Is Information Intrinsic?

- Discrete entropy is intrinsic
- Differential entropy is extrinsic

- Differences in entropy are scale - independent, but extrinsic

- Information is the expected difference in entropy before and after an observation... is information intrinsic?

- recall:

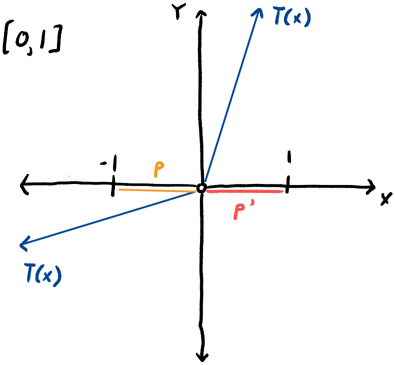
1. Scaling: if $Y = AX + b$ then $h[Y] = h[X] + \log(|\det(A)|)$
   for any invertible A
2. $\Delta h \perp\!\!\!\perp$ to scaling: if $Y = AX + b$, $Y' = AX' + b$, $X \sim p$, $X' \sim p'$
   then:
   $$h[Y] - h[Y'] = h[X] - h[X'] + \underbrace{\log(|\det(A)|) - \log(|\det(A)|)}_{\text{cancel !}}$$
   $$= \underbrace{h[X] - h[X']}_{\text{does not depend on } A \text{ or } b}$$
3. generic transform: $Y = T(X)$ for T invertible, differentiable (T is a diffeomorphism)
   - let $\frac{d}{dx} T(x) =$ Jacobian of T at x
   - then (change of density): $f_Y(y) = f_X(x) \, |\det(\frac{d}{dx} T(x))|^{-1}$ at $x = T^{-1}(y)$ ($y = T(x)$)
   - so:
   $$h[Y] = -\mathbb{E}_Y[\log(f_Y(Y))] = -\mathbb{E}_X[\log(f_X(X) \, |\det(\ldots)|^{-1})] = -\mathbb{E}_X[\log(f_X(X))] - \mathbb{E}_X[\log(|\det(\ldots)|^{-1})]$$
   $$= h[X] - \mathbb{E}_X[\log(|\det(\ldots)|^{-1})] = h[X] + \mathbb{E}_X[\log(|\det(\tfrac{d}{dx} T(x))|)]$$
   - that is:
   $$h[Y] = h[T(X)] = h[X] + \underbrace{\mathbb{E}_X[\log(|\det(\tfrac{d}{dx} T(x))|)]}_{\text{expected expansion factor}}$$
4. Δh is extrinsic: $Y = T(X)$, $Y' = T(X')$, $X \sim p$, $X' \sim p'$

   then: $$h[Y] - h[Y'] = h[X] - h[X'] + \underbrace{\mathbb{E}_{X \sim p}[\log(|\det(\tfrac{d}{dx} T(X))|)] - \mathbb{E}_{X' \sim p'}[\log(|\det(\tfrac{d}{dx} T(X'))|)]}_{\text{average expansion factor over } p \text{ and } p'}$$

   $p \neq p'$ so averages may differ...

- Ex: $\mathcal{X} = [-1, 1]$, $p = \text{Uniform}[-1, 0]$, $p' = \text{Uniform}[0, 1]$

$$T(x) = \begin{cases} 3x & \text{if } x > 0 \\ \frac{1}{3}x & \text{if } x < 0 \end{cases}$$



$$\frac{d}{dx} T(x) = \begin{cases} 3 & \text{if } x > 0 \\ 1/3 & \text{if } x < 0 \end{cases}$$

so
$$\left.\begin{aligned} \mathbb{E}_{x \sim p}\left[\log(|\det(\tfrac{d}{dx}T(x))|)\right] &= \log(1/3) \\ \mathbb{E}_{x' \sim p'}\left[\quad \dots \quad\right] &= \log(3) \end{aligned}\right\} \text{difference is } \log(1/3) - \log(3) = -\log(9) \neq 0.$$

- Information is a difference in entropies so it must be scale independent, is it extrinsic or intrinsic?

- Consider: $Y = T(X)$, $X \sim p_X$

  observe Z, $X|Z=z \sim p_{X|Z=z}$

Then:
$$\begin{aligned}
I[Y;Z] &= h[Y] - h[Y|Z] \\
&= h[Y] - \mathbb{E}_Z\left[h[Y|Z=z]\right] \\
&= h[X] + \mathbb{E}_X\left[\log(|\det(\tfrac{d}{dx}T(x))|)\right] - \mathbb{E}_Z\left[h[X|Z=z] + \mathbb{E}_{X|Z=z}\left[\log(|\det(\tfrac{d}{dx}T(x))|)\right]\right] \\
&= \quad \dots \quad - h[X|Z] - \underbrace{\underbrace{\mathbb{E}_Z\Big[\mathbb{E}_{X|Z=z}}_{\mathbb{E}_{X,Z} \text{ jointly}}\underbrace{\left[\log(|\det(\tfrac{d}{dx}T(x))|)\right]\Big]}_{\perp\!\!\!\perp \text{ of } Z}}_{= \mathbb{E}_X} \\
&= h[X] + \mathbb{E}_X\left[\log(|\det(\tfrac{d}{dx}T(x))|)\right] - h[X|Z] - \mathbb{E}_X\left[\log(|\det(\tfrac{d}{dx}T(x))|)\right] \\
&= I[X;Z] + \underbrace{\mathbb{E}_X\left[\log(|\det(\tfrac{d}{dx}T(x))|)\right] - \mathbb{E}_X\left[\log(|\det(\tfrac{d}{dx}T(x))|)\right]}_{\text{Cancel!!}} \\
&= I[X;Z]
\end{aligned}$$

so, if T is a diffeomorphism (invertible, differentiable)

then $I[Y;Z] = I[T(X);Z] = I[X;Z]$ ← does not depend on T
therefore, independent of representation
(up to diffeomorphisms)

· Conclusion:

(i) differences in h are scale independent, but extrinsic
(ii) information is intrinsic (& scale independent), at least up to diffeomorphism!

· Ex: $I[X;Y]$ is independent of $Var[X]$ and $Var[Y]$
(see Gaussian examples)

· Suggests there should be a unifying definition of discrete and differential info...

· if $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, let P be a partition of $\mathcal{X}$
Q ... of $\mathcal{Y}$

· let $X^{(P)}$, $Y^{(Q)}$ be the coarse-grained X & Y

· Def: $\underbrace{I(X;Y)}_{\text{possibly diff.}} = \sup\limits_{P,Q} \{ \underbrace{I(X^{(P)}; Y^{(Q)})}_{\text{discrete}} \}$

↑ "max over all partitions"

, unlike h, I is a limit of discrete approximations

and, coarse-graining never increases information shared by X and Y (usually discards some info)