■ Common formulae of information theory

Hugo Touchette (htouchet@mit.edu) Updated: August 31, 2004

Taken from T. Cover, J. Thomas, *Elements of Information Theory*, Wiley, 1991.

Entropy. $H(X) = -\sum_x p(x) \log p(x)$

$H(X) \geq 0$
$H_b(X) = \log_b a H_a(X)$
$H(X) \leq \log |\mathcal{X}|$
$d_\alpha \sum_x p(x)^\alpha \big|_{\alpha=1} = H(X)$

$\log_a x = \log_a b \log_b x = \dfrac{\log_b x}{\log_b a}$

$(a^x)' = a^x \ln a, \qquad (\log_a x)' = \dfrac{1}{x \ln a} = \dfrac{\log_a e}{x}$

Joint entropy. $H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y)$

$H(X_1, \ldots, X_n) \leq \sum_{i=1}^n H(X_i)$ eq. iff X_i ind.

Conditional entropy. $H(Y|X) = \sum_x H(Y|x) p(x) = -\sum_{x,y} p(x,y) \log p(y,x)$

$H(Y|X) \geq 0$ eq. if $Y = f(X)$, or iff $Y|x$ is deterministic for all $x \in \mathrm{supp}(X)$
$H(Y|X) \leq H(Y)$ eq. iff X, Y ind.

Relative entropy. $D(p||q) = \sum_x p(x) \log \dfrac{p(x)}{q(x)}$

$D(p||q) \geq 0$
$D(p||u) = \log |\mathcal{X}| - H(X),\ u(x) = |\mathcal{X}|^{-1}$

Mutual information. $I(X;Y) = \sum_{x,y} p(x,y) \log \dfrac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y))$

$I(X;Y) \geq 0$ eq. iff X, Y ind.
$I(X;Y) = I(Y;X)$
$I(X;X) = H(X)$
$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
$I(X;Y) = H(X) + H(Y) - H(X,Y)$
$\max I(X;Y) = \min(H(X), H(Y))$

Information metric. $\Delta(X,Y) = H(X|Y) + H(Y|X)$
$\Delta(X,Y) = H(X) + H(Y) - 2I(X;Y) = H(X,Y) - I(X,Y)$

Chain rules.

$$
\begin{aligned}
H(X,Y) &= H(X) + H(Y|X) \\
H(X,Y|Z) &= H(X|Z) + H(Y|X,Z) \\
H(X_1, \ldots, X_n) &= \sum_{i=1}^n H(X_i | X_{i-1}, \ldots, X_1)
\end{aligned}
$$

$$
\begin{aligned}
I(X_1,\ldots,X_n;Y) &= \sum_{i=1}^{n} I(X_i;Y|X_{i-1},X_{i-2},\ldots,X_1)\\
D(p(x,y)||q(x,y)) &= D(p(x)||q(x)) + D(p(y|x)||q(y|x))
\end{aligned}
$$

Conditioning.

$D(p(y\|x)\|\|q(y\|x)) = \sum_x p(x) \sum_y p(y\|x) \log \frac{p(y\|x)}{q(y\|x)}$	Conditional relative entropy
$I(X;Y\|Z) = H(X\|Z) - H(X\|Y,Z)$	Conditional mutual information

Convexity properties.

$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ or $f''(x) \geq 0$	Convex function
$E[f(X)] \geq f(E[X])$	Jensen's inequality
$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$	log sum inequality

Differential entropy. $h(X) = -\int p(x) \log p(x)\ dx$

$h(X+a) = h(X)$	$H(X^{\Delta}) + \log \Delta \to h(X)$
$h(AX) = h(X) + \log \|A\|$	$D(P\|\|Q) \geq 0$
$h(X) \leq \log \operatorname{supp} X$	$D(P\|\|Q) = \sup_{\Delta} D(P^{\Delta}\|\|Q^{\Delta})$
$h(X) = \frac{1}{2}\log 2\pi e \sigma^2$ when $X \sim N(m,\sigma^2)$	

Correlations and causation.

$A \to B \leftarrow C$	$I(A;C) = 0$
	$I(A;C\|B) \neq 0$ in general
$A \leftarrow B \to C$ or	$I(A;C) \neq 0$ in general
$A \to B \to C$	$I(A;C\|B) = 0$ (bottleneck)

Chains of random variables.

$X \to Y \to Z$	$I(X;Y) \geq I(X;Z)$
	$I(Z;Y) \geq I(Z;X)$ eq. iff $I(X;Y\|Z) = 0$
	$I(X;Y) \geq I(X;g(Y))$
	$I(X;Y\|Z) \leq I(X;Y)$
$X^{(n)} \to Y^{(k)} \to Z^{(m)}$	$I(X;Z) \leq \log k$
$k < n, k < m$	$I(X;Z) = 0$ if $k = 1$

Asymptotic equipartition theorem.

$X_1 X_2 \ldots X_n \sim p(x)$ iid.

$$
\begin{aligned}
-\frac{1}{n}\log p(x_1,x_2,\ldots,x_n) &= -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i)\\
&\to -E[\log p(x_i)] = H(X) \text{ (in probability)}
\end{aligned}
$$

$A_\varepsilon^n = \{x^n \in \mathcal{X}^n : 2^{-n(H+\varepsilon)} \leq p(x^n) \leq 2^{-n(H-\varepsilon)}\}$	Typical set
$-\frac{1}{n}\log p(x^n) = H(X)$ (within ε)	
$\Pr\{A_\varepsilon^n\} > 1-\varepsilon$	(from above result)
$\|A_\varepsilon^n\| \leq 2^{n(H+\varepsilon)}$	$\|A_\varepsilon^n\| \doteq 2^{nH}$ (within ε exponentially)
$\|A_\varepsilon^n\| \geq (1-\varepsilon)2^{n(H-\varepsilon)}$	$p(x^n) \doteq 2^{-nH}$

Method of types.

$X_1X_2\dots X_n,\ x_1x_2\dots x_n = x^n = \boldsymbol{x} \in \mathcal{X}^n$	$\mathcal{X} = \{a_1, a_2, \dots, a_{\lvert\mathcal{X}\rvert}\}$	
$P_{\boldsymbol{x}}(a) = N(a\lvert\boldsymbol{x})/n$	$\sum_a P_{\boldsymbol{x}}(a) = 1$	Type
$\mathcal{P}_n = \{P_{\boldsymbol{x}} : \lvert\boldsymbol{x}\rvert = n\}$		Set of types n
$T(P) = \{\boldsymbol{x} \in X^n : P_{\boldsymbol{x}} = P\}$		Type class
$\lvert\mathcal{P}_n\rvert \le (n+1)^{\lvert\mathcal{X}\rvert}$	$\lvert\mathcal{X}^n\rvert \sim \lvert\mathcal{X}\rvert^n$ $\lvert\mathcal{P}_n\rvert \sim n^{\lvert\mathcal{X}\rvert}$	
$Q^n(\boldsymbol{x}) = 2^{-n[H(P_x)+D(P_x\|\|Q)]}$	$X_1X_2\dots X_n \sim$ iid $Q(x)$	
$Q^n(\boldsymbol{x}) = 2^{-nH(Q)}$	$\boldsymbol{x} \in T(Q)$	Type Q seqs.
$\frac{1}{(n+1)^{\lvert\mathcal{X}\rvert}} 2^{nH(P)} \le \lvert T(P)\rvert \le 2^{nH(P)}$	$\lvert T(P)\rvert \doteq 2^{nH(P)}$	Type class size
$\frac{1}{(n+1)^{\lvert\mathcal{X}\rvert}} 2^{-nD(P\|\|Q)} \le Q^n(T(P)) \le 2^{-nD(P\|\|Q)}$	$Q^n(T(P)) \doteq 2^{-nD(P\|\|Q)}$	Type class prob.
$\Pr\{D(P_{\boldsymbol{x}}\|\|Q) > \varepsilon\} \le 2^{-n[\varepsilon - \lvert\mathcal{X}\rvert \frac{\log n+1}{n}]}$	$X_1X_2\dots X_n \sim$ iid $Q(x)$	WLLN
$\Pr\{D(P_{\boldsymbol{x}}\|\|Q) > \varepsilon\} \le n^{\lvert\mathcal{X}\rvert} 2^{-n\varepsilon} \sim 2^{-n\varepsilon}$	$D(P_{\boldsymbol{x}}\|\|Q) \to 0$ (ip)	
$Q^n(E) \le (n+1)^{\lvert\mathcal{X}\rvert} 2^{-nD(P^*\|\|Q)}$	$X_1X_2\dots X_n \sim$ iid $Q(x)$	Sanov
$P^* = \arg\min_{P\in E} D(P\|\|Q)$	$P \subseteq \mathcal{P}$	

Rate distortion theory.

$$X^n \to f(X^n) \longrightarrow g(f(X^n)) \to \hat{X}^n$$

$d : \mathcal{X} \times \widehat{\mathcal{X}} \to \mathbb{R}^+$	$d(x^n, \hat{x}^n) = \frac{1}{n}\sum_i d(x_i, \hat{x}_i)$ $\max d(x, \hat{x}) < \infty$	Distance, distortion
$d = \{\begin{smallmatrix}0, & x=\hat{x}\\ 1, & x\neq\hat{x}\end{smallmatrix}$	$E[d(X,\hat{X})] = \Pr\{X \neq \hat{X}\}$	Hamming distance
$R(D) = \min_{p(x\lvert x): E[d(X^n,\hat{X}^n)] \le D} I(X;\hat{X})$	$f(X^n) \in \{1, 2, \dots, 2^{nR}\}$	Rate (# bits needed)
$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & 0 \le D \le \sigma^2 \\ 0, & D > \sigma^2 \end{cases}$		Gaussian channel

Elements of probability theory.

$X_i \sim$ iid, $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{\text{ip}} E[X]$	WLLN
$Y = g(X),\ X = h(Y) = g^{-1}(Y)$	$f_Y(y) = f_X(h(y))\,\lvert h'(y)\rvert = f_X(x)\left\lvert\frac{\partial h(y)}{\partial y}\right\rvert$
$Y = g(X) = \alpha X + \beta$	$\mathcal{N}(\mu, \sigma^2) \xrightarrow{g} \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$
$Y = X + Z$	$X \sim \mathcal{N}(x, \sigma_X^2)$ $Z \sim \mathcal{N}(z, \sigma_Z^2)$ $Y \sim \mathcal{N}(x + z, \sigma_X^2 + \sigma_Z^2)$