Appendix A

Probability and Information Theory

A.1 Random Variables and Distributions

A **finite probability space** is given by a non-empty finite set Ω and a **probability function** $P : \Omega \to [0,1]$ with $\sum_{\omega\in\Omega} P(\omega) = 1$. The subsets of Ω are called *events*. The **probability** of an event $\Lambda \subseteq \Omega$ is given by $P[\Lambda] := \sum_{\omega\in\Lambda} P(\omega)$, and for two events Λ and Γ with $P[\Gamma] > 0$, the **conditional probability** $P[\Lambda \,|\, \Gamma]$ is defined as $P[\Lambda \,|\, \Gamma] := P[\Lambda \cap \Gamma]/P[\Gamma]$.

A **random variable** is a function $X : \Omega \to \mathcal{X}$, where we may assume the range $\mathcal{X}$ to be finite. The **distribution** of X is the function $P_X : \mathcal{X} \to [0,1]$ defined as $P_X(x) = P[X = x]$, where $X = x$ is a shorthand for the event $\{\omega \in \Omega \,|\, X(\omega) = x\}$. We write P_{XY} for the **joint distribution** of two random variables X and Y, i.e. $P_{XY}(x,y) = P[X = x \cap Y = y]$, and similarly for more than two random variables. Also, we write $P_{X|\Gamma}(x) = P[X = x \,|\, \Gamma]$ and $P_{X|Y}(x|y) = P_{X|Y=y}(x) = P[X = x \,|\, Y = y]$ for the respective **conditional distributions** (conditioned on an event Γ, and a random variable Y).

Two random variables X and Y are **independent** if $P_{XY} = P_X \cdot P_Y$, in the sense that $P_{XY}(x,y) = P_X(x) \cdot P_Y(y)$ for all x and y's in the corresponding ranges of X and Y.

The **expectation** of a real-valued random variable Y is defined as $E[Y] := \sum_y P_Y(y) \cdot y$. If Y is of the form $Y = f(X)$ for a random variable X and a real-valued function f, then this equals $E[f(X)] = \sum_x P_X(x) \cdot f(x)$, which we also write as $E_{x \leftarrow X}[f(x)]$ in some occasions.

Throughout, we leave the probability space (Ω, P) implicit. Whenever we consider a random variable (or several random variables), we understand an underlying finite probability space (Ω, P) to be given; as such, the (joint) distribution of the random variable(s) is assumed to be given as well. We may also specify a random variable (or several random variables) by means of an "experiment", which uniquely determines the (joint) distribution of the random variable(s).

Finally, in a general context, i.e., when not necessarily associated to any particular random variable, a **distribution** is simply an arbitrary function $Q : \mathcal{X} \to [0,1]$ on a finite set $\mathcal{X}$ with the property that $\sum_x Q(x) = 1$.

Definition A.1. *The* **statistical distance** *between two distributions P and Q with common domain $\mathcal{X}$ is defined as*

$$\delta(P,Q) := \frac{1}{2} \sum_{x\in\mathcal{X}} \bigl|P(x) - Q(x)\bigr|$$

If the distributions describing two experiments have small statistical distance, then this can be interpreted as that the experiments behave in exactly the same way except with small "error" probability. This is formalized as follows.

Lemma A.1. *Let Q and Q' be two distributions with common domain $\mathcal{X}$. Then there exists a joint distribution $P_{XX'}$ for random variables X and X' such that $P_X = Q$ and $P_{X'} = Q'$, and such that $P[X \neq X'] = \delta(Q, Q')$.*

Corollary A.2. *Let Q and Q' be two distributions with common domain $\mathcal{X}$, and let $T \subseteq \mathcal{X}$ be an arbitrary subset. Then, $|Q(T) - Q'(T)| \leq \delta(Q, Q')$.*

The latter means that for any *test* to decide whether a sample x was chosen according to Q or according to Q', if its probability to give the *correct* answer when x was chosen according to Q is p, then its probability to give the *wrong* answer when x was chosen according to Q' is at least $p - \delta(Q, Q')$. Thus, if p is large then so is $p - \delta(Q, Q')$ if $\delta(Q, Q')$ is small.

A.2 Shannon Entropy

Let X and Y be random variables with respective ranges $\mathcal{X}$ and $\mathcal{Y}$. Throughout, log denotes the *binary* logarithm.

Definition A.2. *The* **(Shannon) entropy** *of X is defined as*

$$\mathrm{H}(X) := -\sum_{x} P_X(x) \log P_X(x)\,,$$

where the sum is over all $x \in \mathcal{X}$ with $P_X(x) > 0$.

It is not hard to see that $0 \leq \mathrm{H}(X) \leq \log |\mathcal{X}|$, with equality on the left if and only if X is constant, i.e. if there is no uncertainty at all in X, and with equality on the right if and only if X is uniform over $\mathcal{X}$, i.e. has maximal uncertainty.

Note that $\mathrm{H}(X)$ is actually a function of the *distribution* P_X of X, and thus we may also write $\mathrm{H}(Q)$ for any distribution Q. Thus, the notion naturally extends to

$$\mathrm{H}(XY) := \mathrm{H}(P_{XY}) = -\sum_{x,y} P_{XY}(x,y) \log P_{XY}(x,y)\,,$$

$$\mathrm{H}(X|Y\!=\!y) := \mathrm{H}(P_{X|Y=y}) = -\sum_{x} P_{X|Y}(x|y) \log P_{X|Y}(x|y)\,,$$

etc.

Definition A.3. *The* **conditional (Shannon) entropy** *of X given Y is defined as*

$$\mathrm{H}(X|Y) := \sum_{y} P_Y(y)\, \mathrm{H}(X|Y\!=\!y)\,,$$

where the sum is over all $y \in \mathcal{Y}$ with $P_Y(y) > 0$.

The following rules hold; the first two rules are called **monotonicity** and **strong subadditivity**, respectively, and the third rule is called **chain rule**.

Lemma A.3. *For any random variables X, Y and Z:*

1. $\mathrm{H}(XY|Z) \geq \mathrm{H}(X|Z)$,
2. $\mathrm{H}(X|Z) \geq \mathrm{H}(X|YZ)$, *and*
3. $\mathrm{H}(X|YZ) = \mathrm{H}(XY|Z) - \mathrm{H}(Y|Z)$.

The Shannon entropy has proven to be the right measure for "uncertainty" in communication theory. For instance, it captures to what extent data can be compressed, but also tells us how much information can be reliably communicated over a noisy communication channel.

A.3 Beyond Shannon Entropy

The Shannon entropy is mainly useful in the context of average-case properties of information, like how much can data be compressed *on average*. On case of *one-shot* properties, other measures typically take over. We discuss some here, most notably the **min-entropy**.

As above, let X and Y be random variables with respective ranges $\mathcal{X}$ and $\mathcal{Y}$.

Definition A.4. *The* **guessing probability** *and the* **min-entropy** *of X are respectively defined as*

$$\mathrm{Guess}(X) := \max_x P_X(x) \qquad \text{and} \qquad \mathrm{H}_\infty(X) := -\log\bigl(\mathrm{Guess}(X)\bigr) = -\log\bigl(\max_x P_X(x)\bigr)\,.$$

Like the Shannon entropy, the min-entropy is 0 if and only if X is constant, and maximal, i.e., $\log|\mathcal{X}|$ if and only if X is uniform on $\mathcal{X}$, but in-between these two extremes, the min-entropy behaves differently (actually: more conservatively).

Also here, $\mathrm{Guess}(X)$ and $\mathrm{H}_\infty(X)$ are actually functions of the *distribution* P_X of X, and thus we may also write $\mathrm{Guess}(Q)$ and $\mathrm{H}_\infty(Q)$ for any distribution Q, and, as such, $\mathrm{Guess}(XY)$, $\mathrm{H}_\infty(X|Y\!=\!y)$, etc. are naturally defined.

Definition A.5. *The* **conditional guessing probability** *and the* **conditional min-entropy** *of X given Y are respectively defined as*

$$\mathrm{Guess}(X|Y) := \sum_y P_Y(y)\,\mathrm{Guess}(X|Y\!=\!y) \qquad \text{and} \qquad \mathrm{H}_\infty(X|Y) := -\log\bigl(\mathrm{Guess}(X|Y)\bigr)\,.$$

Warning: Different notions of *conditional* min-entropy can be found in the literature. The one we are using here is nowadays considered to be “the right one”.

By replacing the guessing probability by the collision probability, we obtain the notion of (conditional) collision entropy as follows.

Definition A.6. *The* **collision probability** *and the* **collision entropy** *of X are respectively defined as*

$$\mathrm{Col}(X) := \sum_x P_X(x)^2 \qquad \text{and} \qquad \mathrm{H}_2(X) := -\log\bigl(\mathrm{Col}(X)\bigr) = -\log\Bigl(\sum_x P_X(x)^2\Bigr)\,,$$

and the **conditional collision probability** *and* **entropy** *of X given Y as*

$$\mathrm{Col}(X|Y) := \Bigl(\sum_y P_Y(y)\sqrt{\mathrm{Col}(X|Y\!=\!y)}\Bigr)^2 \qquad \text{and} \qquad \mathrm{H}_2(X|Y) := -\log\bigl(\mathrm{Col}(X|Y)\bigr)\,,$$

where naturally $\mathrm{Col}(X|Y\!=\!y) := \sum_x P_{X|Y}(x|y)^2$.

Monotonicity and strong subadditivity still hold, plus a weak form of chain rule.

Lemma A.4. *For any random variables X, Y and Z, and for $\alpha \in \{2,\infty\}$:*

1. $\mathrm{H}_\alpha(XY|Z) \geq \mathrm{H}_\alpha(X|Z)$,
2. $\mathrm{H}_\alpha(X|Z) \geq \mathrm{H}_\alpha(X|YZ)$, *and*
3. $\mathrm{H}_\alpha(X|YZ) \geq \mathrm{H}_\alpha(XY|Z) - \log(|\mathcal{Y}|) \geq \mathrm{H}_\alpha(X|Z) - \log(|\mathcal{Y}|)$.

Finally, the different entropy notions compare to each other as follows.

Lemma A.5. *For any random variables X and Z:* $\mathrm{H}_\infty(X|Z) \leq \mathrm{H}_2(X|Z) \leq \mathrm{H}(X|Z)$.

The Shannon entropy, the min-entropy and the collision entropy are all special cases of the so-called **Rényi entropy** of order α, where $0 \leq \alpha \leq \infty$. It is defined as $\mathrm{H}_\alpha(X) := -\log(\mathrm{Ren}_\alpha(X))$, respectively as $\mathrm{H}_\alpha(X|Y) := -\log(\mathrm{Ren}_\alpha(X|Y))$ for the conditional version, where

$$\mathrm{Ren}_\alpha(X) := \Big(\sum_x P_X(x)^\alpha\Big)^{\frac{1}{\alpha-1}} \quad \text{and} \quad \mathrm{Ren}_\alpha(X|Y) := \Big(\sum_y P_Y(y)\,\mathrm{Ren}_\alpha(X|Y\!=\!y)^{\frac{\alpha-1}{\alpha}}\Big)^{\frac{\alpha}{\alpha-1}}$$

for any α for which the expressions are well-defined. It is easy to see that for $\alpha = 2$ the Rényi entropy H_α coincides with the collision entropy, and one can show that $\mathrm{H}_\alpha(X) \to \mathrm{H}_\infty(X)$ for $\alpha \to \infty$ and $\mathrm{H}_\alpha(X) \to \mathrm{H}(X)$ for $\alpha \to 1$, and similarly for the conditional versions. In the limit $\alpha \to 0$ we obtain $\mathrm{H}_0(X) := \log|\mathrm{supp}(P_X)|$. Furthermore, Lemma A.4 above holds for the Rényi entropy H_α of any order α, and the Rényi entropy is monotonically decreasing in α, generalizing Lemma A.5 above.

Using the standard notion of the the p-norm $\|\cdot\|_p$ for real-valued functions with finite domain, $\mathrm{Ren}_\alpha(X)$ and $\mathrm{Ren}_\alpha(X|Y)$ can be nicely written as

$$\mathrm{Ren}_\alpha(X) = \|P_X\|_\alpha^{\frac{\alpha}{\alpha-1}} \quad \text{and} \quad \mathrm{Ren}_\alpha(X|Y) := \Big(\sum_y P_Y(y)\,\|P_{X|Y=y}\|_\alpha\Big)^{\frac{\alpha}{\alpha-1}}$$

for $0 < \alpha \neq 1$. Finally, with respect to the quantum generalization, it is useful to observe that

$$\mathrm{H}_\alpha(X|Y) = \max_{Q_Y} \frac{1}{1-\alpha} \log \sum_{x,y} \frac{P_{XY}(x,y)^\alpha}{Q(y)^{\alpha-1}}\,,$$

where the max is over all distributions $Q_Y : \mathcal{Y} \to [0,1]$. The equality can be shown by solving the optimization problem, e.g. using Lagrange multipliers.