"Numerical" random variables

If $X \sim P$ is RV with values in $\mathcal{A} \subseteq \mathbb{R}$:

Expectation value (mean): $EX = E[X] = \sum_x P(x) \cdot x$

* $E[f(X)] = \sum_x P(x) \cdot f(x)$ "law of the unconscious statistician"

* $E[cX] = c \cdot E[X]$ & $E[X+Y] = E[X] + E[Y]$ (A)

* If X, Y independent: $E[XY] = E[X] \cdot E[Y]$ $\sum_{x,y} p(x) p(y)\, xy$

  $\lightning$ $X \sim \text{Uniform}(\{\pm 1\})$, $Y = -X$ $\Rightarrow E[XY] = -1$, $E[X] = E[Y] = 0$

  NOT indep

Examples

P	Bernoulli(f)	Binomial(n,f)
E	f	$n \cdot f$ (A)
Var	$f(1-f)$	$n \cdot f \cdot (1-f)$ (B)

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

Variance: $\mathrm{Var}(X) = E[(X-EX)^2]$
$= \sum_x P(x)(x-EX)^2 = E[X^2] - E[X]^2$

* $\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$

* If X, Y independent:

  $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ (B)

  use that $E[XY] = E[X] \cdot E[Y]$

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

$$\boxed{\begin{array}{ll} \text{If } f \text{ is convex:} & E[f(Z)] \geq f(EZ) \\ \text{If } f \text{ is concave:} & E[f(Z)] \leq f(EZ) \end{array}}$$

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

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:

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

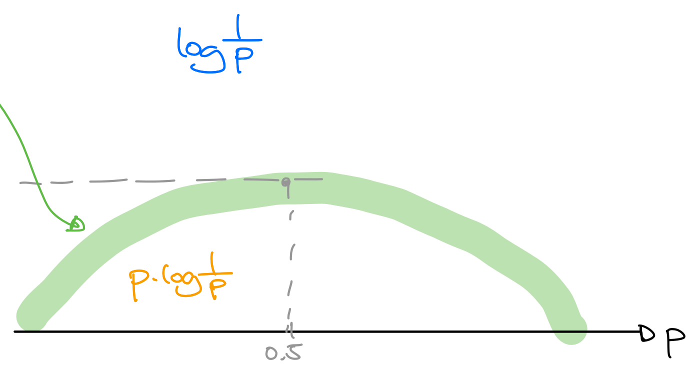
unit "bit"

$0 \cdot \log \frac{1}{0} = 0$ always base 2

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 $\quad p \cdot \log\frac{1}{p} \geq 0 \;\; \forall p \in [0,1]$, $=$ 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. $P(x)>0, P(y)>0 \Rightarrow P(x) = P(y)$ $\square$

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

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

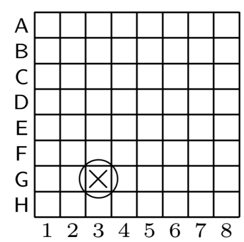
  Pf: Since $P(x,y) = P(x)P(y)$ we have $\log\frac{1}{P(x,y)} = \log\frac{1}{P(x)} + \log\frac{1}{P(y)}$
  $\rightsquigarrow$ take expectation values. $\square$

$\boxed{\text{Interpretation?}}$ 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 $\{0, \ldots, 255\}$: $H(X) = \log_2 256 = 8$ bit



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

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

$P(\text{miss}) = \frac{63}{64} \longrightarrow h(\text{miss}) \approx 0.0227$ bit $\quad$ learned little (63 remaining)

③ "Wenglish" has 2^{15} words in $\{A, \ldots, Z\}^5$ 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. $p(W_1 = Z) = \underset{\ll \frac{1}{26}}{0.1\%} \Rightarrow h(W_1 = Z) \approx 10$ bit $\leftarrow$ no contradiction; we learn less info from the rest since few words start with Z