3: Randomness and Entropy

A Uncertainty and randomness

In the previous two lectures, we identified entropy as a measure of uncertainty and therefore, the information revealed on observing $X \sim P$ is $H(P)$.

Information = reduction of uncertainty

But does uncertainty equal randomness?

How much randomness is there in $X \sim P$?

(two possible definitions)

how many random bits are needed to generate $X \sim P$?

how many random bits can be generated from $X \sim P$?

Some examples:

Example 1. How many independent unbiased coin flips are needed to generate samples from:

(a) $P = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$ (b) $P = \left(\frac{2}{3}, \frac{1}{3}\right)$

avg. no.

worst case

Example 2. How many independent samples on avg. from $\mathrm{Ber}(p)$ are needed to generate one sample from $\mathrm{Ber}(1/2)$?

B Measuring distance between distributions:

Total variation distance $d(P,Q)$

$$d(P,Q) := \frac{1}{2}\sum_{x} |P(x)-Q(x)|$$

Properties

1. $$d(P,Q) = \sup_{A} P(A)-Q(A) = \sup_{A} |P(A)-Q(A)|$$
$$= \sup_{B} Q(B)-P(B) = \sup_{B} |Q(B)-P(B)|$$

where the sup in the first line is attained by

$$A^* = \{x : P(x) > Q(x)\}$$

and that in the second line by

$$B^* = \{x : Q(x) > P(x)\}$$

2. $$0 \le d(P,Q) \le 1,$$

with equality on the left-side iff $P = Q$

and equality on the right-side iff $supp(P) \cap supp(Q) = \emptyset$.

3. We can extend the definition to distributions with densities:

$$d(P,Q) = \frac{1}{2}\int |\underbrace{f(x)}_{\text{density of } P} - \underbrace{g(x)}_{\text{density of } Q}|\, dx$$
$$= \sup_{A} P(A)-Q(A)$$

where the equality in the second line attained by
$A^* = \{x : f(x) > g(x)\}$.

C Generating almost random bits

Given a sample $X \sim P$, we want to generate $U \sim \mathrm{unif}(\{0,1\}^\ell)$.

Specifically, for $0 < \varepsilon < 1$, what is the largest ℓ s.t. we can find $f : \mathcal{X} \to \{0,1\}^\ell$ s.t.

$$d(P_{f(X)}, P_U) \le \varepsilon \, ?$$

A scheme

Let $\mathcal{T}_\lambda = \{x : -\log p(x) > \lambda\}$ and suppose that $P(\mathcal{T}_\lambda) \ge 1 - \varepsilon$.

Consider a partition of $\mathcal{T}_\lambda$ into parts $\mathcal{X}_1, \dots, \mathcal{X}_M, \mathcal{X}_{M+1}$ that each part $\mathcal{X}_i$ satisfies

(#1) $$2^{-\lambda} \cdot N < P(\mathcal{X}_i) \le 2^{-\lambda} \cdot (N+1), \quad \text{for } 1 \le i \le M.$$

and

(#2) $$P(\mathcal{X}_{M+1}) \le 2^{-\lambda} \cdot N$$

we can only find such a partition because $p(x) \le 2^{-\lambda} \; \forall x \in \mathcal{T}_\lambda$.

Thus,

$$\left(\frac{1 - \varepsilon - 2^{-\lambda} \cdot N}{N+1} \right) 2^\lambda \le M < \frac{2^\lambda}{N}.$$

Let $U \sim \mathrm{unif}(\{1, \dots, M'\})$ where $M' = \frac{2^\lambda}{N}$ and

$$f(x) = \begin{cases} i, & \text{if } x \in \mathcal{X}_i \text{ for some } i, \\ \perp, & \text{if } x \notin \mathcal{T}_\lambda. \end{cases}$$

We now analyse our scheme.

$$2\,d(P_{f(x)},P_U) = \sum_{m=1}^{M}\left|P_{f(x)}(m)-\frac{1}{M'}\right| + \sum_{m=M+1}^{M'}\left|P_{f(x)}(m)-\frac{1}{M'}\right| + P_{f(x)}(\perp)$$

$$P_{f(x)}(\perp) = P(J_\lambda^c) \le \varepsilon$$

$$P_{f(x)}(M+1) \le \frac{2^{-\lambda}}{N}, \quad P_U(M+1) = \frac{2^{-\lambda}}{N} = \frac{1}{M'}$$

$$P_{f(x)}(m) = 0, \quad P_U(m) = \frac{2^{-\lambda}}{N}, \text{ for } m \in \{M+2,\dots,M'\}.$$

$$\Rightarrow \sum_{m=M+1}^{M'}\left|P_{f(x)}(m)-\frac{1}{M'}\right| = \frac{M'-M}{M'} - P_{f(x)}(M+1)$$

$$\le 1-\frac{M}{M'} \le 1-\frac{(1-\varepsilon-2^{-\lambda}\cdot N)}{N+1}\cdot N$$

$$\sum_{m=1}^{M}\left|P_{f(x)}(m)-\frac{1}{M'}\right| = \sum_{m=1}^{M}\left|P(\mathcal{X}_m)-\frac{1}{M'}\right|$$

$$\le M\cdot 2^{-\lambda}$$

$$\le \frac{1}{N},$$

where we used (#1) and (#2).

On combining the bounds above, we get

$$2\,d(P_{f(x)},P_U) \le \frac{1}{N} + \frac{1}{N+1} + 2\varepsilon + \varepsilon\cdot 2^{-\lambda}$$

(5)

Setting $N = \frac{1}{\varepsilon} - 1$,

$$2\,d\left(P_{f(x)}, P_U\right) \le 2\left(\varepsilon + \varepsilon + \frac{2^{-\lambda}}{\varepsilon}\right) + \varepsilon$$

$$\Rightarrow d\left(P_{f(x)}, P_U\right) \le \frac{5\varepsilon}{2} + \frac{2 \cdot 2^{-\lambda}}{\varepsilon}.$$

Namely, we have the following theorem:

Theorem. For $\lambda > 0$ and $0 < \varepsilon < 1$, suppose that

$$P\left(\{x : -\log p(x) > \lambda\}\right) \ge 1 - \varepsilon.$$

Then, for $U \sim$ uniformly over at least $\boxed{r = \lambda - \log\frac{1}{\varepsilon}}$ bits, we have

$$d\left(P_{f(x)}, P_U\right) \le \frac{5}{2}\varepsilon + 2 \cdot 2^{-r}.$$

In particular, if $\lambda \ge 2\log\frac{1}{\varepsilon}$, we get

$$d\left(P_{f(x)}, P_U\right) \le \frac{7}{2}\varepsilon.$$

We have generated $\lambda - \log \frac{1}{\varepsilon}$ bits.

So, what is the largest λ we can have?

It's roughly the entropy!

You can use the Chebyshev's inequality to show:

$$P\left(-\log P(x) > \underline{H(P)} - \sqrt{\frac{\mathrm{Var}(-\log P(x))}{\varepsilon}}\right) \ge 1 - \varepsilon.$$

D Generating a sample $X \sim P$ using uniform bits

What is the least number of bits l such that for $U \sim \text{unif}(\{0,1\}^l)$ we can find a function f s.t. $d(P_{f(U)}, P) \leq \varepsilon$?

A scheme

Consider $U \sim \text{unif}(\{1, \ldots, M\})$ and suppose that

$$P\left(\{x : -\log p(x) \leq \lambda\}\right) \geq 1 - \varepsilon.$$

Denoting $\mathcal{J}_\lambda = \{x : -\log p(x) \leq \lambda\}$, each $x \in \mathcal{J}_\lambda$ satisfies $p(x) \geq 2^{-\lambda}$. Suppose that $M > 2^{-\lambda}$. Consider a partition $\mathcal{Y}_1, \ldots, \mathcal{Y}_{M'+1}$ of $\{1, \ldots, M\}$ such that $M' = |\mathcal{J}_\lambda|$ and

$$p(x_i) < P_U(\mathcal{Y}_i) \leq p(x_i) + \frac{1}{M}, \quad 1 \leq i \leq M',$$

where $x_1, \ldots, x_{M'}$ denote the elements of $\mathcal{J}_\lambda$. Further, let $x_{M'+1}$ denote an arbitrary element outside $\mathcal{J}_\lambda$.

Define

$$f(y) = x_i \quad \text{if } y \in \mathcal{Y}_i, \quad 1 \leq i \leq M'+1.$$

Then,

$$2d(P_{f(U)}, P) \leq \sum_{x \in \mathcal{J}_\lambda} |P_{f(U)}(x) - p(x)| + P(\mathcal{J}_\lambda^c) + P_U(\mathcal{Y}_{M'+1})$$

For the first term, by our construction,

$$\left| P_{f(U)}(x_i) - p(x_i) \right| = \left| P_U(y_i) - p(x_i) \right|$$
$$\leq \frac{1}{M} \quad \text{for } i \in \{1, \dots, M'\}.$$

Thus,
$$\sum_{x \in J} \left| P_{f(U)}(x) - p(x) \right| \leq \frac{M'}{M}.$$

For the third term, for each $i \in \{1, \dots, M'\}$,

$$p(x_i) \leq P_U(y_i),$$

which on summing over $i \in \{1, \dots, M\}$ gives

$$P(J_\lambda) \leq P_U\left(\bigcup_{i=1}^{M'} y_i\right) = 1 - P_U(y_{M'+1})$$

$$\Rightarrow P_U(y_{M'+1}) \leq P(J_\lambda^c).$$

Thus, noting that $P(J_\lambda^c) \leq \varepsilon$,

$$d(P_{f(U)}, P) \leq \frac{M'}{2M} + \varepsilon = \frac{|J_\lambda|}{2M} + \varepsilon$$
$$\leq \frac{2^\lambda}{2 \cdot M} + \varepsilon \qquad (\text{why?})$$

In particular, upon choosing $M = \frac{2^\lambda}{\varepsilon}$, we get

$d(P_{f(U)}, P) \leq \frac{3\varepsilon}{2}$. We have shown the following result:

Theorem Suppose that

$$P\left(\{x : -\log p(x) \le \lambda\}\right) \ge 1-\varepsilon .$$

Then, for $\boxed{\ell = \lambda + \log \frac{1}{\varepsilon}}$, using $U \sim \mathrm{unif}(\{0,1\}^{\ell})$ we can find a function f s.t.

$$d\left(P_{f(U)}, P\right) \le \frac{3}{2}\varepsilon .$$

We saw in the previous lecture that a good estimate for λ required in the theorem above is $\underline{\underline{\frac{H(P)}{\varepsilon}}}$.

E Typical sets and Entropy

In all our proofs, the following sets appeared:

$\mathcal{J}_\lambda^{(1)} = \{x : -\log p(x) \le \lambda\}$

$\mathcal{J}_\lambda^{(2)} = \{x : -\,\mathrm{lo} \quad x) > \lambda\}$

We want λ both the sets have probability exceeding $1-\varepsilon$.

We saw that a good choice is roughly $\underline{\underline{H(P)}}$.

* $|\mathcal{J}_\lambda^{(1)}| \le 2^{\lambda}$
* If $P(\mathcal{J}_\lambda^{(2)}) \ge 1-\varepsilon$, $|\mathcal{J}_\lambda^{(2)}| \ge 2^{\lambda}(1-\varepsilon)$

These sets are called <u>typical sets</u> since they contain elements which occur with large probability.

Furthermore, probabilities of elements in these sets are "close" to a uniform distribution on this set.

* <u>The case of iid source</u>

$X \equiv (X_1, \ldots, X_n) \sim$ iid P

Then, for $\lambda = n(H(P)+\eta)$ and $Z_i = -\log P(X_i)$

$$P(\mathcal{J}_\lambda^{(1)}) = P\left(\sum_{i=1}^{n} Z_i \leq n(H(P)+\eta)\right)$$
$$\geq 1 - \sqrt{\frac{\mathrm{Var}(Z_1)}{n\eta}}.$$

and for $\lambda = n(H(P)-\eta)$,

$$P(\mathcal{J}_\lambda^{(2)}) = P\left(\sum_{i=1}^{n} Z_i > n(H(P)-\eta)\right)$$
$$\geq 1 - \sqrt{\frac{\mathrm{Var}(Z_1)}{n\eta}}.$$

* <u>Entropy is additive</u>

For independent X_1, X_2,

$$H(X_1, X_2) = H(X_1) + H(X_2)$$

This is perhaps the most important property of Shannon entropy.