14-08-2017.

lec 3: Convex functions, Inequalities.

Recap

- Entropy
- Properties
  - $H(X) \geq 0$.
  - $H_b(X)$
  - $H(p) \to$ Bernoulli r v
  - example (guessing)
  - (shows up in channel capacity of Binary symmetric channel, Binary erasure channel)
- Joint & Conditional entropy $H(X,Y)$ and $H(X|Y)$, $H(Y|X)$.
  - example
- Relative Entropy $D(p\|q)$
- Mutual information $I(X;Y)$

$$-\sum_x p(x) \sum_y (\log p(y|x))\, p(y|x)$$

Today

- Chain Rule $= H(X_1, \ldots X_n)$
- Conditional mutual information
- Chain rule for mutual information
- conditional relative entropy.
- chain rule for relative entropy
- convex fns
  - $f''(\cdot) \geq 0$ test
- Jensen's inequality
- $D(p\|q) \geq 0$.
  - $I(X;Y) \geq 0$.
- $D(p(y|x) \| q(y|x)) \geq 0$
  - $I(X;Y|Z) \geq 0$.
- $H(X) \leq \log|\mathcal{X}|$
- $H(X_1, \ldots X_n) \leq \sum_{i=1}^{n} H(X_i)$

Theorem (Chain rule for entropy) (Analogue of grouping axiom)

$$H(X_1, \ldots X_n) = \sum_{i=1}^{n} H(X_i | X_1, \ldots X_{i-1})$$

Pf:

$$= -\sum_{\underline{x}} p(\underline{x}) \log p(x_1) \prod_{i=2}^{n} p(x_i | x_{[i-1]})$$

$$\left.\begin{aligned} X_{[\ell]} &= X_1, \ldots X_\ell \\ x_{[\ell]} &= x_1, \ldots x_\ell \\ \underline{x} &= x_1, \ldots x_n \end{aligned}\right\} \text{NOTATION.}$$

$$= -\sum_{\underline{x}} p(\underline{x}) \log p(x_1) - \sum_{\underline{x}} \sum_{i=2}^{n} \left(\log p(x_i | x_{[i-1]})\right) p(\underline{x})$$

$$= H(X_1) + \sum_{i=2}^{n} H(X_i | X_{[i-1]})$$

Conditional Mutual Information (MI)

Defn:

$$I(X;Y|Z) \triangleq H(X|Z) - H(X|Y,Z)$$

thm (chain rule for conditional M.I)

$$I(X_1, X_2, \ldots X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{[i-1]})$$

Pf: $I(X_{[n]};Y) = H(X_{[n]}) - H(X_{[n]}|Y)$

$$= \sum_{i=1}^{n} H(X_i|X_{[i-1]}) - \sum_{i=1}^{n} H(X_i|X_{[i-1]},Y)$$

$$= \sum_{i=1}^{n} I(X_i;Y|X_{[i-1]}).$$

Conditional relative entropy

$$D(p(y|x)\,\|\,q(y|x)) \triangleq \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)}.$$

Note $p(x)$ not explicitly mentioned on LHS.
Can be seen to equal

$$E_{p(x,y)}\left\{ \log \frac{p(Y|X)}{q(Y|X)} \right\}$$

Thm (Chain rule for relative Entropy)

$$D(p(x,y)\,\|\,q(x,y)) = D(p(x)\,\|\,q(x)) + D(p(y|x)\,\|\,q(y|x))$$

Proof:

$$\text{LHS} = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x)\,p(y|x)}{q(x)\,q(y|x)}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x)}{q(x)} + \sum_{x,y} p(x)\,p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

$$= D(p(x)\,\|\,q(x)) + D(p(y|x)\,\|\,q(y|x))$$

Convex functions

Defn: A function $f(x)$ is said to be convex over interval (a,b) if for every $(x_1, x_2) \subseteq (a,b)$ and $0 \le \lambda \le 1$, we have

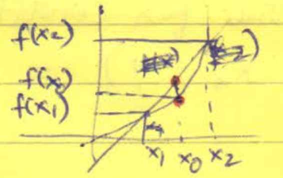
$$f(x_1\lambda + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2).$$

$f(\cdot)$ is said to be strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$.

Note

$$\overbrace{\underbrace{x_1 \quad\quad}_{(1-\lambda)(x_2-x_1)} x_0 \underbrace{\quad\quad x_2}_{\lambda(x_2-x_1)}}^{x_2 - x_1}$$

$x_0 = \lambda x_1 + (1-\lambda)x_2$.



$$\lambda = \frac{x_2 - x_0}{x_2 - x_1}$$

Thm: If $f(\cdot)$ has a 2nd derivative $f''(\cdot)$ that is non-negative (positive) over an interval, then $f(\cdot)$ is (strictly convex) convex.

Pf:
$$f(x_1) = f(x_0) + (x_1 - x_0) f'(x_0) + \frac{(x_1-x_0)^2}{2!} f''(\varepsilon_1)$$
$$f(x_2) = f(x_0) + (x_2 - x_0) f'(x_0) + \frac{(x_2-x_0)^2}{2!} f''(\varepsilon_1')$$

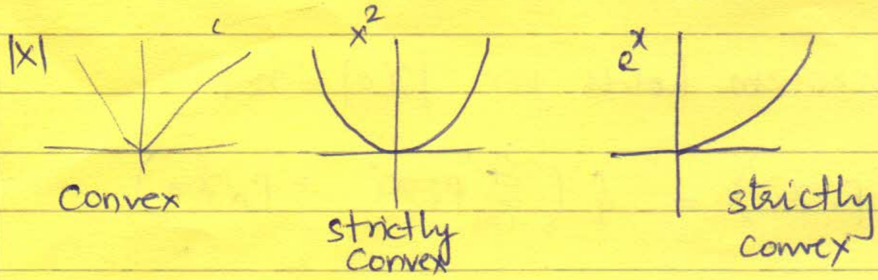
$\varepsilon_1 \in (x_0, x_1)$

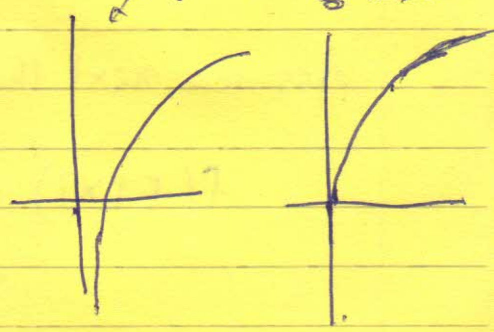
$\varepsilon_1' \in (x_0, x_2)$

$\varepsilon_1, \varepsilon_1'$ some points in (x_0, x_1), (x_0, x_2) respectively.

Eg: Convex functions: x^2, $|x|$, e^x, $x \log x$.

Eg: concave functions: $\log x$, $\sqrt{x}$ $x \geq 0$.





Continuation of proof:

The expantion is given by Taylor's theorem (see Rudin) given by Mean Value theorem

Now by substituting $x_1 - x_0 = -(1-\lambda)(x_2 - x_1)$ and $x_2 - x_0 = \lambda(x_2 - x_1)$ respectively we get and using the fact that $f''(\varepsilon_1), f''(\varepsilon_1') \geq 0$

$$f(x_1) \geq f(x_0) - (1-\lambda)(x_2 - x_1) f'(x_0)$$
$$f(x_2) \geq f(x_0) + \lambda(x_2 - x_1) f'(x_0).$$

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(x_0) = f(\lambda x_1 + (1-\lambda) x_2).$$

Thm (Jensen's Inequality): If f is a convex function and X is a R.V, $E[f(X)] \geq f(E[X])$. — ①

Moreover if f is strictly convex, equality in ① $\rightarrow$ X is a constant.

Note: $E[X] = \sum_{x \in \mathcal{X}} x\, p(x)$

$$E[f(X)] = \sum_{x \in \mathcal{X}} f(x) p(x)$$

Pf: Will assume that X is a discrete RV taking on values from $\mathcal{X}_0$ and that $|\mathcal{X}_0| = n < \infty$

We will induct on the size of $\mathcal{X}_0$. If $|\mathcal{X}_0| = 2$, then ① corresponds to saying that

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

$$\Downarrow$$

$$f(E[X]) \leq E[f(X)].$$

Assume next that the Theorem holds for $|\mathcal{X}_0| = n-1$.

$$f(E[X]) = f\left(\sum_{i=1}^{n} p_i x_i\right) = f\left(\sum_{i=1}^{n-1} p_i x_i + p_n x_n\right).$$

$$= f\left((1-p_n) \sum_{i=1}^{n-1} \left(\frac{p_i}{1-p_n}\right) x_i + p_n x_n\right)$$

let $x_0 = \sum_{i=1}^{n-1} \left(\frac{p_i}{1-p_n}\right) x_i$ (a weighted average of $\{x_i\}_{i=1}^{n-1}$).

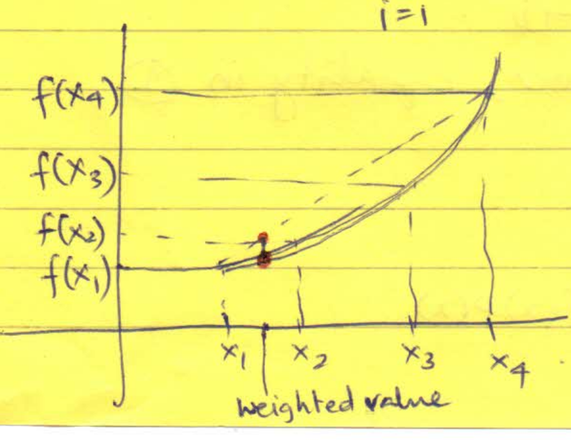
$$f(E[X]) = f((1-p_n)x_0 + p_n x_n)$$

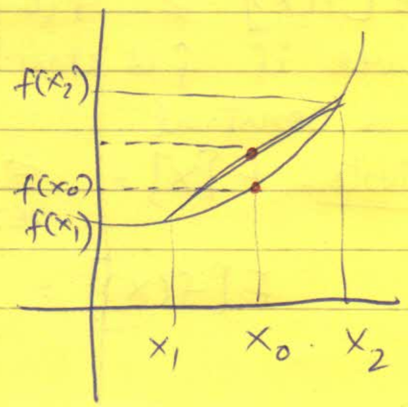
$$\leq (1-p_n) f(x_0) + p_n f(x_n) \rightarrow ①$$

$$= (1-p_n) f\left(\sum_{i=1}^{n-1} \left(\frac{p_i}{1-p_n}\right) x_i\right) + p_n f(x_n)$$

$$\leq (1-p_n) \sum_{i=1}^{n-1} \left(\frac{p_i}{1-p_n}\right) f(x_i) + p_n f(x_n) \rightarrow ②$$

$$= \sum_{i=1}^{n} p_i f(x_i) = E[f(X)].$$





① should also be an
Assume that this is an equality, ~~the~~ equality ~~is~~
if $f(\cdot)$ is strictly convex, the equality holds in
$$f(E[X]) \le (1-p_n) f(x_0) + p_n f(x_n)$$
only if $p_n = 0$ or $p_n = 1$.
If $p_n = 1$ then $X = \{ x_n$ w.p 1 $\therefore$ X is a constant.
Otherwise. if $p_n = 0$.
$$f\left(\sum_{i=1}^{n-1} p_i x_i\right) \le \sum_{i=1}^{n-1} p_i f(x_i).$$
the equality above holds if X is constant (by induction on $|\mathcal{X}_0|$).