Lecture Notes for "Stochastic Modeling and Computations"

M. Chertkov (lecturer), S. Belan and V. Parfeneyv (recitation instructors)
M.Sc. and Ph.D. level course at Skoltech
Moscow, March 28 - May 28, 2016
`https://sites.google.com/site/mchertkov/courses`

The course offers a soft and self-contained introduction to modern applied probability covering theory and application of stochastic models. Emphasis is placed on intuitive explanations of the theoretical concepts, such as random walks, law of large numbers, Markov processes, reversibility, sampling, etc., supplemented by practical/computational implementations of basic algorithms. In the second part of the course, the focus shifts from general concepts and algorithms per se to their applications in science and engineering with examples, aiming to illustrate the models and make the methods of solution, originating from physics, chemistry, machine learning, control and operations research, clear.

Theme #1. Basic Concepts from Statistics

A. Lecture #1. Random Variables: Characterization & Description.

1. *Probability of an event*

Discrete vs Continuous events. State/phase/sample space (for events), Σ.

Example of discrete events: two states, $\Sigma = \{0, 1\}$ -also called Bernoulli random variable (derived from a "process", i.e. dynamics - to be discussed later in the course a lot). Probability of a state, σ,

$$\forall \sigma \in \Sigma: \quad \text{Prob}(\sigma) = P(\sigma) \tag{.1}$$

$$0 \le P(\sigma) \le 1 \tag{.2}$$

$$\sum_{\sigma \in \Sigma} P(\sigma) = 1 \tag{.3}$$

For Bernoulli process, $P(1) = \beta$, $P(0) = 1 - \beta$.

Question: Can you give an example of the Bernoulli distribution from life/science?

Answer: A biased coin.

Another important discrete event distribution is the Poisson. An event can occur $k = 0, 1, 2, \cdots$ times in an interval. The average number of events in an interval is λ - called event rate. The probability of observing k events within the interval is

$$\forall k \in \mathbb{Z}^* = \{0\} \bigcup \mathbb{Z}: \quad P(k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{.4}$$

(Check that the probability is properly normalized, in the sense of Eq. (.3).) The distribution is also called exponential distribution (for obvious reason).

Questions: Are Bernoulli and Poisson distributions related? Can you "design" Poisson from Bernoulli? Can you give an example of the Poisson process from life/science?

Answer: Consider repeating Bernoulli - thus drawing a Bernoulli process. You get sequence of zeros and ones. Then check only for ones and record times/slots associated with arrivals of ones. Study probability distribution of t arrivals in n step, and then analyze $n \to \infty$, to get the Poisson distribution. We will discuss it in details in Lect.# 5. Example of Poisson — arrival of customers at the shop.

The domain can be continuous, bounded or unbounded. Example of distribution which is bounded - is uniform distribution from the $[0, 1]$ interval:

$$\forall x \in [0, 1]: \quad p(x) = 1, \tag{.5}$$

$$\int_0^1 dx p(x) = 1, \tag{.6}$$

where $p(x)$ is the probability density. Gaussian distribution is the most important continuous distribution:

$$\forall x \in \mathbb{Z}: \quad p(x|\sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{.7}$$

$p_{\sigma,\mu}(x)$ another possible notation. It is also called "normal distribution" - where "normality" refers to significance of the distribution for the central limit theorem (law of large numbers), which we will be discussing shortly. The distribution is parameterized by μ and λ - what is the significance of the two parameters? (mean and variance) Standard notation in math for the Gaussian/normal distribution is $N(\mu, \sigma^2)$.

There are many more 'standard' distributions but Bernoulli, Poisson and Gaussian are the 'golden' three. One can generate practically any other distribution from the 'golden set' (possibly extended by the uniform distribution.

Some discussion of notations, e.g. $P_X(x)$, $\mathbb{E}\left[\cdots\right] = \langle\cdots\rangle$.

2. Sampling. Histograms.

Random process generation. Random process is generated/sampled. Any computational package/software contains a random number generator (in fact a number of these). Designing a good random generation is important. In this course, however, we will mainly using the random number generators (in fact pseudo-random generators) already created by others.

To illustrate let us switch to IJulia notebook of the first two lectures.

Histogram. To show distributions graphically, you may also "bin" it in the domain - thus generating the histogram, which is a convenient way of showing $p(\sigma)$ (plot with Julia: breaking $[0, 1]$ interval in $N > 1$ bins). Use it as an opportunity to introduce statistical computational package (Julia should have one too).

3. Moments. Generating Function.

Expectations.

$$\mathbb{E}\left[A(\sigma)\right]_p = \langle A(\sigma)\rangle_p = \sum_{\sigma\in\Sigma} A(\sigma)p(\sigma).$$

Examples: mean,

$$\mathbb{E}\left[\sigma\right],$$

variance,

$$\text{Var}[\sigma] = \mathbb{E}\left[(\sigma - \mathbb{E}[\sigma])^2\right].$$

We have already discussed these for the Gaussian process. What are mean and variance for Bernoulli process?

Moments.

$$k = 0, \cdots, \quad \mathbb{E}\left[\sigma^k\right]_p = \langle\sigma^k\rangle_p = \sum_{\sigma\in\Sigma} \sigma^k p(\sigma) = m_k(\Sigma)$$

Moment Generating Function.

$$M_X(t) = \mathbb{E}\left[\exp(tx)\right], \quad t \in \mathbb{R}$$

One can also view it as a Laplace transform of the probability density function,

$$M_X(t) = \int dx p(x)\exp(tx)$$

. Examples of the moment generating functions for aforementioned (and other) distributions — derived it yourself, see tables online ... and it will also be discussed at the recitations. Characteristic function is a related object — Fourier transform of the probability density:

$$\mathbb{E}\left[\exp(itx)\right] = \int dx p(x)\exp(itx)$$

where $i^2 = -1$.

4. Probabilistic Inequalities.

Here are some useful probabilistic inequalities.

- (Markov Inequality)

$$P(x \geq c) \leq \frac{\mathbb{E}[x]}{c} \tag{.8}$$

- (Chebyshev's inequality)

$$P(|x-\mu| \geq k) \leq \frac{\sigma^2}{k^2} \tag{.9}$$

- (Chernoff bound)

$$P(x \geq a) = P(e^{tx} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tx}]}{e^{ta}} \tag{.10}$$

where μ and σ are mean and variance of x.

We will get back to discussion of these and some additional inequalities in the third lecture.

Exercise: Play in IJulia checking the three inequalities for the distributions mentioned through out the lecture.

Exercise: Prove the Markov inequality. Chebyshev inequality will follow from the Markov, prove it too. Chernoff is trickier, can you prove it too? [See `http://jeremykun.com/2013/04/15/probabilistic-bounds-a-primer/` to check your answers]

Exercise: Provide examples of the distributions for which the tree inequalities are saturated (becomes equalities)?

5. Recitation. Random Variables. Moments. Characteristic Function.

B. Lecture #2. Random Variables: from one to many.

1. Law of Large Numbers

Take n samples $x_1, \cdots, x_n$ generated i.i.d. from a distribution with mean μ and variance, $\sigma > 0$, and compute $y_n = \sum_{i=1}^n x_i/n$. What is Prob(y_n)? $\sqrt{n}(y_n - \mu)$, converges in distribution to Gaussian with mean, μ, and variance, σ:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n x_i - \mu\right) \xrightarrow{d} N(0,\ \sigma^2). \tag{.11}$$

This is so-called Weak Version of the Central Limit Theorem. (Large Deviation Theorem is an alternative name.)

Let us prove the weak-CLT (.11) in a simple case $\mu = 0$, $\sigma = 1$. (Generalization is obvious.) Obviously, $m_1(Y_n\sqrt{n}) = 0$. Compute

$$m_2(Y_n\sqrt{n}) = \mathbb{E}\left[\left(\frac{x_1 + \cdots + x_n}{\sqrt{n}}\right)^2\right] = \frac{\sum_i \mathbb{E}\left[x_i^2\right]}{n} + \frac{\sum_{i\neq j}\mathbb{E}\left[x_i x_j\right]}{n} = 1.$$

Now the third moment:

$$m_3(Y_n\sqrt{n}) = \mathbb{E}\left[\left(\frac{x_1 + \cdots + x_n}{\sqrt{n}}\right)^3\right] = \frac{\sum_i \mathbb{E}\left[x_i^3\right]}{n^{3/2}} \to 0,$$

at $n \to \infty$, assuming $\mathbb{E}\left[x_i^3\right] = O(1)$. Can you guess what will happen with the fourth moment? $m_4(Y_n\sqrt{n}) = 3 = 3m_2(Y_n)$ - Wick theorem (physics jargon). And how about higher odd/even moments?

Exercise: Check IJulia notebok for the lecture and experiment with the law of large numbers for different distributions.

The theorem holds for independent but not identically distributed variables too.

If one is interested in not only the asymptotic itself, $n \to \infty$, but also in how the asymptotic is approached, the so-called strong version of CLT (can also be found in some literature under the name of Cramér theorem) states

$$\forall z > \mu : \quad \lim_{n\to\infty} \frac{1}{n} \log \text{Prob}\left(y_n \geq z\right) = -\Phi^*(z) \tag{.12}$$

$$\Phi^*(z) \doteq \sup_{\lambda\in\mathbb{R}} \left(\lambda z - \Phi(z)\right) \tag{.13}$$

$$\Phi(\lambda) \doteq \log\left(\mathbb{E}\exp\left(\lambda z\right)\right) \tag{.14}$$

$\Phi^*(z)$ is a convex function (also called Cramér function). This was a formal (mathematical) statement. A less formal (physical) version of Eq. (.12) is

$$n \to \infty : \quad \text{Prob}\left(y_n\right) \propto \exp\left(-n\Phi^*(x)\right) \tag{.15}$$

One of our journal club projects is on this subject.

Note, that the weak version of the CLT (.11) is equivalent to approximating the Cramer function (asymptotically exact) by a Gaussian around its minimum.

Exercise (bonus): Prove the strong-CLT (.12,.13). [Hint: use saddle point/stationary point method to evaluate the integrals.]

Exercise: Give an example of an expectation for which not only vicinity of the minimum but also other details of $\Phi^*(x)$ are significant at $n \to \infty$? More specifically give an example of the object which behavior is controlled solely by left/right tail of $\Phi^*(x)$? $\Phi^*(0)$ and its vicinity?

Example of Bernoulli process – a (possibly unfair) coin toss

$$x = \begin{cases} 0 & \text{with probability } 1-p \\ 1 & \text{with probability } p \end{cases} \tag{.16}$$

Then

$$\Phi(\lambda) = \log(pe^{\lambda} + 1 - p) \tag{.17}$$

$$0 < x < 1 : \quad \Phi^*(z) = z\log\frac{z}{p} + (1-z)\log\frac{1-z}{1-p} \tag{.18}$$

Eqs. (.17,.18) are noticeable for two reasons. First of all it leads (after some algebraic manipulations) to the famous Stirling formula for the asymptotic of a factorial

$$n! = \sqrt{2\pi n} n^n e^{-n}\left(1 + O(1/n)\right)$$

. Do you see how? Second, the $z\log z$ structure is an "entropy" which will appear few more times in the course - stay tuned.

2. *Multivariate Distribution. Marginalization. Conditional Probability.*

Consider an n-component vector build of components each taking a value from a set, Σ, $\sigma = (\sigma_i \in \Sigma | i = 1, \cdots, n)$. Σ may be discrete, e.g. $\Sigma = \{0, 1\}$, or continuous, e.g. $\Sigma = \mathbb{R}$. Assume that any state, σ, occur with the probability, $P(\sigma)$, where $\sum_\sigma P(\sigma) = 1$.

Consider a simple example of bi-variate distribution.

$$\sigma = (\sigma_i = \pm 1 | i = 1, \cdots, n) : P(\sigma) = Z^{-1} \prod_{i=1}^{n-1} \exp(J\sigma_i\sigma_{i+1}) \tag{.19}$$

$$Z = \sum_\sigma \prod_{i=1}^{n-1} \exp(J\sigma_i\sigma_{i+1}) \tag{.20}$$

where Z is the normalization constant (also called partition function in physics), introduced to guarantee that the sum over all the states is unity. For $n = 2$ we can also write

$$P(\sigma) = P(\sigma_1, \sigma_2) = \frac{\exp(J\sigma_1\sigma_2)}{4\cosh(J)}. \tag{.21}$$

$P(\sigma)$ is also called a joint distribution function of the σ vector components, $\sigma_1, \cdots, \sigma_n$. It is also useful to consider conditional distribution, say for the example above with $n = 2$,

$$P(\sigma_1|\sigma_2) = \frac{P(\sigma_1, \sigma_2)}{\sum_{\sigma_1} P(\sigma_1, \sigma_2)} = \frac{\exp(J\sigma_1\sigma_2)}{2\cosh(J\sigma_2)} \tag{.22}$$

is the probability to observe σ_1 under condition that σ_2 is known. Notice that, $\sum_{\sigma_1} P(\sigma_1|\sigma_2) = 1,\ \forall\sigma_2$.

Let us now marginalize the multivariate (joint) distribution over a subset of variables. For example,

$$P(\sigma_1) = \sum_{\sigma\setminus\sigma_1} P(\sigma) = \sum_{\sigma_2,\cdots,\sigma_n} P(\sigma_1, \cdots, \sigma_n). \tag{.23}$$

We will repeat exercises (joint, conditional, marginal) with multivariate Gaussian distribution at the recitations. The Gaussian distributions are remarkably unique, because application of any of the aforementioned operations, joint-to-conditional and joint-to-marginal, will also be Gaussian ... not to mention that the Gaussian emerges naturally in the result of the CLT.

3. Bayes Theorem

We already saw how to get conditional distribution and marginal distribution from the joint one

$$P(x|y) = \frac{P(x, y)}{P(y)}, \quad P(y|x) = \frac{P(x, y)}{P(x)}. \tag{.24}$$

Combining the two formulas to exclude the joint probability distribution we arrive at the famous Bayes formula

$$P(x|y)P(y) = P(y|x)P(x). \tag{.25}$$

Here, in Eqs. (.24,.26) both x and y may be multivariate.

Rewriting Eq. (.26) as

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}, \tag{.26}$$

one often refers (in the field of the so-called Bayesian inference/reconstruction) to $P(x)$ as the "prior" probability – degree of initial "belief" in x, $P(x|y)$ - the "posterior" - degree of belief having accounted for y, and quotient $\frac{P(y|x)}{P(y)}$ representing "support/knowledge" y provides for x.

A good illustration of the notion of conditional probability can be found at `http://setosa.io/ev/conditional-probability/`

Let us conclude the lecture playing with a made up bi-variate binary (with the total of 2^2 states) case.

4. Recitation. Properties of Gaussian Distributions. Laws of Large Numbers.

C. Lecture #3. Information-Theoretic View on Randomness

1. Entropy.

Entropy is defined as an expectation of $-log$-probability

$$S(X) = -\mathbf{E}\left[\log(P(x))\right] = -\sum_{x\in\mathcal{X}} P(x)\log(P(x)). \tag{.27}$$

Intuitively, entropy is a measure of uncertainty. Simple illustration that entropy of a deterministic process (when a state happens with probability 1) is 0 ($\lim_{p\to 0} p\log p = 0$). Note, that following statistical physics tradition we use S for entropy, while it is also custom in information theory to use H for the same object.

Importantly, logarithm of the probability distribution is chosen as a measure of information in the definition of entropy (and not another function) because it is **additive** for independent sources.

Let us familiarize ourselves with the concepts of entropy on the example of the Bernoulli $\{0,1\}$ process (.16)

$$S(X) = -p\log p - (1-p)\log(1-p). \tag{.28}$$

If we plot the entropy as the function of p. It has a bell-like shape with the maximum at $p = 1/2$ - fare coin has the largest entropy (most uncertain). Entropy is zero at $p = 0$ and $p = 1$ - the two cases are deterministic, i.e. fully certain.

Entropy (.27), $S(X)$, has the following properties (some can be interpreted as alternative definitions):

- $S(X) \geq 0$
- $S(X) = 0$ iff x is deterministic.
- $S(X) \leq \log(|\mathcal{X}|)$ and $S(X) = \log(|\mathcal{X}|)$ iff x is distributed uniformly over the set $\mathcal{X}$.
- Choice of the logarithm base is custom - just a re-scaling. (Base 2 is custom in the information , when dealing with binary variables.)
- Entropy is the measure of average uncertainty.
- Entropy is less than the average number of bits needed to describe the random variable (the equality is achieved for uniform distribution). (*)
- Entropy is the lower bound on the average length of the shortest description of the random variable

(*) requires a clarification. Take integers which are smaller or equal then n, and represent them in the binary system. We will need $\log_2(n)$ binary variables (bits) to represent any of the integers. If all the integers are equally probable then $\log_2(n)$ is exactly the entropy of the distribution. If the random variable is distributed non-uniformly than the entropy is less than the estimate.

Exercise: Order the following three cases in terms of entropy: (a) 5 equally probable states; b) 3 states which happens with the probabilities $1/2, 1/6, 1/3$; c) 6 states which happen with the probabilities $1/2, 1/10, 1/10, 1/10, 1/10, 1/10$.

If we have a pair of (discrete for concreteness) variables, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ their joint entropy is

$$S(X,Y) \doteq -\sum_{y\in\mathcal{Y}} p(x,y)\log(p(x,y)). \tag{.29}$$

Conditional entropies are

$$S(Y|X) \doteq -\mathbb{E}_{p(x,y)}\left[\log(p(y|x)\right] = -\sum_{y\in\mathcal{Y}} p(x,y)\log(p(y|x)). \tag{.30}$$

Note, that $S(Y|X) \neq S(X|Y)$.

The so-called chain rule states (check)

$$S(X,Y) = S(X) + S(Y|X). \tag{.31}$$

One can also extend it to the multi-variate case $(X_1, \cdots, X_n) \sim P(x_1, \cdots, x_n)$ (this notation is standard in statistics) becomes

$$S(X_n, \cdots, X_1) = \sum_{i=1}^{n} S(X_i|X_{i-1}, \cdots, X_1). \tag{.32}$$

The name "chain-rule" should become clear from (.32). The chain rule is illustrated in Fig. (1).

2. Independence/Dependence. Mutual Information.

The essence of our next theme is in comparing random numbers, or more accurately thei probabilities. Kublack-Leibler divergence offers a convenient way of measuring two probabilities

$$D(p_1\|p_2) \doteq \sum_{x\in\mathcal{X}} p_1(x)\log\frac{p_1(x)}{p_2(x)}. \tag{.33}$$

Note that the KL difference is not symmetric wrt exchange of the order of the two distribution. Moreover it is not a proper metric of comparison (it does not satisfy the triangle inequality (Any proper metric of a space should be a) positive, b) zero when

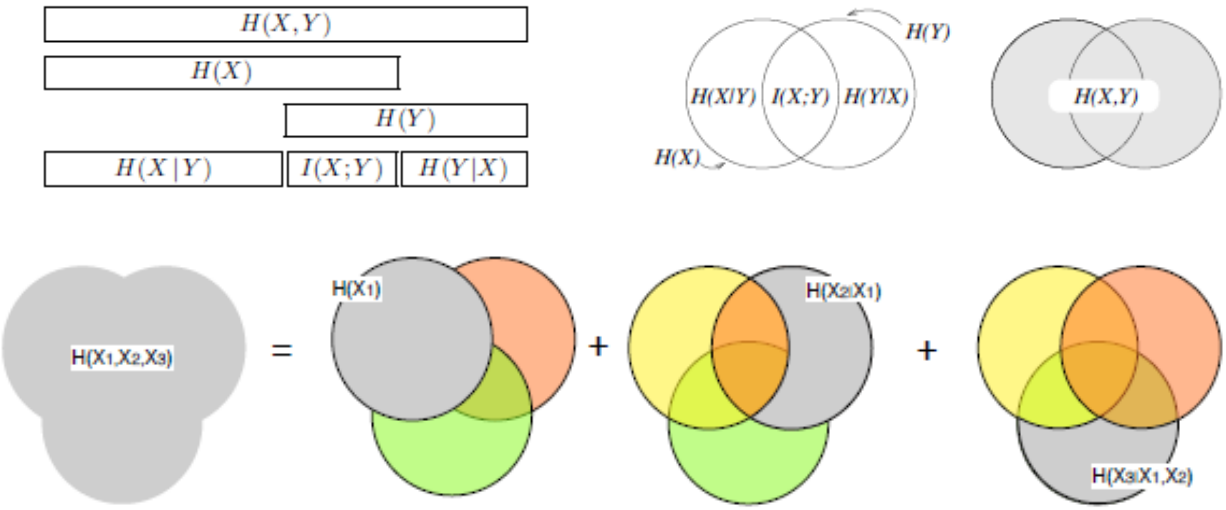

FIG. 1: Venn diagram(s) explaining the chain rule for computing multivariate entropy.

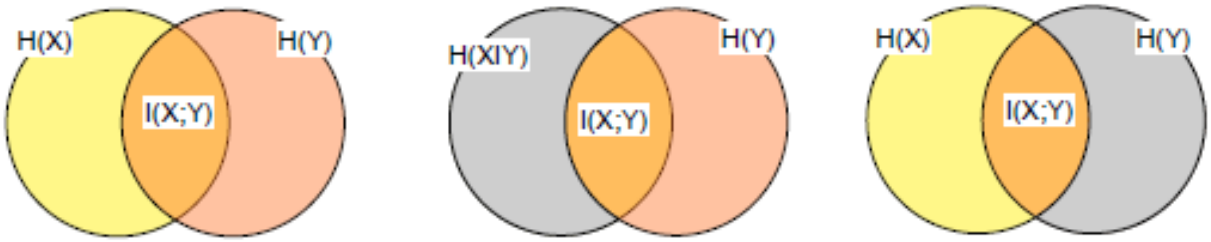

FIG. 2: Venn diagram explaining relations between mutual information and entropies.

comparing identical states; c) symmetric, and d) satisfy the triangle inequality, $d_(a,b) \leq d_{a,c} + d_{bc}$. The last two do not satisfy in the case of the KL divergence. However, an infintesimal version of KL divergence - Hesian of the KL distance, related to the so-called Fisher informarmation.)

Comparing the two information sources, say tracking events x and y, the extreme case is when the probabilities are independent, i.e. $P(x,y) = P(x)P(y)$ and $P(x|y) = P(x)$, $P(y|x) = P(y)$. Mutual information is the measure of dependence

$$I(X;Y) = \mathbb{E}_{P(x,y)}\left[\log \frac{P(x,y)}{P(x)P(y)}\right] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}. \tag{.34}$$

Intuitively the mutual information measures the information that x and y share. In other words, it measures how much knowing one of these variables reduces uncertainty about the other. For example, if x and y are independent, then knowing x does not give any information about y and vice versa - the mutual information is zero. In the other extreme, if x is a deterministic function of y then all information conveyed by x is shared with y. In this case the mutual information is the same as the uncertainty contained in x itself (or y itself), namely the entropy of x (or y).

Back to mutual information. Mutual information is obviously related to entropies,

$$I(X;Y) = S(X) - S(X|Y) = S(Y) - S(Y|X) = S(X) + S(Y) - S(X,Y). \tag{.35}$$

which is illustrated in Fig. (2). It also possesses the following properties

$$I(X;Y) = I(Y;X) \text{ (symmetry)} \tag{.36}$$
$$I(X;X) = S(X) \text{ (self-information)} \tag{.37}$$

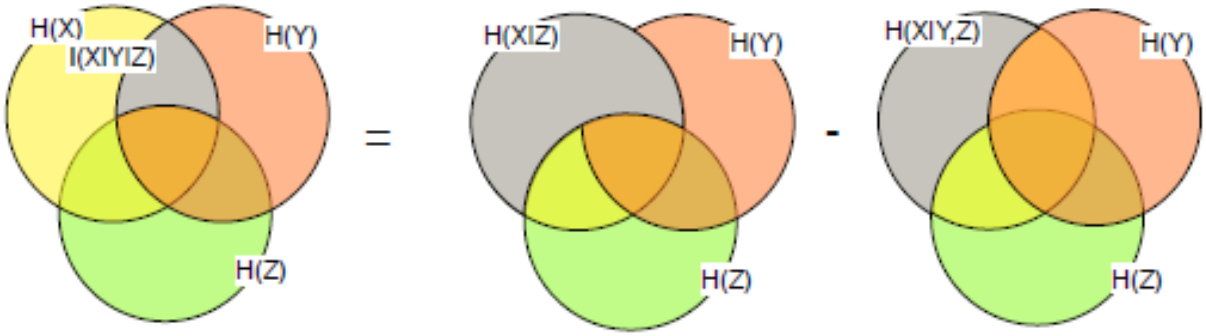

FIG. 3: Venn diagram explaining the chain rules for mutual information.

The conditional mutual information between X and Y given Z is

$$I(X;Y|Z) \doteq S(X|Z) - S(X|Y,Z) = \mathbb{E}_{P(x,y,z)}\left[\log \frac{P(x,y|z)}{P(x|z)P(y|z)}\right] \tag{.38}$$

The entropy chain rule (.31) when applied to the mutual information of $(X_1, \cdots, X_n) \sim P(x_1, \cdots, x_n)$ results in

$$I(X_n, \cdots, X_1; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \cdots, X_1) \tag{.39}$$

See Fig. (3) for the Venn diagram illustration of Eq. (.39).

See [1] for extra discussions on entropy, mutual information and related.

3. *Information Channel*

Information channel, $X \to Y$, through the channel, $P(y|x)$. Information Channel Capacity is

$$C \doteq \max_{p(x)} I(X;Y). \tag{.40}$$

Main - Shannon – theorem of the information theory (channel coding): Maximum rate at which we can communicate reliably over the channel is the information channel capacity C. More at the recitation.

4. *Probabilistic Inequalities for Entropy and Mutual Information*

Jensen's inequality. Let $f(X)$ be a convex function then

$$\mathbb{E}\left[f(X)\right] \geq f\left(\mathbb{E}\left[X\right]\right). \tag{.41}$$

Here convexity of $f(x)$ on an interval $[a,b]$ means (reminder):

$$f(\lambda u + (1-\lambda)v) \leq \lambda f(u) + (1-\lambda)f(v), \quad \forall u, v \in [a,b], \quad 0 < \lambda < 1 \tag{.42}$$

See Fig. (4) with the hint on the proof of the Jensen inequality.

Consequences of the Jensen inequality (for entropy and mutual information):

- (Information Inequality)

$$D(p\|q) \geq 0, \quad \text{with equality iff} p = q$$

- (conditioning reduces entropy)

$$S(X|Y) \leq S(X) \quad \text{with equality iff } X \text{ and } Y \text{ are independent}$$

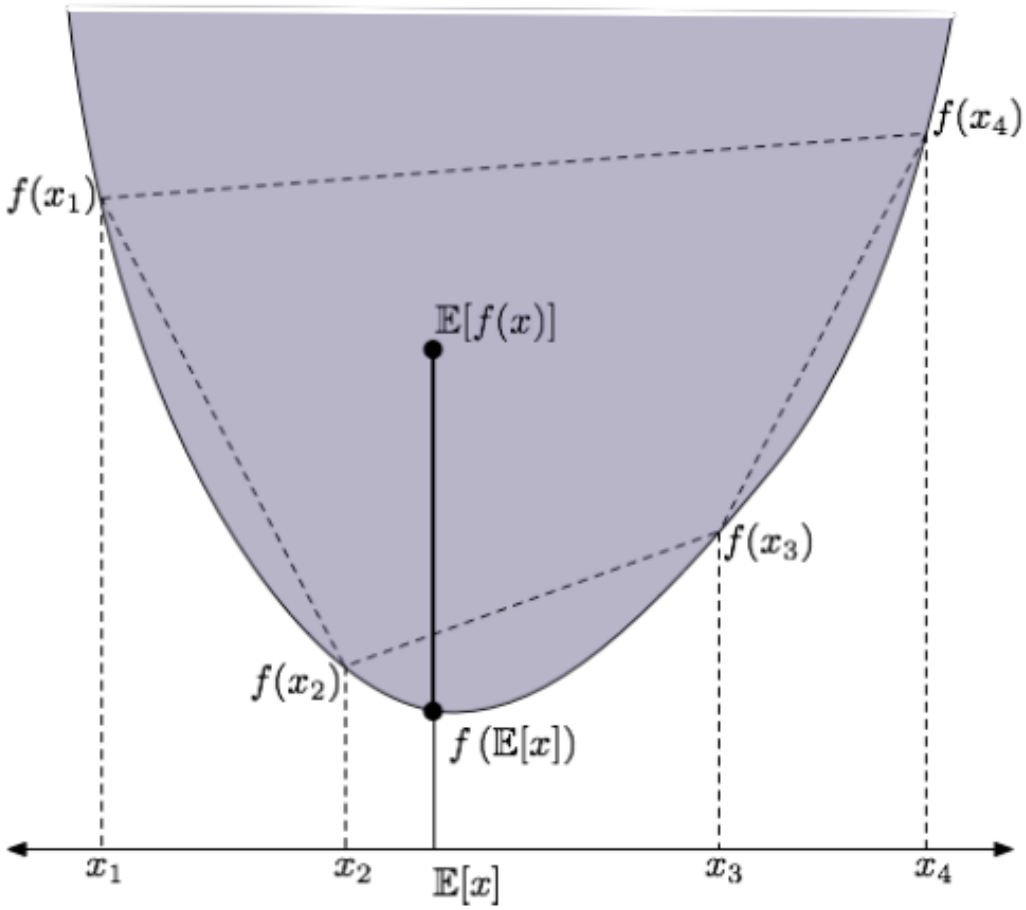

FIG. 4:

- (Independence Bound on Entropy)

$$S(X_1, \cdots, X_n) \leq \sum_{i=1}^{n} S(X_i) \quad \text{with equality iff } X_i \text{are independent}$$

Another useful inequality [Log-Sum Theorem]

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}, \tag{.43}$$

with equality iff a_i/b_i is constant. Convention: $0 \log 0 = 0$, $a \log(a/0) = \infty$ if $a > 0$ and $0 \log 0/0 = 0$. Consequences of the Log-Sum theorem

- (Convexity of Relative Entropy) $D(p\|q)$ is convex in the pair p and q
- (Concavity of Entropy) For $X \sim p(x)$ we have $S(P) \doteq S_P(X)$ (notations are extended) is a concave function of $P(x)$.
- (Concavity of the mutual information in $P(x)$) Let $(X, Y) \sim P(x, y) = P(x)P(y|x)$. Then $I(X; Y)$ is a concave function of $P(x)$ for fixed $P(y|x)$.

- (Concavity of the mutual information in $P(y|x)$) Let $(X, Y) \sim P(x, y) = P(x)P(y|x)$. Then $I(X; Y)$ is a concave function of $P(y|x)$ for fixed $P(x)$.

We will see later (studying Graphical Models) why the convexity/concavity properties are useful.

5. Recitation. Entropy, Mutual Information and Probabilistic Inequalities

---

[1] D. J. C. Mackay, *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press, 2003. [Online]. Available: http://www.inference.phy.cam.ac.uk/itprnn/book.html