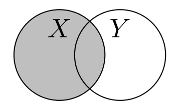
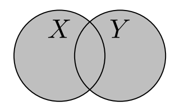
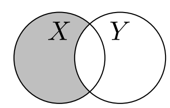
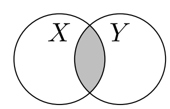
Some measures of information

Xun Zheng
xzheng1@andrew.cmu.edu

1 Some measures of information

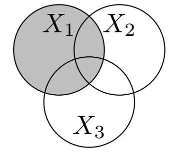
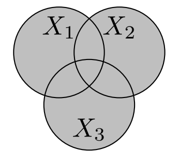
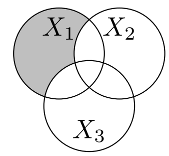
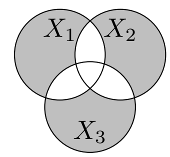
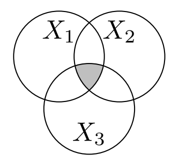
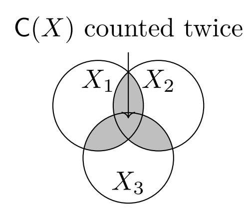
1.1 Bivariate case

We have in mind two random variables (X, Y) with joint distribution $P(X, Y)$.

quantity	definition	comment	i-diagram
entropy	$\mathsf{H}(X)$ $= \mathsf{E}_{x \sim P(X)}\left[\log \frac{1}{P(x)}\right]$		
joint entropy	$\mathsf{H}(X, Y)$ $= \mathsf{E}_{x,y \sim P(X,Y)}\left[\log \frac{1}{P(x,y)}\right]$		
conditional entropy (given $Y = y$)	$\mathsf{H}(X\|Y = y)$ $= \mathsf{E}_{x \sim P(X\|Y=y)}\left[\log \frac{1}{P(x\|y)}\right]$	uncertainty of X knowing $Y = y$	
conditional entropy (average)	$\mathsf{H}(X\|Y)$ $= \mathsf{E}_{y \sim P(Y)}\big[\mathsf{H}(X\|Y = y)\big]$ $= \mathsf{E}_{x,y \sim P(X,Y)}\left[\log \frac{1}{P(x\|y)}\right]$ $= \mathsf{H}(X, Y) - \mathsf{H}(Y)$	average (w.r.t. $P(Y)$) uncertainty of X when the value of Y is known; also from chain rule of entropy	
mutual information	$\mathsf{I}(X; Y)$ $= \mathsf{H}(X) - \mathsf{H}(X\|Y)$ $= \mathsf{H}(X) + \mathsf{H}(Y) - \mathsf{H}(X, Y)$ $= \mathsf{D}\big(P(X, Y) \\| P(X)P(Y)\big)$ $= \mathsf{E}_{y \sim P(Y)}\mathsf{D}\big(P(X\|Y = y) \\| P(X)\big)$	average (w.r.t. $P(Y)$) **reduction** in uncertainty of X when the value of Y is known, or vice versa	

1.2 Multivariate case

Consider a collection of random variables $X = (X_1, \ldots, X_d)$ with joint distribution $P(X)$.

quantity	definition	comment	i-diagram
marginal entropy	$\mathsf{H}(X_j)$ $= \mathsf{E}_{x \sim P(X_j)}\left[\log \frac{1}{P(x_j)}\right]$		
joint entropy	$\mathsf{H}(X)$ $= \mathsf{E}_{x \sim P(X)}\left[\log \frac{1}{P(x)}\right]$ $= \sum_{j=1}^{d} \mathsf{H}(X_j \vert X_{<j})$	generalized chain rule	
conditional entropy	$\mathsf{H}(X_j \vert X_{-j})$ $= \mathsf{E}_{x \sim P(X)}\left[\log \frac{1}{P(x_j \vert x_{-j})}\right]$		
residual entropy	$\mathsf{R}(X)$ $= \sum_{j=1}^{d} \mathsf{H}(X_j \vert X_{-j})$	erasure entropy	
co-information	$\mathsf{C}(X)$ $= \sum_{S \subseteq [d]} -(-1)^{\vert S \vert} \mathsf{H}(X_S)$	amount of information all variables participate in; can be negative	
total correlation	$\mathsf{TC}(X)$ $= -\mathsf{H}(X) + \sum_{j=1}^{d} \mathsf{H}(X_j)$ $= \mathsf{D}(P(X) \Vert P(X_1) \cdots P(X_d))$ $= \mathsf{TC}(X_S) + \mathsf{TC}(X_{S^c}) + \mathsf{I}(X_S; X_{S^c})$ $= \sum_{j=2}^{d} \mathsf{I}(X_j; X_{<j})$	multi-information; non-negative, zero iff X mutually independent; read i-diagram with caution since $\mathsf{C}(X)$ can be negative	
dual total correlation	$\mathsf{DTC}(X)$ $= \mathsf{H}(X) - \mathsf{R}(X)$	binding information; non-negative, zero iff X mutually independent