CS 446/ECE 449: Machine Learning

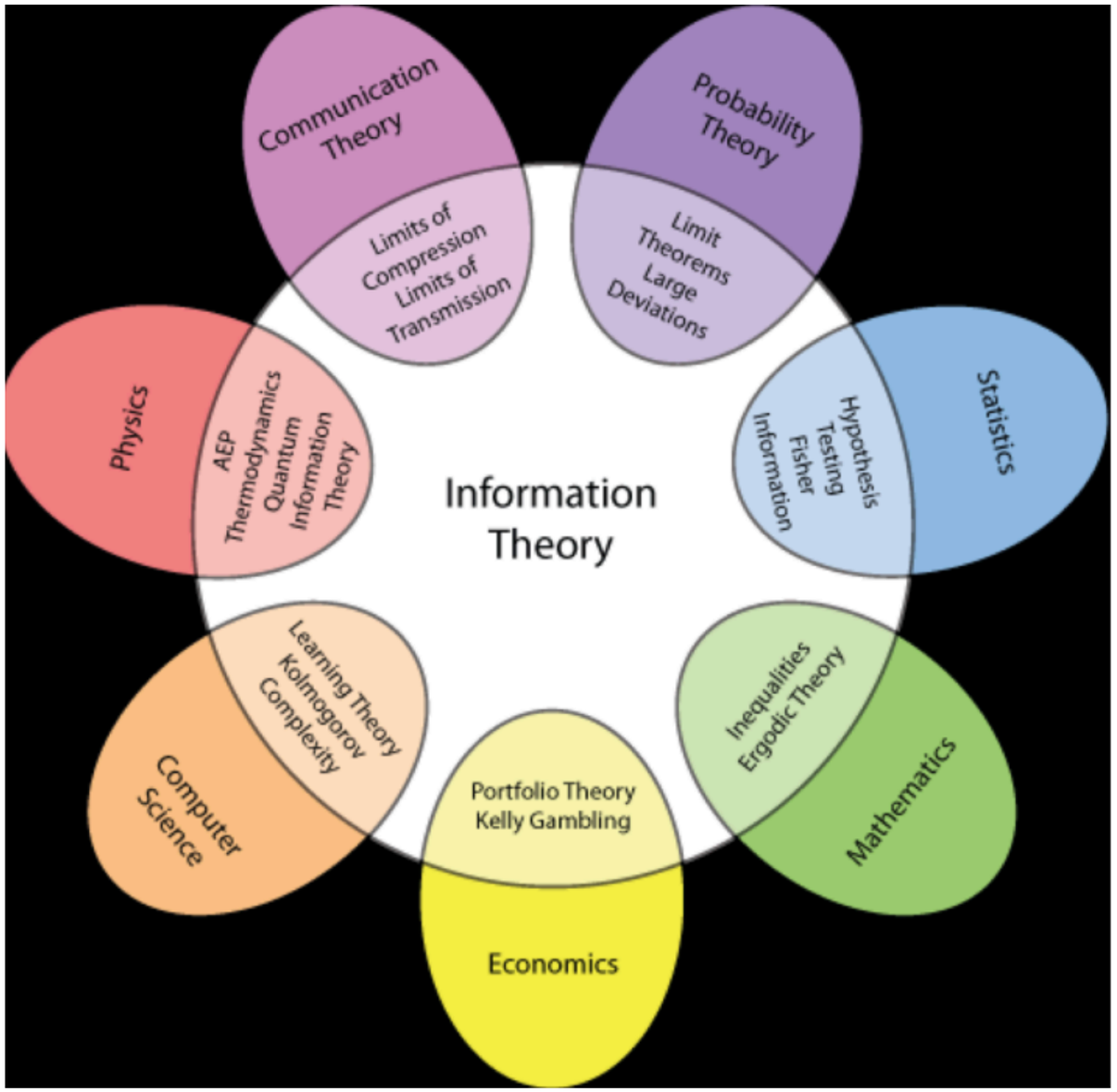
Lecture 19: Information Theory 101

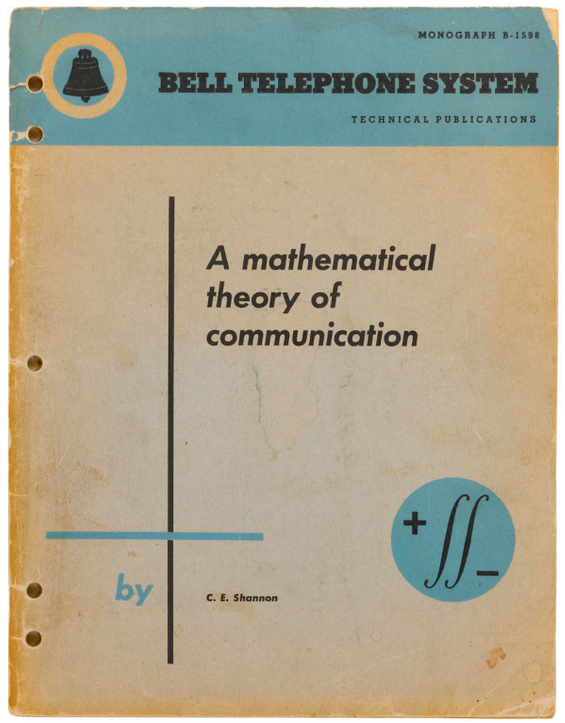
Han Zhao
03/28/2024

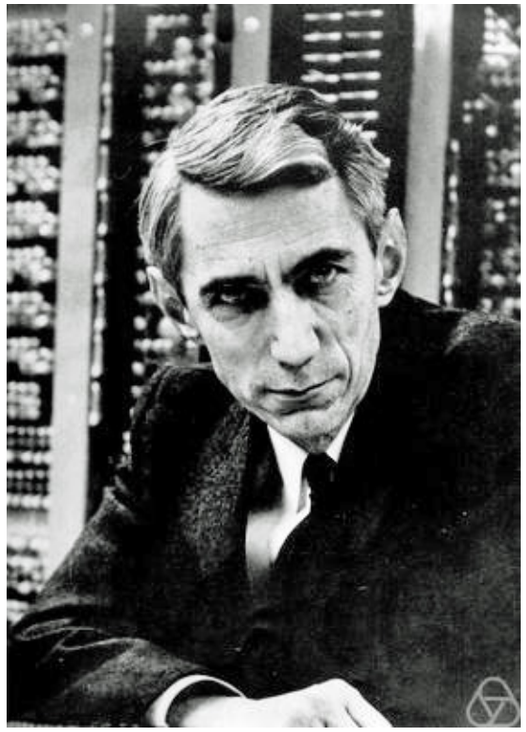
Today

- Entropy, KL-divergence, Mutual Information
- Total Variation distance & the distinguishing game
- Data-processing Inequality

Information Theory 101







Claude E. Shannon, 1948

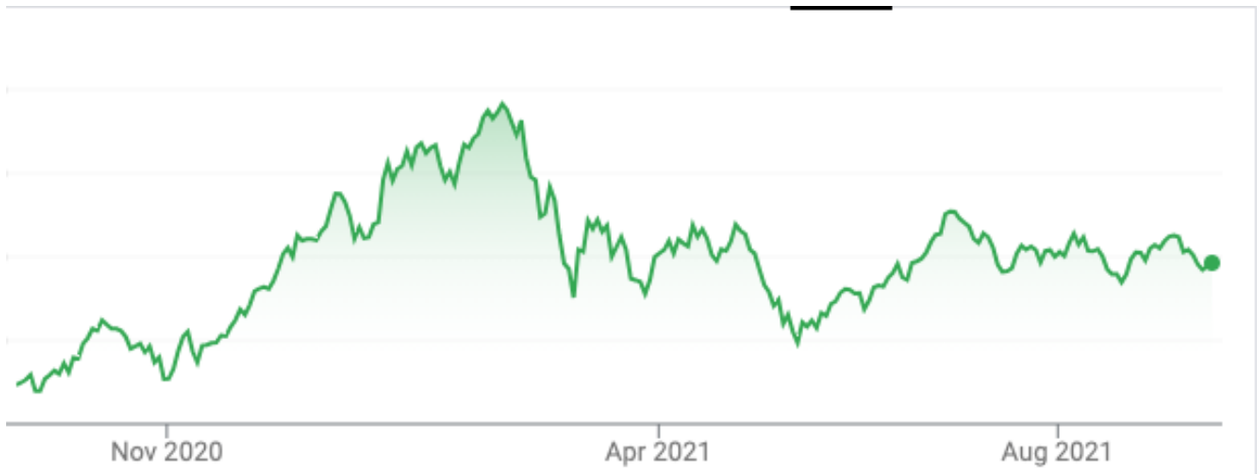
Information theory is the scientific study of the quantification, storage and communication of digital information

—Wikipedia

Information Theory 101

How to quantify information: information content of a random outcome

The sun always rises everyday, no uncertainty here



The price of a stock fluctuates, full of uncertainty

Information Theory 101

How to quantify information:

Let X be a discrete random variable taking n different values:

- Example of X: English letter in a document
- Example of X: last digit of S&P 500 index
- Example of X: result of a coin tossing
- ……

X takes n values with probability:

$$p_1, \ldots, p_n, \sum_{i=1}^{n} p_i = 1, p_i \geq 0$$

What's the uncertainty associated with X ?

Information Theory 101

X takes n values with probability:

$$p_1, \ldots, p_n, \sum_{i=1}^{n} p_i = 1, p_i \geq 0$$

What's the uncertainty associated with X ?

Intuitively:

- It should only depend on $p = (p_1, \ldots, p_n)$, independent of the alphabet
- Let's use $H(X) := H(p_1, \ldots, p_n)$ to denote this measure of uncertainty

Information Theory 101

Monotonicity:

Let $f(n) := H(1/n, \ldots, 1/n)$. If $n \leq m$, then $f(n) \leq f(m)$.

Grouping rule:

Let $p = (p_1, \ldots, p_m)$ be a probability distribution on m elements. Define a new distribution q on $m-1$ elements as $q_i = p_i, \forall i < m-1$ and $q_{m-1} = p_{m-1} + p_m$. We ask that the following equality holds:

$$H(p) = H(q) + (p_{m-1} + p_m) \cdot H\left(\frac{p_{m-1}}{p_{m-1} + p_m}, \frac{p_m}{p_{m-1} + p_m}\right)$$

Continuity

$H(p_1, \ldots, p_n)$ is continuous in $p = (p_1, \ldots, p_n)$.

Theorem (informal, Shannon 1948):

Under the three axioms above, (up to multiplicative constant $c > 0$), the Shannon entropy admits the following unique form:

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i$$

Information Theory 101

X takes n values with probability: an axiomatic approach

$$p_1, \ldots, p_n, \sum_{i=1}^{n} p_i = 1, p_i \geq 0$$

What's the uncertainty associated with X ?

Theorem (informal, Shannon 1948):

Under the three axioms above, (up to multiplicative constant $c > 0$), the Shannon entropy admits the following unique form:

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i$$

Note: 0 log(0) = 0

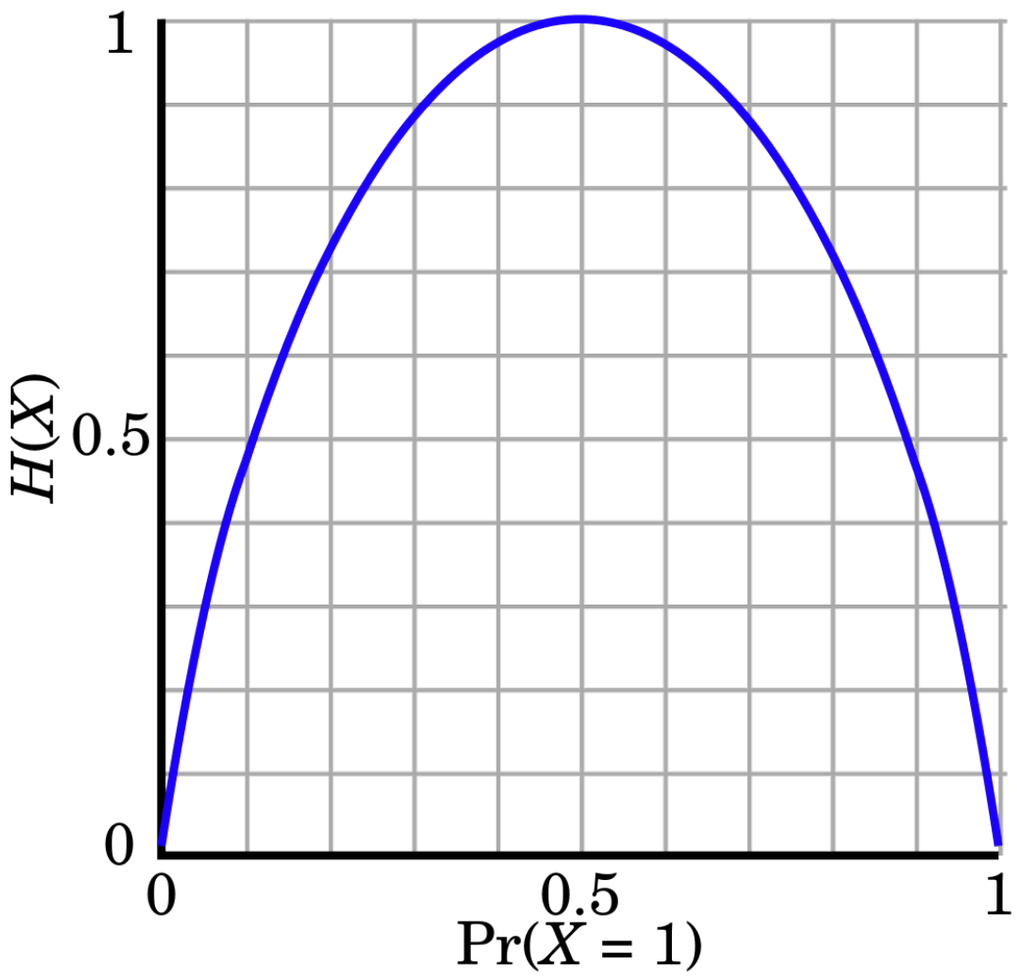
$$H(X) = \mathbb{E}[\log(1/p(X))]$$

Alternative perspective: the optimal average coding length

Information Theory 101

Some properties of entropy:

- For discrete RV, always nonnegative
- Concave in p
- Maximum entropy achieved with uniform distribution



Information Theory 101

Other related notions:

- For a pair of RV (X, Y) with joint probability $p(X, Y)$, the joint entropy is

$$H(X, Y) = -\mathbb{E}[\log p(X, Y)]$$

- Conditional entropy: the uncertainty of one variable given the knowledge of another

$$\begin{aligned} H(Y \mid X) &= \mathbb{E}[H(Y \mid X = x)] \\ &= \sum_x p(x) H(Y \mid X = x) \\ &= -\sum_x p(x) \sum_y p(y \mid x) \log p(y \mid x) \end{aligned}$$

- Chain rule:

$$H(X, Y) = H(X) + H(Y \mid X)$$

Information Theory 101

How to measure the discrepancy between two distributions, P, Q?

Kullback-Leibler divergence (KL-div, aka relative entropy):

Discrete: $$D_{\mathrm{KL}}(P \parallel Q) := \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

Continuous: $$D_{\mathrm{KL}}(P \parallel Q) := \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

Notes:

- Always non-negative
- Not symmetric, hence not a distance
- No upper bound in general (think about one example)

Information Theory 101

Kullback-Leibler divergence (KL-div, aka relative entropy):

Discrete:
$$D_{\mathrm{KL}}(P \parallel Q) := \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

Continuous:
$$D_{\mathrm{KL}}(P \parallel Q) := \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

Notes:

- Additive for independent distributions, i.e., if

$$P(x, y) = P_1(x) P_2(y), \; Q(x, y) = Q_1(x) Q_2(y)$$

then,

$$D_{\mathrm{KL}}(P \parallel Q) = D_{\mathrm{KL}}(P_1 \parallel Q_1) + D_{\mathrm{KL}}(P_2 \parallel Q_2)$$

Information Theory 101

Kullback-Leibler divergence (KL-div, aka relative entropy):

Discrete:
$$D_{\mathrm{KL}}(P \parallel Q) := \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

Continuous:
$$D_{\mathrm{KL}}(P \parallel Q) := \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

Examples:

- For two Bernoulli distributions $\mathrm{Bern}(p)$ and $\mathrm{Bern}(q)$,
$$D_{\mathrm{KL}}(p_1 \parallel p_2) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$$

- For two Gaussian distributions $\mathcal{N}(\mu_0, \Sigma), \mathcal{N}(\mu_1, \Sigma)$,
$$D_{\mathrm{KL}}(\mathcal{N}_0 \parallel \mathcal{N}_1) = \frac{1}{2}(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)$$
$$= \frac{1}{2} \|\mu_0 - \mu_1\|_{\Sigma}^2$$ Mahalanobis distance

Information Theory 101

Mutual Information: the reduction of uncertainty of one RV given the knowledge of another

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \end{aligned}$$

Equivalently,

$$\begin{aligned} I(X;Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= D_{\mathrm{KL}}(p(X,Y) \parallel p(X) \otimes p(Y)) \end{aligned}$$

Some identities:

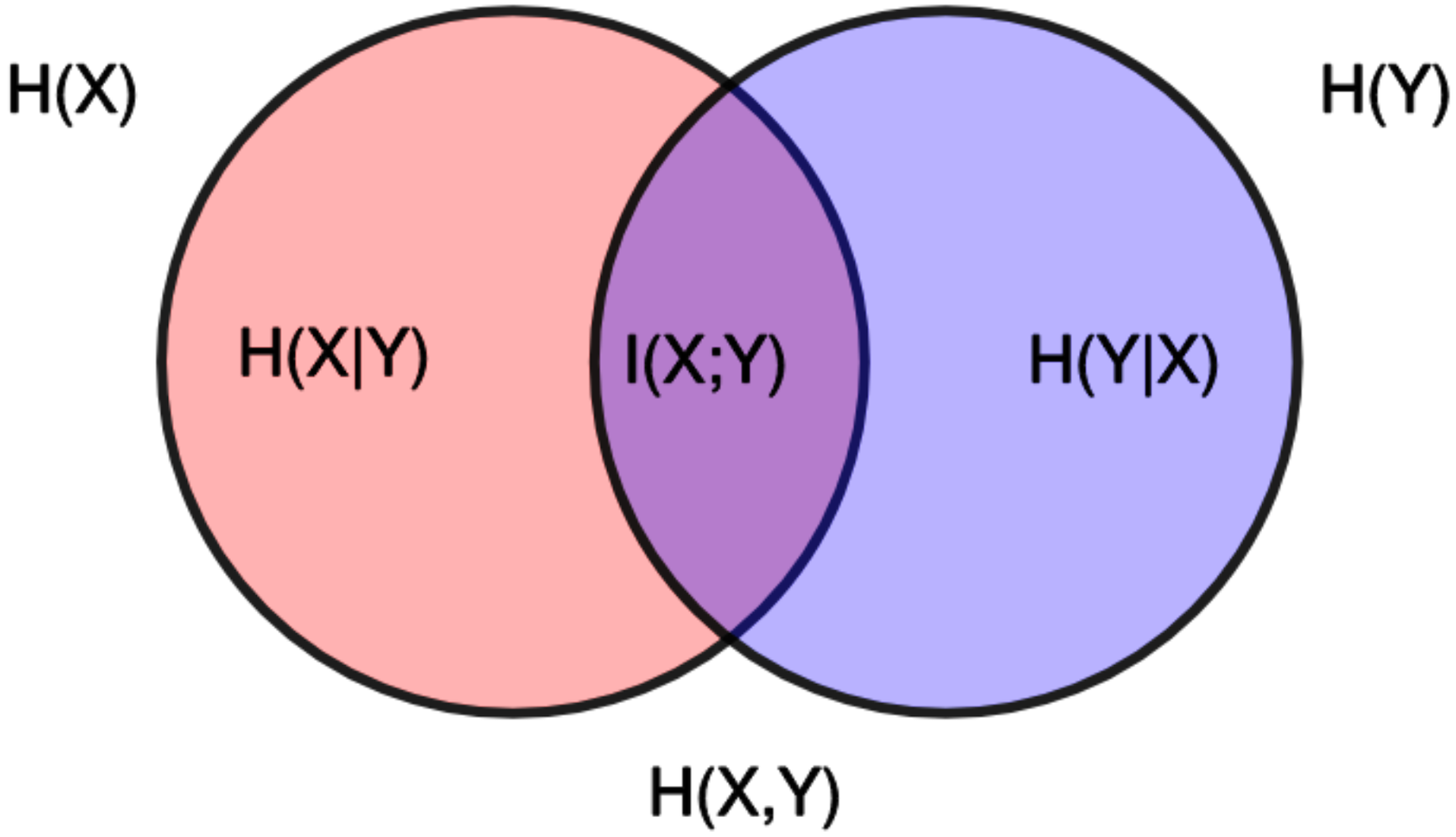
$$I(X;X) = H(X)$$

$$H(X,Y) = H(X) + H(Y) - I(X;Y)$$

$I(X;Y) = 0$ iff X, Y independent with each other.

Information Theory 101

Venn diagram between entropy, conditional entropy, joint entropy and mutual information



Information Theory 101

Total variation (TV) distance:

$$d_{\mathrm{TV}}(P, Q) := \sup_{A \subseteq \mathbb{R}^d} |P(A) - Q(A)|$$

Examples:

- TV-distance between two Bernoulli distributions $\mathrm{Bern}(p_0)$ and $\mathrm{Bern}(p_1)$: $|p_0 - p_1|$

Notes:

- Always non-negative, symmetric and bounded between [0, 1]
- TV-distance does not use the underlying geometry of $\mathbb{R}^d$
- TV-distance is a strong one, e.g., if μ is continuous, then

$$d_{\mathrm{TV}}(\hat{\mu}_n, \mu) = 1, \ \forall n$$

Information Theory 101

Total variation (TV) distance:

$$d_{\mathrm{TV}}(P,Q) := \sup_{A \subseteq \mathbb{R}^d} |P(A) - Q(A)|$$

Proposition:

Let P, Q be two probability distributions over the same space, then

$$d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\sum_{x \in \mathcal{X}} |P(x) - Q(x)| =: \frac{1}{2}\|P - Q\|_1$$

Proof:

Let A^* be the event such that $d_{\mathrm{TV}}(P,Q) = P(A^*) - Q(A^*)$. Then

$$\forall x \in \mathcal{X} \backslash A^*,\ P(x) < Q(x)$$

Hence,

$$\begin{aligned}
\|P - Q\|_1 &= \sum_{x \in \mathcal{X}} |P(x) - Q(x)| \\
&= \sum_{x \in A^*} (P(x) - Q(x)) + \sum_{x \in \mathcal{X} \backslash A^*} (Q(x) - P(x)) \\
&= (P(A^*) - Q(A*)) + (1 - Q(A^*) - (1 - P(A^*))) \\
&= 2\,(P(A^*) - Q(A*)) = 2 d_{\mathrm{TV}}(P,Q)
\end{aligned}$$

Information Theory 101

Total variation (TV) distance:

$$d_{\mathrm{TV}}(P, Q) := \sup_{A \subseteq \mathbb{R}^d} |P(A) - Q(A)|$$

Interpretation of the TV-distance as a binary classification problem:

- Let P, Q be two known distributions
- An adversary chooses a distribution D as follows:

$$D = \begin{cases} P & w.p.\ 0.5 \\ Q & w.p.\ 0.5 \end{cases}$$

- You get to see a sample from D, and then make a (possibly randomized) guess about whether D is P or Q?
- Given knowledge about P, Q, upon receiving the sample x, what's your best strategy?

Information Theory 101

Interpretation of the TV-distance as a binary classification problem:

$$\eta(x) := \Pr(\text{my guess is } P \mid x)$$

Then, the error probability of using strategy $\eta(\cdot)$ when seeing x is:

$$\begin{aligned}\Pr(\text{error}, x) &= \Pr(\text{guess } P, x \sim Q) + \Pr(\text{guess } Q, x \sim P) \\ &= \frac{1}{2}Q(x)\eta(x) + \frac{1}{2}P(x)(1-\eta(x))\end{aligned}$$

Hence, the overall error probability is given by:

$$\begin{aligned}\Pr(\text{ error }) &= \sum_{x\in\mathcal{X}} \Pr(\text{ error }, x) \\ &= \sum_{x\in\mathcal{X}} \frac{1}{2}Q(x)\eta(x) + \frac{1}{2}P(x)(1-\eta(x)) \\ &= \frac{1}{2} + \frac{1}{2}\sum_{x\in\mathcal{X}} \eta(x)(Q(x)-P(x))\end{aligned}$$

Information Theory 101

Interpretation of the TV-distance as a binary classification problem:

$$\eta(x) := \Pr(\text{my guess is } P \mid x)$$

Hence, the overall error probability is given by:

$$\Pr(\text{error}) = \frac{1}{2} + \frac{1}{2} \sum_{x \in \mathcal{X}} \eta(x) \left(Q(x) - P(x)\right)$$

So, clearly, to minimize the overall guessing error, the best strategy is given by:

$$\eta(x) = \begin{cases} 1 & P(x) \geq Q(x) \\ 0 & P(x) < Q(x) \end{cases}$$

Under this optimal strategy, the optimal distinguishing error is:

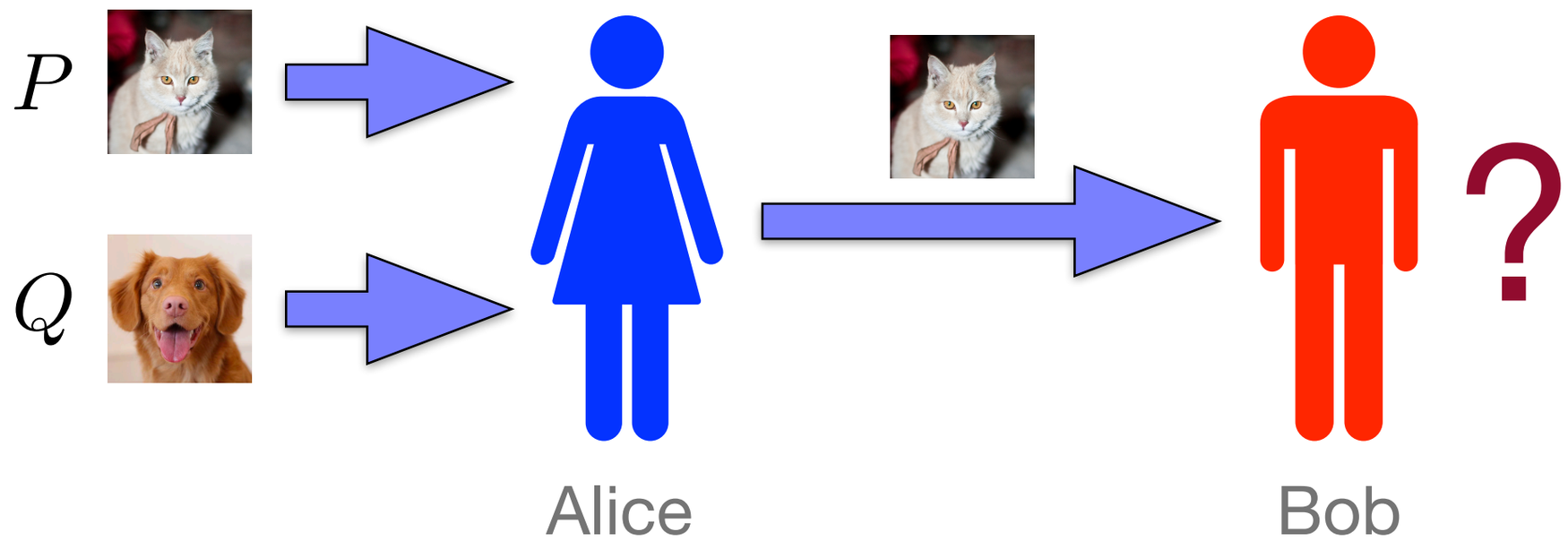
$$\begin{aligned} \Pr(\text{error}) &= \frac{1}{2} + \frac{1}{2} \sum_{x : P(x) \geq Q(x)} Q(x) - P(x) \\ &= \frac{1}{2} - \frac{1}{2} d_{\mathrm{TV}}(P, Q) \end{aligned}$$

Consider some extremal cases of this distinguishing game

Information Theory 101

Total Variation (TV) distance & the distinguishing game

TV distance: $d_{\mathrm{TV}}(P, Q) := \sup_{A \subseteq \mathbb{R}^d} |P(A) - Q(A)|$



Bob's optimal error: $\Pr(\text{error}) = \frac{1}{2} + \frac{1}{2} \sum_{x : P(x) \geq Q(x)} Q(x) - P(x)$

$$= \frac{1}{2} - \frac{1}{2} d_{\mathrm{TV}}(P, Q)$$

Information Theory 101

Relationship between the two divergence measures we discussed so far:

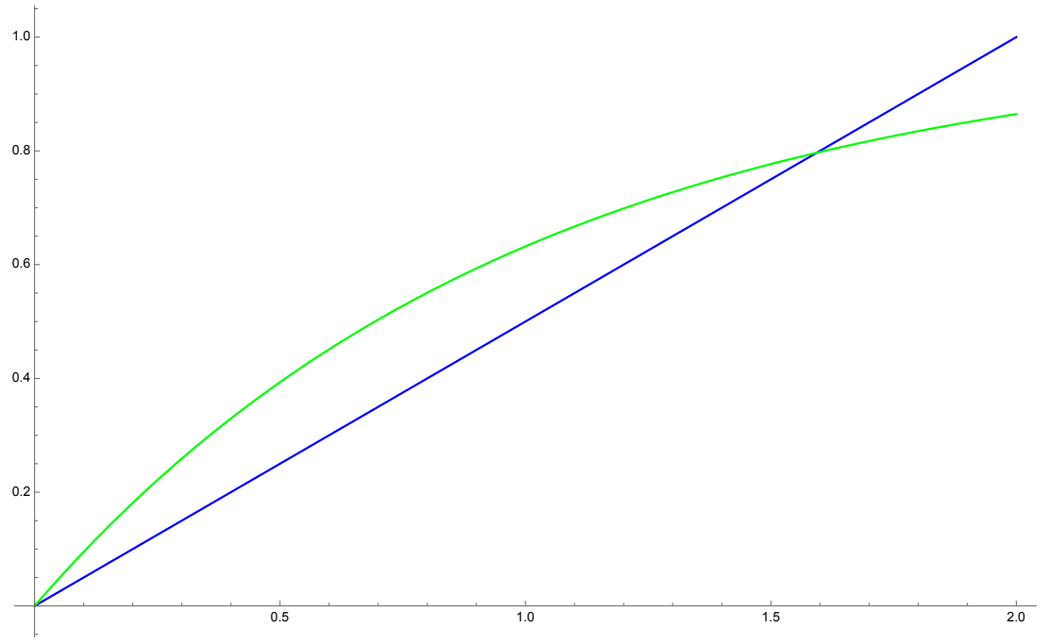
Pinsker's inequality

$$d_{\mathrm{TV}}(\cdot,\cdot) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\cdot\|\cdot)}$$

Additionally, for TV and KL, we have:

$$d_{\mathrm{TV}}(\cdot,\cdot) \leq \sqrt{1-\exp(-D_{\mathrm{KL}}(\cdot \parallel \cdot))}$$

The second one is better when KL is large:



Information Theory 101

Data processing pipeline:

Nature Camera CD

Information Theory 101

Markov chain: we say that X, Y, Z forms a Markov chain in this order, denoted as $X \to Y \to Z$, if

$$p(x, y, z) = p(x)p(y \mid x)p(z \mid y)$$

Intuitively, given the knowledge of Y, X, Z are conditionally independent from each other.

- Example (coin tossing): probability of getting a head is p. Generate a sequence of independent tosses $X_1, \ldots, X_n$, then

$$p \to \{X_1, \ldots, X_n\} \to \bar{X}_n$$

- Example (multilayer perceptron): X the input image. $Z_i = f_i(Z_{i-1})$ the i-th layer representation with $Z_0 = X$, then

$$X \to Z_1 \to Z_2 \to \cdots \to Z_L$$

Information Theory 101

Data Processing Inequality: If $X \to Y \to Z$, then

$$I(X;Y) \geq I(X;Z), \quad I(Y;Z) \geq I(X;Z)$$

Intuitively:

- Further processing of data cannot increase mutual information

Proof:

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y \mid Z) \\ &= I(X;Y) + I(X;Z \mid Y) \end{aligned}$$

But

$$I(X;Z \mid Y) = 0$$

So,

$$I(X;Y) \geq I(X;Z)$$

The other one follows since Z, Y, X also forms a Markov chain