Information Inequalities: (commentary on 3 key inequalities)

- **Def:** $I[X;Y] = D(P_{X,Y} \| P_X P_Y)$

- exploit Jensen, log-sum, & convexity of D to show (proofs in T&C)...

1. Information is Nonnegative: $I[X;Y] \geq 0$ and $= 0$ iff X and Y are independent

* we've seen this before...

(on average, observation reduces uncertainty...
$H[X|Y] \leq H[X]$, $H[X,Y] \leq H[X] + H[Y]$)

2. Data-Processing Inequality:

- **Def:** $X \to Y \to Z$ form a (3-step) Markov chain if Z is conditionally independent of X given Y
  i.e. $P_r(Z=z|Y=y, X=x) = P_r(Z=z|Y=y)$

  more generally: $X(t_1) \to X(t_2) \to \dots \to X(t_n)$ is a discrete-time Markov chain if $X(t_{s+k})$ is conditionally independent of $X(t_{s-k})$ given $X(t_s)$ $\forall\ k>0,\ s>0$
  (future ⫫ of past given present)

- Can use as a model for all data processing... receive Y, perform a deterministic or random procedure on Y that does not use any other info about X (Y is *all* our data regarding X), output Z

  Z represents processed Y

- **Inequality:** If $X \to Y \to Z$ is a Markov chain, then:
  $I[X;Z] \leq I[X;Y]$
  so, if $Z = g(Y)$, $I[X;T(Y)] \leq I[X;Y]$ for any g.
  (transform performed by processing)

  * more general case of the concavity of I... (see T&C 2.7.4)

  w/ equality iff X is conditionally independent of Y given Z

(processing never ↑ information.)

We saw this in the deterministic processing case for entropy... see coin tossing conversion example.
let $Y=X$, $Z = T(Y)$, $I[X;T(X)] \leq I[X;X] = H[X]$
(T some transform)

3. Fano's Inequality: let $X \sim p$, $X \in \mathcal{X}$. Observe Y. Use Y to estimate X via an estimator $\hat{X} = g(Y)$ for some g, possibly stochastic.

- **Inequality:** $P_r(\hat{X} = g(Y) \neq X) \geq \dfrac{H[X] - I[X;Y] - 1}{\log(|\mathcal{X}|)}$

  $P_r(\hat{X} = g(Y) \neq X)$: probability of an error in estimation

  * $H[X] - I[X;Y] = H[X|Y]$. information reduces conditional uncertainty after observing Y, reduces estimation error.

so,

1. Information is nonnegative and only zero if independent
2. Processing never increases information
3. Information reduces the probability of estimation error

1. and 3. are not too surprising.

2. (data-processing) is strikingly strong and somewhat confounding since we usually process data
   - indeed, almost all of stats & data science is about processing data to learn from it...

} so, let's dig into 2. a bit...

(the following section is ~ philosophy/argument, my aim here is to help answer the question: why is information intrinsic?)

- **Data-Processing Interrogated:**
  - let's start with an idea the data-processing inequality clarifies...
    - **Def:** Suppose $X \sim P_\theta$ where P_θ depends on an unknown parameter θ then:

      $g(X)$ is a sufficient statistic if X is conditionally independent of θ given $g(X)$.

    - **then:** (i) $\theta \to X \to g(X)$ is a Markov chain for any g (by construction)

      so, for most statistics g, $I[\theta; g(X)] \leq I[\theta; X]$

      (we lose info about the unknown by processing X to produce $g(X)$, unless...

      (ii) if $g(X)$ is sufficient:

      $\theta \to g(X) \to X$ is a Markov chain (X cond. $\perp\!\!\!\perp$ of θ given $g(X)$)

      then: $I[\theta; X] \leq I[\theta; g(X)]$ by data-processing

      together: $I[\theta; X] \overset{\text{(ii)}}{\leq} I[\theta; g(X)] \overset{\text{(i)}}{\leq} I[\theta; X]$ ... requires $I[\theta; g(X)] = I[\theta; X]$.

      ($I[\theta; X] \leq I[\theta; g(X)]$: sufficient; $I[\theta; g(X)] \leq I[\theta; X]$: statistic (i.e. processed X))

- So, data processing, $X \to g(X)$ loses information about θ unless $g(X)$ is sufficient and, if $g(X)$ is sufficient, processing loses no information...

- $I[\theta; g(X)] = I[\theta; X]$ iff $g(X)$ is sufficient.

  - Ex: $\{X_i\}_{i=1}^n \sim \text{Poisson}(\lambda)$ for unknown λ then $g(X) = \frac{1}{n}\sum_{i=1}^n X_i$ is sufficient

    (indeed, the sample mean is sufficient for $\text{Exp}(\lambda)$, $N(\mu, \sigma^2)$ for known σ

    $\{X_i\}_{i=1}^n \sim \text{Gamma}(r, s)$ for r, s unknown then $g(X) = \left[\frac{1}{n}\sum_{i=1}^n X_i, \left(\prod_{i=1}^n X_i\right)^{1/n}\right]$ is sufficient (sample arithmetic and geometric means).

- What if we can't find a sufficient statistic (the usual case)? then data-processing loses information!

- Contradicts (?) common intuition that we learn from data (extract information) by processing it

  - Ex: low-dimensional embedding/latent space representation

$$\underset{\text{high-D}}{X} \xrightarrow{\text{process}} \underset{\text{low-D}}{g(X)}$$

  - usually s.t. $g(X)$ has "intrinsically" meaningful coordinates
  - i.e. the position of data, axes, or relative position of data after applying g is meaningful (think: PCA, interpretable feature selection, word embedding, etc.)

- Notice: processing is all about changing representation
  information is intrinsic, representation agnostic
  - All of the original information is in X, we just might not know how to read/decode it...
    - We process data to change into a representation we understand...

- Ex: - encryption - all the information is in the encrypted message but need to decrypt (process) to read it

  - translation - you recieve a message in a language you don't read, all the information is there, it just needs to be translated to read

  - feature identification - often we measure many different variables without knowing a priori which are important/ what combination of variables matters again, all the information exploited is in the original data (we can't produce information unless we change our prior model - i.e. bring in outside info), the task is to change representation so that we can interpret the data.

- We process data to "decrypt" the information stored in it
  i.e. change representation so that we can interpret the data

  - this is why we built an intrinsic theory... to distinguish the information stored in data, independent of any encryption (lossless, deterministic, invertible change of representation) from our ability to interpret the information (represation dependent)

- information is only reduced or preserved by processing since processing either:
  - is deterministic and invertible: information is preserved since we can reconstruct X from $g(X)$
  - is sufficient: information is preserved since all desired relations w/ Y are encoded in $g(X)$
  - is stochastic or noninvertible and is not sufficient: information is lost since relevant aspects of X can't be recovered from $g(X)$,
    - if $X \neq X' \rightarrow g(X) = g(X')$ then can't reconstruct X from g
    - if g is noisy, can't reconstruct X w/ 100% certainty.

- thus, information theory cannot address the interpretability of the representation
  - it can provide bounds on the accuracy/limits of idealized decryption processes

- interpretability is subjective (I can't read Sanskrit, "it's all greek to me"; I can't read binary but my computer can)
- information is objective * (really, intrinsic)

- to compare & contrast information and interpretability as separate quantities we'd need a quantitative definition of interpretability

  - an idea: define interpretability as the computational complexity of some interpretation process that aims to decrypt/decode the message encoded in the data

- Ex: bootstrapping:

  $\{X_i\}_{i=1}^n \sim P$, $s: \mathcal{X}^n \to \mathbb{R}$ is some statistic of interest
  want to find the quantiles of $s(X)$ for uncertainty quantification

  - data: $\{X_i\}_{i=1}^n \sim P$ - desired quantity: sampling dist of $s(X_1, \dots X_n)$

  given $\{X_i\}_{i=1}^n$ we only have one sample of $s(X_1, \dots X_n)$,
  can't make a histogram/quantiles out of one sample

  - data processing: try bootstrapping: $\{X_i\}_{i=1}^n \longrightarrow g(\{X_i\}_{i=1}^n) = \{\{Y_j^{(k)}\}_{j=1}^n\}_{k=1}^m$
    (boot strapped samples)

  - interpretation: given $\{\{Y_j^{(k)}\}_{j=1}^n\}_{k=1}^m \longrightarrow \{s(Y_1^{(k)}, \dots Y_n^{(k)})\}_{k=1}^m$
    now have many samples, can compute (estimate) quantiles.

- it seems like we gained information by bootstrapping to get "new" data sets...
  (people will say this)

- But, bootstrapping is just a stochastic/randomized computation technique
  - it's just noisy data processing
- we didn't add new information by adding new randomness
  - all the draws $Y_j^{(k)} \overset{iid}{\sim} \underbrace{\text{Uniform}(\{X_j\}_{j=1}^n)}_{\text{"empirical distribution"}}$, i.e. are from the data we already had
  - $\{X_j\}_{j=1}^n$ fully specifies the dist. of bootstrap samples, thus the sampling dist. of $s(Y_1, \dots Y_n)$.
  - all info used was available in $\{X_j\}_{j=1}^n$

- really, the problem is that we don't know how to "decode" a sampling dist. from one sample
- so, process data to create many datasets $\{\{Y_j^{(k)}\}_{j=1}^n\}_{k=1}^m$
  - we know how to "decode" many sample data sets into quantiles. Cost is $\mathcal{O}(m)$. Cheap for most m.
  - Switching $\{X_j\}_{j=1}^n \to \{\{Y_j^{(k)}\}_{j=1}^n\}_{k=1}^m$ doesn't create info, rather it changes the representation of $\{X_j\}_{j=1}^n$ such that we can easily extract the sampling variability of s.

- in fact, the noisy process (bootstrap) loses information
  - ideally, we would enumerate every possible bootstrap sample and compute s for each, very expensive to fully enumerate, for a given n. Then would lose no information, but the cost would be enormous. Direct evaluation of sampling variability of s w.r.t. the empirical distribution is available w/ no info loss, but too computationally expensive to use

[• trade some information loss in processing for cheaper evaluation (easier interpretation)]

^

this tradeoff is key in cryptography...
safe/private codes should change representation while retaining information, but while requiring expensive decoding