+1} \zeta^\rho\,\mathrm{d}\zeta && (10)\\
&= \frac{1}{|\mathcal{X}|}\,\frac{\zeta^{\rho+1}}{\rho+1}\Big|_1^{|\mathcal{X}|+1} && (11)\\
&= \frac{1}{|\mathcal{X}|}\,\frac{(|\mathcal{X}|+1)^{\rho+1}-1}{\rho+1} && (12)\\
&\le \frac{|\mathcal{X}|+1}{|\mathcal{X}|}\cdot\frac{(|\mathcal{X}|+1)^\rho}{\rho+1}, && (13)
\end{aligned}$$

where in (8) we used the fact that $\mathsf{G}^*(\cdot)$ is a bijection from $\mathcal{X}$ onto the set $\{1,\ldots,|\mathcal{X}|\}$, and where in (9) we have used the fact that $\zeta \mapsto \zeta^\rho$ is monotonically increasing on the interval $[1,\infty)$ irrespective of the value of $\rho > 0$.

The simpler upper bound in (3) follows trivially because $|\mathcal{X}|$ is the worst-case number of necessary guesses.

As to a lower bound, we can use the monotonicity to infer that

$$m^\rho \geq \int_{m-1}^{m} \zeta^\rho \, \mathrm{d}\zeta \tag{14}$$

and hence

$$\begin{aligned}
\mathsf{E}[\mathsf{G}^*(X)^\rho] &= \frac{1}{|\mathcal{X}|} \sum_{m=1}^{|\mathcal{X}|} m^\rho && (15)\\
&= \frac{1}{|\mathcal{X}|} \left(1 + \sum_{m=2}^{|\mathcal{X}|} m^\rho \right) && (16)\\
&\geq \frac{1}{|\mathcal{X}|} \left(1 + \sum_{m=2}^{|\mathcal{X}|} \int_{m-1}^{m} \zeta^\rho \, \mathrm{d}\zeta \right) && (17)\\
&= \frac{1}{|\mathcal{X}|} \left(1 + \int_{1}^{|\mathcal{X}|} \zeta^\rho \, \mathrm{d}\zeta \right) && (18)\\
&= \frac{1}{|\mathcal{X}|} \left(1 + \frac{\zeta^{\rho+1}}{\rho+1} \Big|_1^{|\mathcal{X}|} \right) && (19)\\
&= \frac{1}{|\mathcal{X}|} \left(1 + \frac{|\mathcal{X}|^{\rho+1} - 1}{\rho+1} \right) && (20)\\
&\geq \frac{1}{|\mathcal{X}|} \, \frac{|\mathcal{X}|^{\rho+1}}{\rho+1} && (21)\\
&= \frac{1}{\rho+1} \, |\mathcal{X}|^\rho. && (22)
\end{aligned}$$

□♢

The following proposition establishes that $\mathsf{G}(\cdot)$ is optimal if the guessing order it induces is according to decreasing order of probabilities.

Proposition 2 (Optimal Guessing Functions). *Given a PMF P_X and some $\rho > 0$, a guessing function $\mathsf{G}(\cdot)$ is optimal if, and only if,*

$$\Big(P_X(x) > P_X(x') \Big) \implies \Big(\mathsf{G}(x) < \mathsf{G}(x') \Big) \tag{23}$$

for all $x, x' \in \mathcal{X}$.

Before proving this proposition, we study the implications of swapping the guessing orders of two elements of $\mathcal{X}$.

Lemma 3. *Let X be drawn according to P_X from a finite set $\mathcal{X}$, and let $\mathsf{G}(\cdot)$ be a guessing function for X. Let x, x' be distinct elements of $\mathcal{X}$. Let $\tilde{\mathsf{G}}$ be identical to G except that it swaps the values of $\mathsf{G}(x)$ and $\mathsf{G}(x')$, so*

$$\tilde{\mathsf{G}}(\xi) = \begin{cases} \mathsf{G}(\xi) & \text{if } \xi \in \mathcal{X} \setminus \{x, x'\}, \\ \mathsf{G}(x) & \text{if } \xi = x', \\ \mathsf{G}(x') & \text{if } \xi = x, \end{cases} \qquad \xi \in \mathcal{X}. \tag{24}$$

Then for every $\rho > 0$,

$$\mathsf{E}\big[\tilde{\mathsf{G}}(X)^\rho\big] = \mathsf{E}[\mathsf{G}(X)^\rho] + \big(P_X(x) - P_X(x')\big)\big(\mathsf{G}(x')^\rho - \mathsf{G}(x)^\rho\big). \tag{25}$$

Proof: Clearly $\tilde{\mathsf{G}}(\cdot)$ is also a guessing function and

$$\begin{aligned}
\mathsf{E}\big[\tilde{\mathsf{G}}(X)^\rho\big] &= \sum_{\xi\in\mathcal{X}} P_X(\xi)\,\tilde{\mathsf{G}}(\xi)^\rho && (26)\\
&= \sum_{\xi\in\mathcal{X}\setminus\{x,x'\}} P_X(\xi)\,\tilde{\mathsf{G}}(\xi)^\rho + P_X(x)\,\tilde{\mathsf{G}}(x)^\rho + P_X(x')\,\tilde{\mathsf{G}}(x')^\rho && (27)\\
&= \sum_{\xi\in\mathcal{X}\setminus\{x,x'\}} P_X(\xi)\,\mathsf{G}(\xi)^\rho + P_X(x)\,\mathsf{G}(x')^\rho + P_X(x')\,\mathsf{G}(x)^\rho && (28)\\
&= \sum_{\xi\in\mathcal{X}\setminus\{x,x'\}} P_X(\xi)\,\mathsf{G}(\xi)^\rho + P_X(x)\,\mathsf{G}(x)^\rho + P_X(x')\,\mathsf{G}(x')^\rho\\
&\quad + P_X(x)\,\mathsf{G}(x')^\rho + P_X(x')\,\mathsf{G}(x)^\rho - P_X(x)\,\mathsf{G}(x)^\rho - P_X(x')\,\mathsf{G}(x')^\rho && (29)\\
&= \mathsf{E}[\mathsf{G}(X)^\rho] + \big(P_X(x) - P_X(x')\big)\big(\mathsf{G}(x')^\rho - \mathsf{G}(x)^\rho\big), && (30)
\end{aligned}$$

where the third equality follows from (24), and the fourth by adding and subtracting $P_X(x)\mathsf{G}(x)^\rho + P_X(x')\,\mathsf{G}(x')^\rho$. □

Proof of Proposition 2: With the aid of Lemma 3 the proof of Proposition 2 is straightforward. If $\mathsf{G}(\cdot)$ is optimal, then (23) must hold because in this case $\mathsf{E}\big[\tilde{\mathsf{G}}(X)^\rho\big]$ of the lemma cannot be smaller than $\mathsf{E}[\mathsf{G}(X)^\rho]$.

As to sufficiency, let $\mathsf{G}(\cdot)$ satisfy (23) and let G^* be optimal. As such, it must satisfy (23) when substituting G^* for G. It then follows that for every ξ in the range of $P_X(\cdot)$, the set $\{x\in\mathcal{X} : P_X(x)=\xi\}$ is mapped by G and G^* to the same set. Consequently, we can go from G to G^* by a sequence of swaps of the kind addressed in Lemma 3 with $P_X(x)$ and $P_X(x')$ being equal. For such swaps $\mathsf{E}[\mathsf{G}(X)^\rho]$ and $\mathsf{E}\big[\tilde{\mathsf{G}}(X)^\rho\big]$ are identical, and hence $\mathsf{E}[\mathsf{G}(X)^\rho]$ must equal $\mathsf{E}\big[\mathsf{G}^*(X)^\rho\big]$, and $\mathsf{G}(\cdot)$ must be optimal. □

Corollary 4. *If $\mathsf{G}^*(\cdot)$ is an optimal guessing function for X, then for every $x\in\mathcal{X}$,*

$$\big|\{\xi\in\mathcal{X} : P_X(\xi) > P_X(x)\}\big| \le \mathsf{G}^*(x) \le \big|\{\xi\in\mathcal{X} : P_X(\xi)\ge P_X(x)\}\big|. \tag{31}$$

Moreover,

$$\big|\{\xi\in\mathcal{X} : P_X(\xi)\ge P_X(x)\}\big| \le \frac{1}{P_X(x)}, \tag{32}$$

so

$$\mathsf{G}^*(x) \le \frac{1}{P_X(x)}. \tag{33}$$

Proof: The LHS of (31) follows by noting that, since G^* is optimal, every outcome ξ whose probability exceeds that of x will be guessed before x.

The RHS follows because if ξ is guessed before x, its probability cannot be smaller than that of x.

To prove (32) we note that each element of the set on the LHS of (32) has a probability of at least $P_X(x)$, so there cannot be more than $1/P_X(x)$ of them because otherwise their total probability would exceed 1.

Finally, (33) follows from (31) (the inequality on the right) and (32). □

2 Guessing with Side Information

We next consider guessing in the presence of some side information (SI). We envision that the pair (X,Y) is drawn from the finite set $\mathcal{X}\times\mathcal{Y}$ according to some PMF $P_{X,Y}$, and that

we must guess X after observing Y. For every possible outcome y of Y, we need to propose a guessing function $\mathsf{G}(\cdot|y)$. We seek to minimize

$$\mathsf{E}[\mathsf{G}(X|Y)^\rho],$$

which can also be expressed as

$$\sum_{y\in\mathcal{Y}} P_Y(y)\,\mathsf{E}[\mathsf{G}(X|Y=y)^\rho].$$

The latter form shows that it is optimal to choose $\mathsf{G}(\cdot|y)$ to minimize

$$\mathsf{E}[\mathsf{G}(X|Y=y)^\rho].$$

This optimization problem is similar to the one we encountered when guessing X without side information: the only difference is that the probability we associate with the outcome x is now not $P_X(x)$ but $P_{X|Y}(x|y)$. We thus conclude:

Proposition 5. *To minimize*

$$\mathsf{E}[\mathsf{G}(X|Y)^\rho]$$

it is optimal to guess X after observing that $Y=y$ according to decreasing order of posterior probabilities $P_{X|Y}(x|y)$.

How helpful is Y for guessing X? Clearly observing Y cannot hurt, because we can always ignore the observation and guess according to decreasing order of the prior P_X. Hence, if $\mathsf{G}^*(x|y)$ and $\mathsf{G}^*(x)$ are optimal for guessing X with and without the SI respectively, then

$$\mathsf{E}[\mathsf{G}^*(X|Y)^\rho] \le \mathsf{E}[\mathsf{G}^*(X)^\rho] \tag{34}$$

(with equality whenever X and Y are independent).[1]

While generally useful, the benefits of SI are limited by its support set:

Proposition 6. *If the pair (X,Y) takes values in the finite set $\mathcal{X}\times\mathcal{Y}$ according to the joint PMF $P_{X,Y}$, then for every $\rho>0$*

$$\mathsf{E}[\mathsf{G}^*(X)^\rho] \le |\mathcal{Y}|^\rho\,\mathsf{E}[\mathsf{G}^*(X|Y)^\rho]. \tag{35}$$

Proof: To prove the result we will construct a guessing function $\mathsf{G}(\cdot)$ for X for which

$$\mathsf{E}[\mathsf{G}(X)^\rho] \le |\mathcal{Y}|^\rho\,\mathsf{E}[\mathsf{G}^*(X|Y)^\rho]. \tag{36}$$

The idea is simple, but the notation cumbersome. It is therefore instructive to consider an example. Suppose $\mathcal{X}=\{a,b,c,d,e\}$ and $\mathcal{Y}=\{\uparrow,\downarrow\}$ and that conditionally on $Y=\uparrow$ the probabilities of the different outcomes of X are alphabetically increasing

$$0<P(a|\uparrow)<P(b|\uparrow)<\cdots<P(e|\uparrow), \tag{37}$$

whereas conditionally on $Y=\downarrow$ they are alphabetically decreasing

$$P(a|\downarrow)>P(b|\downarrow)>\cdots>P(e|\downarrow)>0. \tag{38}$$

Having observed $Y=\uparrow$, the optimal guessing order is e,d,c,b,a, whereas when $Y=\downarrow$ it is a,b,c,d,e.

[1] Equality may actually hold for $\rho>0$ even when X and Y are not independent.

We now construct a guessing rule for X by interlacing the two orders. Beginning (arbitrarily) with $\uparrow$, we take the first guess determined by $\mathsf{G}^*(\cdot|Y=\uparrow)$ — namely e — followed by the first guess determined by $\mathsf{G}^*(\cdot|Y=\downarrow)$ — namely a — followed by the second guess determined by $\mathsf{G}^*(\cdot|Y=\uparrow)$ followed by the second guess determined by $\mathsf{G}^*(\cdot|Y=\downarrow)$, etc., until we have exhausted both lists and obtained a list with $|\mathcal{Y}|\cdot|\mathcal{X}|$ elements

$$e, a, d, b, c, c, b, d, a, e. \tag{39}$$

This is the order we are after, but since we do not repeat guesses, we cross out all the guesses that have already appeared to obtain the guessing order

$$e, a, d, b, c. \tag{40}$$

Denoting the guessing function induced by this order $\mathsf{G}(\cdot)$, we have

$$\mathsf{G}(x) \le 2\,\mathsf{G}(x|\uparrow) \tag{41}$$

because already in the order in (39) (which contains repetitions) this holds since x appears in this list in location $2\mathsf{G}(x|\uparrow)-1$ (if our arbitrary order of interlacing was indeed $\uparrow$ followed by $\downarrow$) or in location $2\,\mathsf{G}(x|\uparrow)$ (if the interlacing order was instead $\downarrow$ followed by $\uparrow$). Either way, (41) holds. (In the list (39) x occurs also in another location (which is determined by $\mathsf{G}(x|\downarrow)$ and the interlacing order).

Likewise

$$\mathsf{G}(x) \le 2\,\mathsf{G}(x|\downarrow) \tag{42}$$

and hence

$$\mathsf{G}(x) \le 2\min\bigl\{\mathsf{G}(x|\uparrow), \mathsf{G}(x|\downarrow)\bigr\}. \tag{43}$$

To conclude the proof we need to generalize this interlacing construction. To do so we denote the elements of $\mathcal{Y}$ by $\{y_1, y_2, \ldots, y_{|\mathcal{Y}|}\}$ and the elements of $\mathcal{X}$ by $\{x_1, \ldots, x_{|\mathcal{X}|}\}$. We shall interlace the lists starting with the list determined by $\mathsf{G}^*(\cdot|y_1)$ and ending with the list determined by $\mathsf{G}^*(\cdot|y_{|\mathcal{Y}|})$. To do so we shall first form a list of length $|\mathcal{X}|\cdot|\mathcal{Y}|$ analogous to (39), where each element of $\mathcal{X}$ appears $|\mathcal{Y}|$ times. The element x appears in the list in locations

$$|\mathcal{Y}|\cdot\mathsf{G}^*(x|y_1)-|\mathcal{Y}|+1,\ |\mathcal{Y}|\cdot\mathsf{G}^*(x|y_2)-|\mathcal{Y}|+2,\ \ldots,\ |\mathcal{Y}|\cdot\mathsf{G}^*(x|y_{|\mathcal{Y}|}).$$

Each element $x\in\mathcal{X}$ appears in this list no later than

$$|\mathcal{Y}|\cdot\min\bigl\{\mathsf{G}^*(x|y_1), \ldots, \mathsf{G}^*(x|y_{|\mathcal{Y}|})\bigr\}. \tag{44}$$

If we now erase from the list each entry that has appeared earlier in the list, we obtain a new list of length $|\mathcal{X}|$ in which each element x of $\mathcal{X}$ appears exactly once and in a location that is no later than (44). Guessing the elements of $\mathcal{X}$ according to the order in which they occur in the list yields a guessing function $\mathsf{G}(\cdot)$ satisfying (35) because

$$\mathsf{G}(()x) \le |\mathcal{Y}|\cdot\min\bigl\{\mathsf{G}(x|y_1), \ldots, \mathsf{G}(x|y_{|\mathcal{Y}|})\bigr\}, \tag{45}$$

so

$$\begin{aligned}
\mathsf{E}[\mathsf{G}(X)^\rho] &= \sum_{y\in\mathcal{Y}} P_Y(y)\ \mathsf{E}[\mathsf{G}(X)^\rho\,|\,Y=y] && (46)\\
&\le \sum_{y\in\mathcal{Y}} P_Y(y)\,|\mathcal{Y}|^\rho\,\mathsf{E}[\min\bigl\{\mathsf{G}(X|y_1)^\rho, \ldots, \mathsf{G}(X|y_{|\mathcal{Y}|})^\rho\,|\,Y=y] && (47)\\
&\le \sum_{y\in\mathcal{Y}} P_Y(y)\,|\mathcal{Y}|^\rho \mathsf{G}(X|y)^\rho && (48)\\
&= |\mathcal{Y}|^\rho\,\mathsf{E}[\mathsf{G}^*(X|Y)^\rho]. && (49)
\end{aligned}$$

□

3 Rényi Entropy

The order-α Rényi entropy $\mathsf{H}_\alpha(X)$ of a chance variable X of PMF P_X is defined for every positive α other than 1 as

$$\mathsf{H}_\alpha(X) \triangleq \frac{\alpha}{1-\alpha}\log\left(\left(\sum_{x\in\mathcal{X}}\big(P_X(x)\big)^\alpha\right)^{1/\alpha}\right), \quad \big(\alpha>0,\ \alpha\neq 1\big). \tag{50}$$

The order-α conditional Rényi entropy $\mathsf{H}_\alpha(X|Y)$ of X given Y is defined for every such α as

$$\mathsf{H}_\alpha(X|Y) \triangleq \frac{\alpha}{1-\alpha}\log\left(\sum_{y\in\mathcal{Y}}\left(\sum_{x\in\mathcal{X}}\big(P_{X,Y}(x,y)\big)^\alpha\right)^{1/\alpha}\right), \quad \big(\alpha>0,\ \alpha\neq 1\big). \tag{51}$$

Note 7. The term

$$\left(\sum_{x\in\mathcal{X}}\big(P_X(x)\big)^\alpha\right)^{1/\alpha}$$

is sometimes denoted $\|P_X\|_\alpha$ because it is reminiscent of the L_α "norm" of the mapping[2]

$$P_X\colon \mathcal{X}\to[0,1].$$

Consequently, the Rényi entropy is sometimes written as

$$\mathsf{H}_\alpha(X) = \frac{\alpha}{1-\alpha}\log\|P_X\|_\alpha \tag{52}$$

We will need the following variational characterization of $\mathsf{H}_\alpha(X)$:

Proposition 8. *Let X be a chance variable of PMF $P\in\mathcal{P}(\mathcal{X})$. Then*

$$\max_{Q\in\mathcal{P}(\mathcal{X})}\big\{\rho\,\mathsf{H}(Q)-\mathsf{D}(Q\|P)\big\} = \log\left(\left(\sum_{x\in\mathcal{X}}\big(P(x)\big)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right), \quad \rho>0, \tag{53}$$

$$= \rho\,\mathsf{H}_{\frac{1}{1+\rho}}(X), \quad \rho>0. \tag{54}$$

Proof:

$$\rho\,\mathsf{H}(Q)-\mathsf{D}(Q\|P) = \rho\sum_{x\in\mathcal{X}}Q(x)\log\frac{1}{Q(x)}-\sum_{x\in\mathcal{X}}Q(x)\log\frac{Q(x)}{P(x)} \tag{55}$$

$$= \sum_{x\in\mathcal{X}}Q(x)\log\left(\frac{1}{Q(x)^\rho}\cdot\frac{P(x)}{Q(x)}\right) \tag{56}$$

$$= \sum_{x\in\mathcal{X}}Q(x)\log\frac{P(x)}{Q(x)^{1+\rho}}. \tag{57}$$

Let $R(x)$ be a PMF such that $P(X)$ is proportional to $R(x)^{1+\rho}$, so

$$P(x) = \frac{1}{\gamma}\,R(x)^{1+\rho} \tag{58}$$

where

$$\gamma = \sum_{x\in\mathcal{X}}R(x)^{1+\rho}. \tag{59}$$

[2]We put norm in quotes because this is a norm only when α is greater-equal 1.

More explicitly,

$$R(x) = \gamma^{\frac{1}{1+\rho}}\, P(x)^{\frac{1}{1+\rho}} \tag{60}$$

Then, from (57),

$$\begin{aligned}
\rho\, \mathsf{H}(Q) - \mathsf{D}(Q\,\|\,P) &= \sum_{x\in\mathcal{X}} Q(x) \log \frac{P(x)}{Q(x)^{1+\rho}} && (61)\\
&= \sum_{x\in\mathcal{X}} Q(x) \log\left(\frac{1}{\gamma}\frac{R(x)^{1+\rho}}{Q(x)^{1+\rho}}\right) && (62)\\
&= \log\frac{1}{\gamma} + (1+\rho)\sum_{x\in\mathcal{X}} Q(x)\log\frac{R(x)}{Q(x)} && (63)\\
&= \log\gamma^{-1} - (1+\rho)\,\mathsf{D}(Q\,\|\,R) && (64)\\
&\le \log\gamma^{-1} && (65)
\end{aligned}$$

because $1+\rho$ is nonnegative and so is $\mathsf{D}(Q\,\|\,R)$.

We next use (58) and (59) to express γ in terms of $P(\cdot)$: From (58) it follows that

$$\frac{R(x)}{\gamma^{\frac{1}{1+\rho}}} = P(x)^{\frac{1}{1+\rho}} \tag{66}$$

and hence, upon summing over x,

$$\frac{1}{\gamma^{\frac{1}{1+\rho}}} = \sum_{x\in\mathcal{X}} P(x)^{\frac{1}{1+\rho}} \tag{67}$$

or

$$\frac{1}{\gamma} = \left(\sum_{x\in\mathcal{X}} P(x)^{\frac{1}{1+\rho}}\right)^{1+\rho}. \tag{68}$$

Combining (65) and (68) we obtain that

$$\rho\, \mathsf{H}(Q) - \mathsf{D}(Q\,\|\,P) \le \log\left(\left(\sum_{x\in\mathcal{X}} P(x)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right), \tag{69}$$

with equality if, and only if, we have equality in (65). Since $1+\rho$ is positive, (65) holds with equality if, and only if, $\mathsf{D}(Q\,\|\,R)$ is zero, i.e., if, like $R(x)$, $Q(x)$ is proportional to $P(x)^{1/(1+\rho)}$, i.e.,

$$Q(x) = \frac{P(x)^{\frac{1}{1+\rho}}}{\sum_{x'\in\mathcal{X}} P(x')^{\frac{1}{1+\rho}}}. \tag{70}$$

□

We next turn to sequences and product distributions. Let P be a PMF on $\mathcal{X}$, and let P^n be its n-fold product distribution on $\mathcal{X}^n$,

$$P^n(\mathbf{x}) = \prod_{i=1}^{n} P(x_i), \quad \mathbf{x}\in\mathcal{X}^n. \tag{71}$$

Let the random n-vector $\mathbf{X}$ be drawn according to P^n, with its components $X_1,\ldots,X_n$ thus IID $\sim P$. How many guesses would we need to guess $\mathbf{X}$? Can we say something about its large-n asymptotics?

The number of possible outcomes of $\mathbf{X}$ is $|\mathcal{X}|^n$, i.e., $2^{n\log_2|\mathcal{X}|}$, which is exponential in n. We therefore expect the number of required guesses to also be exponential, but, as we shall see, the exponent will be typically smaller than $\log|\mathcal{X}|$ (unless P is the uniform distribution on $\mathcal{X}$). We shall calculate this exponent using the Method of Types.

We shall solve this problem by first considering a different one where, before we guess the sequence $\mathbf{x}$, its empirical type $\hat{\mathsf{P}}_{\mathbf{x}}$ is revealed to us. This problem falls under the category of guessing with side information. Let Y denote the type of the sequence $\mathbf{X}$. What can we say about

$$\mathsf{E}[\mathsf{G}^*(X|Y)^\rho]?$$

Conditional on its type, the sequence $\mathbf{X}$ is uniformly distributed over the type class $\mathcal{T}(\hat{\mathsf{P}}_{\mathbf{x}})$. Consequently, using the expression for the number of guesses required to guess a uniform chance variable (Example 1) we obtain

$$\mathsf{E}\Big[\mathsf{G}^*\big(X\big|\hat{\mathsf{P}}_{\mathbf{x}}=Q\big)^\rho\Big] \le |\mathcal{T}(Q)|^\rho \tag{72}$$

$$\le 2^{n\,\mathsf{H}(Q)\rho}. \tag{73}$$

The probability that $\mathbf{X}\sim P^n$ will be of type Q is upper-bounded by $2^{-n\,\mathsf{D}(Q\|P)}$ (and is zero if Q is not in $\mathcal{P}_n(\mathcal{X})$). Thus,

$$\mathsf{E}[\mathsf{G}^*(X|Y)^\rho] = \sum_{Q\in\mathcal{P}_n(\mathcal{X})} P^n\big(\hat{\mathsf{P}}_{\mathbf{x}}=Q\big)\,\mathsf{E}\Big[\mathsf{G}^*\big(X\big|\hat{\mathsf{P}}_{\mathbf{x}}=Q\big)^\rho\Big] \tag{74}$$

$$\le \sum_{Q\in\mathcal{P}_n(\mathcal{X})} 2^{-n\,\mathsf{D}(Q\|P)}\,2^{n\,\mathsf{H}(Q)\rho} \tag{75}$$

$$\le |\mathcal{P}_n(\mathcal{X})| \max_{Q\in\mathcal{P}_n(\mathcal{X})} 2^{n(\rho\,\mathsf{H}(Q)-\mathsf{D}(Q\|P))} \tag{76}$$

$$\le (n+1)^{|\mathcal{X}|} \max_{Q\in\mathcal{P}_n(\mathcal{X})} 2^{n(\rho\,\mathsf{H}(Q)-\mathsf{D}(Q\|P))} \tag{77}$$

$$= (n+1)^{|\mathcal{X}|}\,2^{n\max_{Q\in\mathcal{P}_n(\mathcal{X})}\{\rho\,\mathsf{H}(Q)-\mathsf{D}(Q\|P)\}} \tag{78}$$

$$\le (n+1)^{|\mathcal{X}|}\,2^{n\rho\,\mathsf{H}_{1/(1+\rho)}(X)}. \tag{79}$$

Since the side-information (consisting of the type of $\mathbf{X}$) takes on $|\mathcal{P}_n(\mathcal{X})|$ values, it follows by Proposition 6 and (79) that

$$\mathsf{E}[\mathsf{G}^*(X)^\rho] \le (n+1)^{|\mathcal{X}|(1+\rho)}\,2^{n\rho\,\mathsf{H}_{1/(1+\rho)}(X)}. \tag{80}$$

The exponent with which the ρ-th moment of the number of guesses grows with n is thus upper-bounded by

$$\rho\,\mathsf{H}_{1/(1+\rho)}(X).$$

We next obtain a lower bound that shows that this is the correct exponent. First recall that $\mathsf{E}[\mathsf{G}^*(X)^\rho]$ is lower-bounded by $\mathsf{E}[\mathsf{G}^*(X|Y)^\rho]$ (see (34)), so it suffices to obtain a lower bound on $\mathsf{E}[\mathsf{G}^*(X|Y)^\rho]$.

For every $Q\in\mathcal{P}_n(\mathcal{X})$, the LHS of (2) establishes that

$$\mathsf{E}\big[\mathsf{G}^*\big(X\big|\hat{\mathsf{P}}_{\mathbf{x}}=Q\big)^\rho\big] \ge \frac{1}{1+\rho}\cdot|\mathcal{T}(Q)|^\rho$$

$$\ge \frac{1}{1+\rho}\,\frac{1}{(n+1)^{\rho|\mathcal{X}|}}\,2^{n\rho\,\mathsf{H}(Q)},\quad Q\in\mathcal{P}_n(\mathcal{X}).$$

We shall now average this over the empirical type of $\mathbf{X}$ as follows: For every $Q \in \mathcal{P}_n(\mathcal{X})$, the probability that $\mathbf{X}$ has empirical type Q is lower-bounded by

$$\frac{1}{(n+1)^{|\mathcal{X}|}}\, 2^{-n\,\mathsf{D}(Q\|P)},$$

so

$$\begin{aligned}
\mathsf{E}[\mathsf{G}^*(X|Y)^\rho] &= \sum_{Q\in\mathcal{P}_n(\mathcal{X})} P^n\big(\hat{\mathsf{P}}_{\mathbf{x}} = Q\big)\, \mathsf{E}\Big[\mathsf{G}^*\big(X\big|\hat{\mathsf{P}}_{\mathbf{x}} = Q\big)^\rho\Big] && (81)\\
&\geq \frac{1}{1+\rho}\,\frac{1}{(n+1)^{(1+\rho)|\mathcal{X}|}} \sum_{Q\in\mathcal{P}_n(\mathcal{X})} 2^{-n\,\mathsf{D}(Q\|P)}\, 2^{n\rho\,\mathsf{H}(Q)} && (82)\\
&\geq \frac{1}{1+\rho}\,\frac{1}{(n+1)^{(1+\rho)|\mathcal{X}|}} \max_{Q\in\mathcal{P}_n(\mathcal{X})} 2^{n(\rho\,\mathsf{H}(Q)-\mathsf{D}(Q\|P))} && (83)\\
&= \frac{1}{1+\rho}\,\frac{1}{(n+1)^{(1+\rho)|\mathcal{X}|}}\, 2^{n\max_{Q\in\mathcal{P}_n(\mathcal{X})}\{\rho\,\mathsf{H}(Q)-\mathsf{D}(Q\|P)\}}. && (84)
\end{aligned}$$

The variational distance between any $Q \in \mathcal{P}(\mathcal{X})$ and the PMF in $\mathcal{P}_n(\mathcal{X})$ that is closest to it tends to zero as $n \to \infty$. Consequently, since $\rho\,\mathsf{H}(Q) - \mathsf{D}(Q\|P)$ is continuous in Q in the variational distance,

$$\begin{aligned}
\lim_{n\to\infty} \max_{Q\in\mathcal{P}_n(\mathcal{X})} \{\rho\,\mathsf{H}(Q) - \mathsf{D}(Q\|P)\} &= \max_{Q\in\mathcal{P}(\mathcal{X})} \{\rho\,\mathsf{H}(Q) - \mathsf{D}(Q\|P)\} && (85)\\
&= \rho\,\mathsf{H}_{\frac{1}{1+\rho}}(X) && (86)
\end{aligned}$$

and the lower and upper bound are exponentially tight. We thus conclude that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathsf{E}\big[\mathsf{G}^*(X^n)^\rho\big] = \rho\,\mathsf{H}_{\frac{1}{1+\rho}}(X). \tag{87}$$

4 Rényi's Order-α Divergence

Rényi's order-α divergence $\mathsf{D}_\alpha(P\|Q)$ is defined for PMFs P, Q on the finite set $\mathcal{X}$ as

$$\mathsf{D}_\alpha(P\|Q) \triangleq \frac{1}{\alpha-1} \log \sum_{x\in\mathcal{X}} P(x)^\alpha\, Q(x)^{1-\alpha}. \tag{88}$$

Lemma 9 (Key Properties of $\mathsf{D}_\alpha(P\|Q)$).

a) $\mathsf{D}_\alpha(P\|Q)$ *is nonnegative and is zero if, and only if,* $P = Q$.

b) $\lim_{\alpha\to 1} \mathsf{D}_\alpha(P\|Q) = \mathsf{D}(P\|Q)$.

Proposition 10. *Given any PMF* P *on* $\mathcal{X}$ *and any* $\rho > 0$*, the maximum*

$$\max_{Q,R} \big\{\rho\,\mathsf{H}(Q) - \mathsf{D}(R\|P)\big\} \tag{89}$$

over all PMFs Q *and* R *over* $\mathcal{X}$ *that satisfy*

$$\mathsf{H}(Q) + \mathsf{D}(Q\|P) \leq \mathsf{H}(R) + \mathsf{D}(R\|P) \tag{90}$$

is achieved when Q *and* R *are equal; it is thus equal to*

$$\max_{\tilde{Q}} \big\{\rho\,\mathsf{H}\big(\tilde{Q}\big) - \mathsf{D}\big(\tilde{Q}\big\|P\big)\big\}. \tag{91}$$

Proof: When R and Q are equal the constraint (90) is clearly satisfied. We thus need to show that we can replace any pair (R, Q) satisfying (90) with some pair $(\tilde{Q}, \tilde{Q})$ without decreasing the target function. Our choice of $\tilde{Q}$ is

$$(\tilde{Q}, \tilde{Q}) = \begin{cases} (R, R) & \text{if } \mathsf{H}(R) > \mathsf{H}(Q), \\ (Q, Q) & \text{otherwise.} \end{cases} \tag{92}$$

To see that this choice does not decrease the target function, we consider two cases:

Case I: $\mathsf{H}(R) > \mathsf{H}(Q)$. In this case it is clear from (89) that replacing (R, Q) with (R, R) increases the target function.

Case II: $\mathsf{H}(Q) \geq \mathsf{H}(R)$. In this case we replace R with Q with the result of the target function being

$$\rho\, \mathsf{H}(Q) - \mathsf{D}(Q\|P)$$

instead of

$$\rho\, \mathsf{H}(Q) - \mathsf{D}(R\|P).$$

We have to show that $\mathsf{D}(R\|P) \geq \mathsf{D}(Q\|P)$. But this follows from the assumption $\mathsf{H}(Q) \geq \mathsf{H}(R)$ (the case under consideration) and the constraint (90) which can be rewritten as

$$\mathsf{D}(R\|P) - \mathsf{D}(Q\|P) \geq \mathsf{H}(Q) - \mathsf{H}(R). \tag{93}$$

□

5 Conditional Rényi Entropy

We next consider the case where the pairs

$$(X_1, Y_1),\ (X_2, Y_2),\ \ldots, (X_n, Y_n)$$

are IID according to the joint PMF $P_{X,Y}$ on the finite set $\mathcal{X} \times \mathcal{Y}$, and where we need to guess the sequence $\mathbf{X} = X_1^n$ after observing $\mathbf{Y} = Y_1^n$. The sequence $\mathbf{Y}$ thus serves as side information for guessing $\mathbf{X}$. We shall refrain from using Proposition 6 because $\mathbf{Y}$ takes value in the set $\mathcal{Y}^n$ whose size is exponential in n, and the bound (35) turns out to be loose. We will however use an inequality related to (35).

Proposition 11. *Let X, Y, Z be of some joint PMF $P_{X,Y,Z}$ on the finite set $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Then*

$$\mathsf{E}[\mathsf{G}^*(X|Y)^\rho] \leq |\mathcal{Z}|^\rho\, \mathsf{E}[\mathsf{G}^*(X|Y,Z)^\rho]. \tag{94}$$

Proof: Conditioning on $Y = y$ and applying Proposition 6, we obtain

$$\mathsf{E}[\mathsf{G}^*(X|Y=y)^\rho] \leq |\mathcal{Z}|^\rho\, \mathsf{E}[\mathsf{G}^*(X|Y=y,Z)^\rho], \tag{95}$$

where the expectation on the LHS is with respect to $P_{X|Y=y}$ and on the RHS with respect to $P_{X,Z|Y=y}$. Taking expectations on both sides with respect to P_Y yields the desired inequality.

□

We shall apply this inequality by substituting X_1^n for X, Y_1^n for Y, and the joint type of X_1^n and Y_1^n for Z. We begin the evaluation of

$$\mathsf{E}[\mathsf{G}^*(X|Y,Z)^\rho]$$

by evaluating the ρ-th moment of the number of guesses conditional on Y_1^n being equal to y_1^n and the joint type of (X_1^n, Y_1^n) being $R(y) \cdot V(x|y)$ for some $R \in \mathcal{P}(\mathcal{Y})$ and $V \in \mathcal{P}(\mathcal{X}|\mathcal{Y})$. (Here, $R(y)$ is the empirical type of y_1^n.) Under this conditioning, X_1^n is uniformly distributed over the V-shell of y_1^n, and the ρ-th moment of the number of guesses required to guess X_1^n is upper-bounded by

$$|\mathcal{T}_V(y_1^n)|^\rho$$

and hence by

$$2^{n\rho\,\mathsf{H}(V|R)}. \tag{96}$$

Since this depends only on V and R (and ρ), the averaging over $\mathcal{Y}^n$ (with respect to the conditional law of Y_1^n given that the joint type of (X_1^n, Y_1^n) is $R \circ V \triangleq R(y)V(x|y)$) yields

$$2^{n\rho\,\mathsf{H}(V|R)}. \tag{97}$$

Averaging (97) over the joint type of (X_1^n, Y_1^n) and recalling that when (X_1^n, Y_1^n) are drawn IID $\sim P_{X,Y}$ the probability that their joint type equals $R \circ V$ is upper-bounded by $2^{-n\,\mathsf{D}(R\circ V\|P_{X,Y})}$ yields that

$$\begin{aligned}
\mathsf{E}\big[\mathsf{G}^*(X_1^n|Y_1^n,Z)^\rho\big] &\le \sum_{R\circ V\in\mathcal{P}_n(\mathcal{X}\times\mathcal{Y})} 2^{-n\,\mathsf{D}(R\circ V\|P_{X,Y})}\, 2^{n\rho\,\mathsf{H}(V|R)} && (98)\\
&\le (n+1)^{|\mathcal{X}||\mathcal{Y}|}\, 2^{n\max_{R,V}\{\rho\,\mathsf{H}(V|R)-\mathsf{D}(R\circ V\|P_{X,Y})\}} && (99)\\
&= (n+1)^{|\mathcal{X}||\mathcal{Y}|}\, 2^{n\rho\,\mathsf{H}_{1/(1+\rho)}(X_1|Y_1)}, && (100)
\end{aligned}$$

where the last equality holds because of the following proposition.

Proposition 12.

$$\mathsf{H}_{\frac{1}{1+\rho}}(X_1|Y_1) = \max_{\substack{R\in\mathcal{P}(\mathcal{Y})\\ V\in\mathcal{P}(\mathcal{X}|\mathcal{Y})}} \big\{\mathsf{H}(V|R) - \rho^{-1}\,\mathsf{D}(R\circ V\|P_{X,Y})\big\} \tag{101}$$

Proof: We first show that $\mathsf{H}(V|R)-\rho^{-1}\,\mathsf{D}(R\circ V\|P_{X,Y}) \le \mathsf{H}_{1/(1+\rho)}(X_1|Y_1)$ for every $R \in \mathcal{P}(\mathcal{Y})$ and $V \in \mathcal{P}(\mathcal{X}|\mathcal{Y})$. This is clearly true when $\mathsf{D}(R \circ V\|P_{X,Y}) = \infty$, so we may assume that $P_{X,Y}(x,y) = 0$ implies $R(y)V(x|y) = 0$, and hence that $P_Y(y) = 0$ implies $R(y) = 0$. Now observe that

$$\begin{aligned}
&\mathsf{H}(V|R) - \rho^{-1}\,\mathsf{D}(R\circ V\|P_{X,Y})\\
&\quad= \sum_{y\in\mathcal{Y}} R(y) \sum_{x\in\mathcal{X}} V(x|y)\log\frac{1}{V(x|y)} - \frac{1}{\rho}\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} R(y)\,V(x|y)\log\frac{R(y)\,V(x|y)}{P_{X,Y}(x,y)} && (102)\\
&\quad= \sum_{y\in\mathcal{Y}} R(y) \sum_{x\in\mathcal{X}} V(x|y)\log\frac{P_{X|Y}(x|y)^{\frac{1}{\rho}}}{V(x|y)^{1+\frac{1}{\rho}}} - \frac{1}{\rho}\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} R(y)\,V(x|y)\log\frac{R(y)}{P_Y(y)} && (103)\\
&\quad= \frac{1+\rho}{\rho}\sum_{y\in\mathcal{Y}} R(y) \sum_{x\in\mathcal{X}} V(x|y)\log\frac{P_{X|Y}(x|y)^{\frac{1}{1+\rho}}}{V(x|y)} - \frac{1}{\rho}\sum_{y\in\mathcal{Y}} R(y)\log\frac{R(y)}{P_Y(y)} && (104)
\end{aligned}$$

$$\le \frac{1+\rho}{\rho} \sum_{y\in\mathcal{Y}} R(y) \log\left(\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)^{\frac{1}{1+\rho}}\right) - \frac{1}{\rho}\sum_{y\in\mathcal{Y}} R(y)\log\frac{R(y)}{P_Y(y)} \tag{105}$$

$$= \frac{1}{\rho}\sum_{y\in\mathcal{Y}} R(y)\log\frac{P_Y(y)\Big(\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)^{\frac{1}{1+\rho}}\Big)^{1+\rho}}{R(y)} \tag{106}$$

$$\le \frac{1}{\rho}\log\left(\sum_{y\in\mathcal{Y}} P_Y(y)\left(\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)^{\frac{1}{1+\rho}}\right)^{1+\rho}\right) \tag{107}$$

$$= \mathsf{H}_{\frac{1}{1+\rho}}(X_1|Y_1), \tag{108}$$

where (105) and (107) follow from Jensen's Inequality. The proof is completed by noting that equality is attained in both inequalities by the choice

$$R(y) = \frac{P_Y(y)\Big(\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)^{\frac{1}{1+\rho}}\Big)^{1+\rho}}{\sum_{y'\in\mathcal{Y}} P_Y(y')\Big(\sum_{x'\in\mathcal{X}} P_{X|Y}(x'|y')^{\frac{1}{1+\rho}}\Big)^{1+\rho}} \tag{109}$$

and

$$V(x|y) = \frac{P_{X|Y}(x|y)^{\frac{1}{1+\rho}}}{\sum_{x'\in\mathcal{X}} P_{X|Y}(x'|y)^{\frac{1}{1+\rho}}}, \quad R(y) > 0. \tag{110}$$

(Note that $P_Y(y) > 0$ when $R(y) > 0$ so the RHS of (110) makes sense. How we define $V(x|y)$ when $R(y) = 0$ does not matter.) □

Next we turn to find a lower bound. Since the side information Z cannot hurt,

$$\mathsf{E}\big[\mathsf{G}^*(X|Y)^\rho\big] \ge \mathsf{E}\big[\mathsf{G}^*(X|Y,Z)^\rho\big], \tag{111}$$

and it remains to lower-bound the RHS. As in the upper bound, we substitute X_1^n for X, Y_1^n for Y, and the joint type of X_1^n and Y_1^n for Z, and we lower-bound the ρ-th moment of the number of guesses conditional on Y_1^n being equal to y_1^n and the joint type of (X_1^n, Y_1^n) being $R\circ V$ for some $R\in\mathcal{P}(\mathcal{Y})$ and $V\in\mathcal{P}(\mathcal{X}|\mathcal{Y})$. (Here, $R(y)$ is the empirical type of y_1^n.) Under this conditioning, X_1^n is uniformly distributed over the V-shell of y_1^n, and the ρ-th moment of the number of guesses required to guess X_1^n is lower-bounded by

$$\frac{1}{\rho+1}\,|\mathcal{T}_V(y_1^n)|^\rho$$

and hence, for $R\circ V$ in $\mathcal{P}_n(\mathcal{X}\times\mathcal{Y})$, by

$$\frac{1}{\rho+1}\,\frac{1}{(n+1)^{\rho|\mathcal{X}||\mathcal{Y}|}}\,2^{n\rho\,\mathsf{H}(V|R)}. \tag{112}$$

Since this depends only on V and R (and ρ), the averaging over $\mathcal{Y}^n$ (with respect to the conditional law of Y_1^n given that the joint type of (X_1^n, Y_1^n) is $R\circ V \triangleq R(y)V(x|y)$) yields

$$\frac{1}{\rho+1}\,\frac{1}{(n+1)^{\rho|\mathcal{X}||\mathcal{Y}|}}\,2^{n\rho\,\mathsf{H}(V|R)}. \tag{113}$$

Averaging over the joint type of (X_1^n, Y_1^n) and recalling that when $R\circ V$ is in $\mathcal{P}_n(\mathcal{X}\times\mathcal{Y})$ and (X_1^n, Y_1^n) are drawn IID $\sim P_{X,Y}$ the probability that their joint type equals $R\circ V$ is lower-bounded by $(n+1)^{-|\mathcal{X}||\mathcal{Y}|}\,2^{-n\,\mathsf{D}(R\circ V\|P_{X,Y})}$, we obtain

$$\mathsf{E}\big[\mathsf{G}^*(X_1^n|Y_1^n,Z)^\rho\big] \ge \frac{1}{\rho+1}\,\frac{1}{(n+1)^{\rho|\mathcal{X}||\mathcal{Y}|}}\,2^{n\max_{R\circ V\in\mathcal{P}_n(\mathcal{X}\times\mathcal{Y})}\{\rho\,\mathsf{H}(V|R)-\mathsf{D}(R\circ V\|P_{X,Y})\}}. \tag{114}$$

The variational distance between any $R \circ V \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and the PMF in $\mathcal{P}_n(\mathcal{X} \times \mathcal{Y})$ that is closest to it tends to zero as $n \to \infty$. Consequently, using a continuity argument and Proposition 12, we can show that the RHS of (114) grows exponentially with n with an exponent that equals $\rho\, \mathsf{H}_{\frac{1}{1+\rho}}(X_1|Y_1)$.

From this and from (100), we thus obtain

$$\lim_{n\to\infty} \frac{1}{n} \log \mathsf{E}\left[\mathsf{G}^*(X_1^n|Y_1^n)^\rho\right] = \rho\, \mathsf{H}_{\frac{1}{1+\rho}}(X_1|Y_1). \tag{115}$$