Wrapping up the Probability Recap

Recall: For a "numerical" random variable X, we defined

* expectation value or mean: $EX = E[X] = \sum_x P(x) \cdot x$
* variance: $\mathrm{Var}(X) = E\left[(X-EX)^2\right]$

$$\overset{\text{ex}}{=} E[X^2] - (EX)^2$$

Examples

P	Bernoulli(f)		Binomial(n,f)
E	f	$\overset{\text{ex}}{\Longrightarrow}$	$n \cdot f$
Var	$f(1-f)$		$n \cdot f \cdot (1-f)$

$E[(X-EX)^2] = E[(X-f)^2]$
$= f(1-f)^2 + (1-f)(0-f)^2 = f(1-f)$

Three results that give these meaning:

Markov inequality: If $X \geq 0$: $\boxed{\Pr(X \geq t) \leq \frac{E[X]}{t} \quad (\forall t > 0)}$

Pf: $\Pr(X \geq t) = \sum_{x \geq t} P(x) \leq \sum_{x \geq t} P(x) \frac{x}{t} \leq \frac{E[X]}{t}$ $\square$

Chebyshev inequality: $\boxed{\Pr(|X - EX| \geq \varepsilon) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}}$

With high probability (WHP) deviation from mean is of order $\sqrt{\mathrm{Var}(X)}$

Pf: Apply Markov to $Y = (X-EX)^2$. $\square$

Law of large numbers: Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} P$ with $\begin{cases} \text{mean } \mu, \\ \text{variance } \sigma^2. \end{cases}$

Let $\overline{X} := \frac{1}{n}(X_1 + \dots + X_n)$. Then:

$$\boxed{\Pr(|\overline{X} - \mu| \geq \varepsilon) \leq \frac{1}{n} \frac{\sigma^2}{\varepsilon^2}}$$

WHP: empirical average $\approx$ expectation value

Pf: $E\overline{X} = \mu$ & $\mathrm{Var}(\overline{X}) = \frac{1}{n^2} \mathrm{Var}(X_1 + \dots + X_n) = \frac{\sigma^2}{n}$. $\rightsquigarrow$ Chebyshev. $\square$

Convex and Concave functions (§2.7)

Suppose $f: I \to \mathbb{R}$ is function on interval $I = (a,b)$ $a=-\infty$ or $b=\infty$ allowed

We say f is **Convex** if $f'' \geq 0$ exp, x^2, ...

Concave if $f'' \leq 0$ log, $\sqrt{x}$, ...

Jensen's inequality: Let Z be a RV.

If f is convex:	$E[f(Z)] \geq f(EZ)$
If f is concave:	$E[f(Z)] \leq f(EZ)$

i.e. $\sum_z P(z) f(z) \overset{\geq}{\leq} f\left(\sum_z P(z) z\right)$

If $f'' > 0$ or $f'' < 0$: "=" holds only if Z is constant!

Entropy (§2.4)

Entropy of a random variable (RV) X with distribution P:

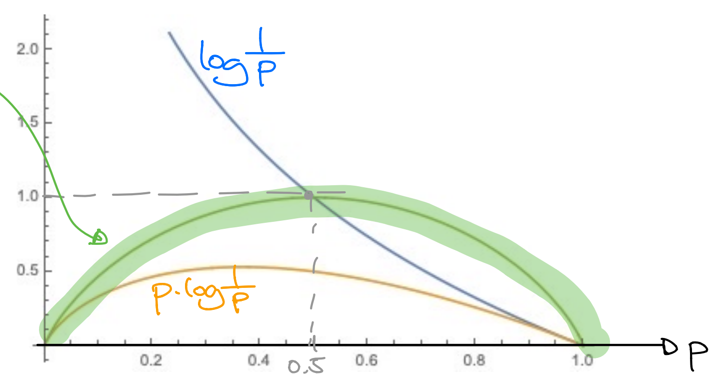
$$H(X) := H(P) := \sum_x P(x) \cdot \log \frac{1}{P(x)} = E\left[\log \frac{1}{P(X)}\right]$$

$0 \cdot \log \frac{1}{0} \equiv 0$ always base 2   Unit "bit"

e.g. $X \sim \text{Bernoulli}(p)$: **binary entropy**

$$H(X) = p \cdot \log \frac{1}{p} + (1-p) \cdot \log \frac{1}{1-p}$$

$\in [0,1]$



Properties:

* $H(X) \geq 0$, = iff constant $p \cdot \log \frac{1}{p} \geq 0 \quad \forall p \in [0,1]$, = iff $p=0$ or $p=1$

* $H(X) \leq \log \#\{x : P(x) > 0\} \leq \log \#\mathcal{A}_X$

  $H(X) = \log \#\mathcal{A}_X \iff X$ uniformly random

Pf: Apply Jensen with $f = \log$ and $Z = \frac{1}{P(X)}$:

$$E\left[\log \frac{1}{P(X)}\right] \leq \log E\left[\frac{1}{P(X)}\right]$$

with equality iff $P(X)$ constant, i.e. $P(x)>0, P(y)>0 \Rightarrow P(x) = P(y)$ □

* **NOTATION**: $H(X,Y) = H(XY)$ = entropy of joint distribution $P(x,y)$

If X,Y **independent**: $H(X,Y) = H(X) + H(Y)$

Pf: Since $P(x,y) = P(x)P(y)$ we have $\log \frac{1}{P(x,y)} = \log \frac{1}{P(x)} + \log \frac{1}{P(y)}$
⇝ take expectation values. □

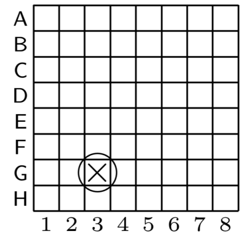
Interpretation? Let us call $h(x) = h(X=x) = \log_2 \frac{1}{P(x)}$ the **information content** (or "surprisal") of an outcome $x \in \mathcal{A}_X$.

$\Rightarrow H(X) = E[h(X)]$ is average information content.

Why is this a good definition? Three suggestive examples:

① **Uniformly random number** in $\{0,\dots,255\}$: $H(X) = \log_2 256 = 8$ bit

② **Poor man's submarine game**: Single submarine hidden, other player asks if submarine in some square → hit/miss



1st move: $P(\text{hit}) = \frac{1}{64} \to h(\text{hit}) = 6$ bit   learned precise location (64 options)

we skipped this

$P(\text{miss}) = \frac{63}{64} \to h(\text{miss}) \approx 0{,}0227$ bit   learned little (63 remaining)

2nd move (if 1st missed): $P(\text{miss}) = \frac{62}{63} \to h(\text{miss}) \approx 0{,}0230$ bit

after 32 misses: $\sum h(\text{miss}) = \log \frac{64}{63} + \dots + \log \frac{33}{32} = \log \frac{64}{32} = 1$ bit   localized to 1/2 of squares

after 48 misses: $\sum h(\text{miss}) = \log \frac{64}{16} = 2$ bit   localized to 1/4 of the squares

hit in 49th round: $h(\text{hit}) = \log \frac{16}{1} = 4$ bit $\to \sum = 6$ bit $= H(\text{position})$

More generally: If we hit when n squares remaining

$$\sum h(\text{miss}) + h(\text{hit}) = \log \frac{64}{63} + \dots + \log \frac{n+1}{n} + \log \frac{n}{1} = \log 64 = 6 \text{ bit}$$

③ "**Wenglish**" has 2^{15} words in $\{A,\dots,Z\}^5$ s.th. frequence of single letters matches English. Let W be uniformly random word in this list.

$H(W) = 15$ bit, i.e. on average 3 bit/letter

but e.g. $p(W_1 = Z) = 0.1\% \Rightarrow h(W_1 = Z) \approx 10$ bit ← no contradiction; we learn less info from the rest since few words start with Z

$\ll \frac{1}{26}$

Compression and Symbol Codes (§5)

Consider data source modeled by RV X. Assume we know distribution P_X.

E.g. X could be a letter and we assume $P(x) = P_{english}(x)$

How well can we compress?

Today + on Thursday we consider symbol codes, which compress one symbol (letter, source message, ...) at a time:

WANT:

$$X \longrightarrow \boxed{\;C\;} \longrightarrow \{0,1\}^+ \longrightarrow \boxed{\;D\;} \longrightarrow \hat{X} = X$$

one or more bits

compression — decompression

> GOAL: Show that lossless compression one symbol at a time can achieve $H(X) \le L < H(X) + 1$, where L = average length of codeword.

at worst one more bit than entropy

NOTATION: $S^+ = \bigcup_{N \geq 1} S^N$ = nonempty strings over S

$\ell(w)$ = length of string $w \in S^+$

Symbol code: $C: \mathcal{A} \to \{0,1\}^+$ for alphabet $\mathcal{A}$

* average length: $L(C,P) = L(C,X) = \sum_{x \in \mathcal{A}} P(x)\, \ell(C(x)) = E\left[\ell(C(X))\right]$

  want to minimize

* extended code: $C^+: \mathcal{A}^+ \to \{0,1\}^+$, $C^+(x_1 \cdots x_N) := C(x_1) \cdots C(x_N)$

Two important classes of codes:

* C is called uniquely decodable (UD) if

  $C^+(w) = C^+(w') \Rightarrow w = w' \quad \forall w, w' \in \mathcal{A}^+$ — we really want this!

* C is called a prefix code if no codeword $C(x)$ is prefix of any other

Any prefix code is UD!