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

1. Entropy.

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

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

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 (logarithm and not some other function) because it is **additive** for independent sources.

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

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

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. (See the accompanied ijulia file.)

Entropy (I.27), $S(X)$, has the following properties (some of these 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 ascending order with respect to 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

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

Conditional entropies are

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

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

The so-called chain rule states (check)

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

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) as follows

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

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

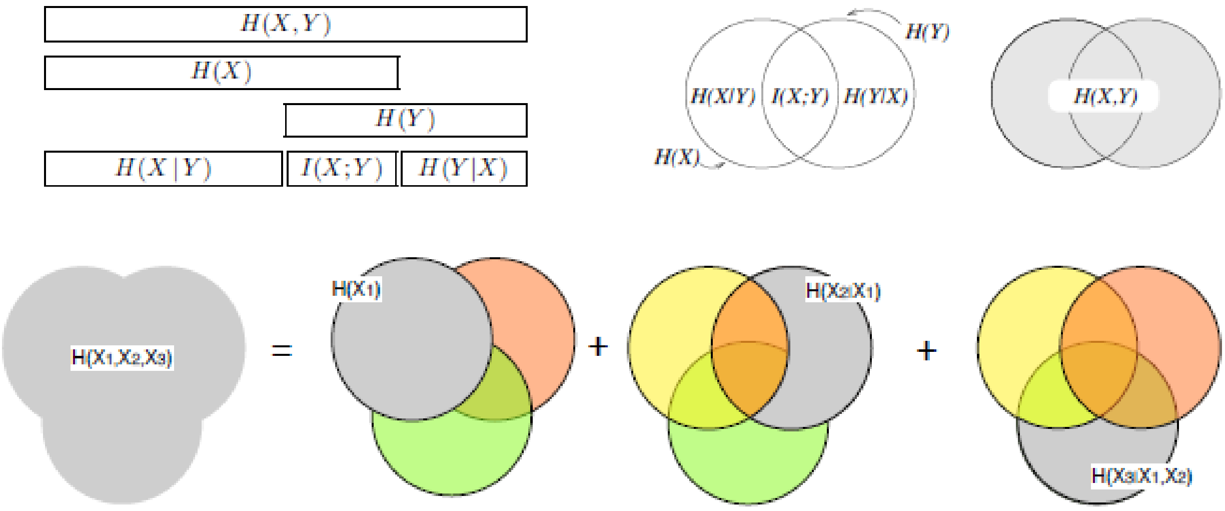

FIG. 1: Venn diagram(s) explaining the chain rule for computing multivariate entropy.

2. *Independence/Dependence. Mutual Information.*

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

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

Note that the KL difference is not symmetric wrt exchange of the order of the two distributions. Moreover it is not a proper metric of comparison as it does not satisfy the so-called triangle inequality. Any proper metric of a space should be a) positive, b) zero when comparing identical states; c) symmetric, and d) satisfy the triangle inequality, $d_(a,b) \leq d_{a,c} + d_{bc}$. The last two conditions do not hold in the case of the KL divergence. However, an infinitesimal version of KL divergence - Hessian of the KL distance around its minimum, also called Fisher information, constitutes a proper metric.

Exercise: Show that $D(p_1\|p_2) \geq 0$ and that minimum of the functional is achieved at $p_1 = p_2$.

Comparing the two information sources, say tracking events x and y, one (and rather dramatic) assumption may be that 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 introduced as the measure of dependence

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

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).

Mutual information is obviously related to entropies,

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

The relation is illustrated in Fig. (2). Mutual Information also possesses the following properties

$$I(X;Y) = I(Y;X) \text{ (symmetry)} \tag{I.36}$$

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

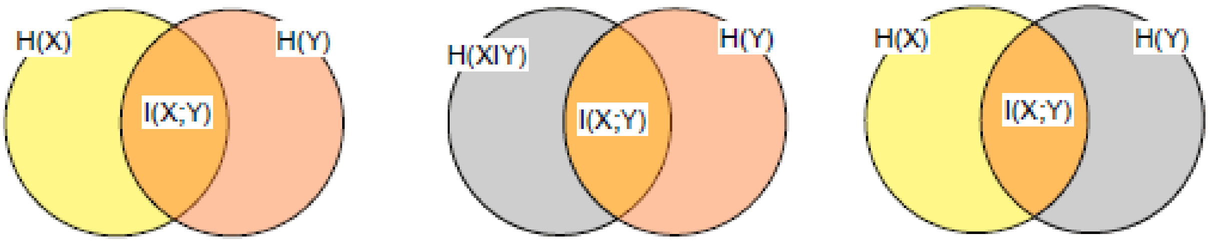

FIG. 2: Venn diagram explaining relations between mutual information and entropies.

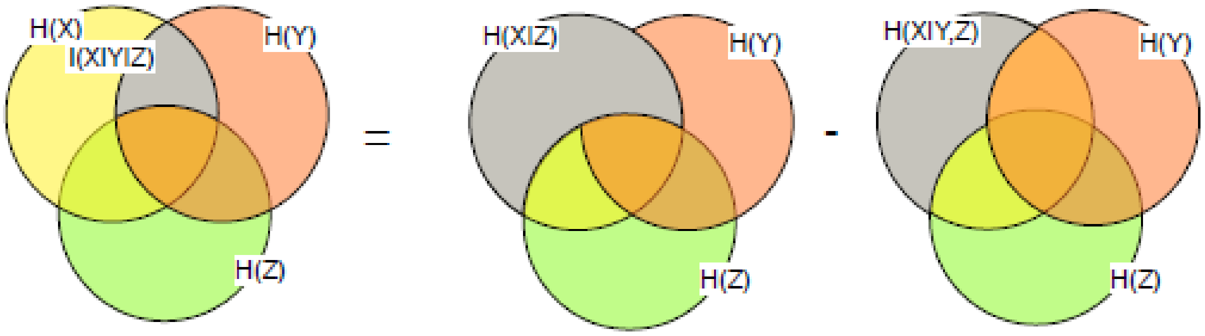

FIG. 3: Venn diagram explaining the chain rules for mutual information.

The conditional mutual information between X and Y given Z is

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

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

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

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

See [2] for extra discussions on entropy, mutual information and related.

3. *Information Channel*

A (noisy) information channel is described through the input ⇒ output map, $X \to Y$. The conditional probability, $P(y|x)$, describes the information channel. An important characteristic is the Channel Capacity

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

Main theorem of the information theory, the channel coding theorem of Shannon, states: Maximum rate at which we can communicate reliably over the channel is the information channel capacity C. More details on this and related concepts will be provided at the recitation linked to this lecture.

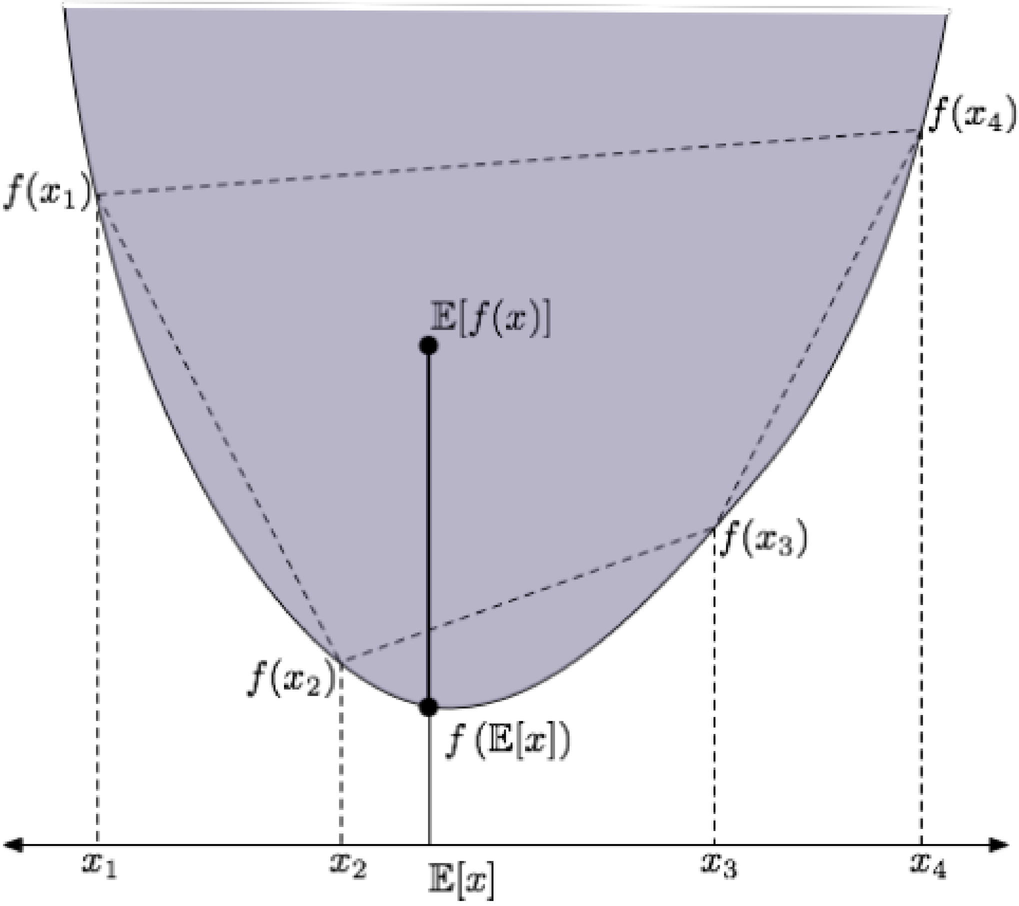

FIG. 4:

4. Probabilistic Inequalities for Entropy and Mutual Information

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

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

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

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

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)

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

- (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]

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

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 (discussing Graphical Models) why the convexity/concavity properties are useful.

5. *Recitation. Entropy, Mutual Information and Probabilistic Inequalities*

II. THEME # 2. STOCHASTIC PROCESSES

A. Lecture #4: Markov Chains [discrete space, discrete time].

1. *Transition Probabilities*

So far we have studied random variables and events often assuming that these are i.i.d. = independent identically distributed. However, in real world we ”jump” from one random state to another so that the consecutive states are dependent. The memory may last for more than one jump, however there is also a big family of interesting random processes which do not have long memory - only current state influences where we jump to. This is the class of random processes described by Markov Chains (MCs).

MCs can be explained in terms of directed graphs, $\mathcal{G}=(\mathcal{V},\mathcal{E})$, where the set of vertices, $\mathcal{V}=(i)$, is associated with the set of states, and the set of directed edges, $\mathcal{E}=(j\leftarrow i)$, correspond to possible transitions between the states. Note that we may also have ”self-loops”, $(i\leftarrow i)$ included in the set of edges. To make description complete we need to associate to each vertex a transition probability, $p_{j\leftarrow i}=p_{ji}$ from the state i to the state j. Since p_{ji} is the probability, $\forall(j\leftarrow i)\in\mathcal{E}:\quad p_{ji}\geq 0$, and

$$\forall i:\quad\sum_{j:(j\leftarrow i)\in\mathcal{E}}p_{ji}=1.\tag{II.1}$$

Then, the combination of $\mathcal{G}$ and $p\doteq(p_{ji}|(j\leftarrow i)\in\mathcal{E})$ defines a MC. Mathematically we also say that the tuple (finite ordered set of elements), $(\mathcal{V},\mathcal{E},p)$, defines the Markov chain. We will mainly consider in the following stationary Markov chains, i.e. these with p_{ji} constant - not changing in time. However, for many of the following statements/considerations generalization to the time-dependent processes is straightforward.

MC generates a random (stochastic) dynamic process. Time flows continuously, however as a matter of convenient abstraction we consider discrete times (and sometimes, actually quite often, events do happen discreetly). One uses $t=0,1,2,\cdots$ for the times when jumps occur. Then a particular random trajectory/path/sample of the system will look like

$$i_1(0),i_2(1),\cdots,i_k(t_k),\quad\text{where}\quad i_1,\cdots,i_k\in\mathcal{V}$$

We can also generate many samples (many trajectories)

$$n=1,\cdots,N:\quad i_1^{(n)}(0),i_2^{(n)}(1),\cdots,i_k^{(n)}(t_k),\quad\text{where}\quad i_1,\cdots,i_k\in\mathcal{V}$$

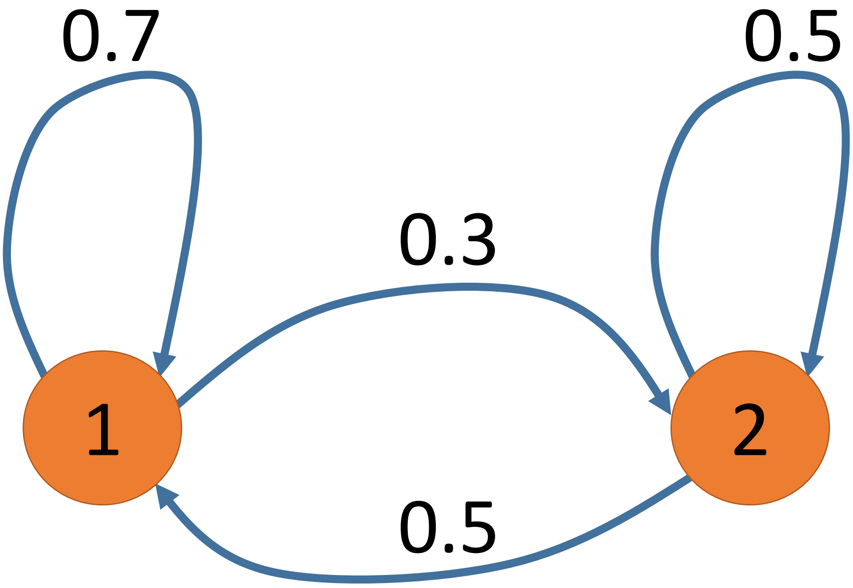

FIG. 5: Markov Chain (MC) - Example #1.

where N is the number of trajectories.

How does one relates the directed graph with weights (associated to the transition probabilities) to samples? The relation, actually, has two sides. The direct one - is about how one generates samples. The samples are generated by advancing the trajectory from the current-time state flipping coin according to the transition probability p_{ij}. The inverse side is about reconstructing characteristics of Markov chain from samples or verifying if the samples where indeed generated according to (rather restrictive) MC rules.

Now let us get back to the direct problem where a MC is described in terms of $(\mathcal{V}, \mathcal{E}, p)$. However, instead of characterizing the system in terms of the trajectories/paths/samples, we can pose the question following evolution of the “state probability vector”, or simply the ”state vector”:

$$\forall i \in \mathcal{V}, \quad \forall t = 0, \cdots : \quad \pi_i(t+1) = \sum_{j:(i \leftarrow j) \in \mathcal{E}} p_{ij} \pi_j(t). \tag{II.2}$$

Here, $\pi(t) \doteq (\pi_i(t) \geq 0 | i \in \mathcal{V})$ is the vector built of components each representing probability for the system to be in the state i at the moment of time t. Thus, $\sum_{i \in \mathcal{V}} \pi_i = 1$. We can also rewrite Eq. (II.2) in the vector/matrix form

$$\pi(t+1) = p\pi(t), \tag{II.3}$$

where $\pi(t)$ the column/state and $p(t)$ is the transition-probability matrix, which satisfies the so-called ”stochasticity” property (II.1). Sequential application of Eq. (II.3) results in

$$\pi(t+k) = p^k \pi(t), \tag{II.4}$$

and we are interested to analyze properties of p^k, characterizing the Markov chain acting for k sequential periods.

Let us first study it on example of the simple MC illustrated in Fig. (5). In this case, call it example #1, p^k is 2×2 matrix which dependence on k as follows

$$p^1 = \begin{pmatrix} 0.7 & 0.5 \\ 0.3 & 0.5 \end{pmatrix}, \quad p^2 = \begin{pmatrix} 0.64 & 0.6 \\ 0.36 & 0.4 \end{pmatrix}, \quad p^{10} \approx p^{100} \approx \begin{pmatrix} 0.625 & 0.625 \\ 0.375 & 0.375 \end{pmatrix}. \tag{II.5}$$

2. *Properties of Markov Chains*

The MC is **irreducible** if one can access any state from any state, formally

$$\forall i,j \in \mathcal{V}: \quad \exists k > 1, \quad \text{s.t.} \quad (p^k)_{ij} > 0. \tag{II.6}$$

Example #1 is obviously irreducible. However, if we replace $0.3 \to 0$ and $0.7 \to 1$ the MC becomes reducible – state 1 is not accessible from 2.

A state i has period k if any return to the state must occur in multiples of k. If $k = 1$ than the state is **aperiodic**. MC is **aperiodic** if all states are aperiodic. An irreducable MC only needs one aperiodic state to imply all states are aperiodic. Any MC with at least one self-loop is aperiodic. Example #1 is obviously aperiodic. However, it becomes periodic with period two if the two self-loops are removed.

A state i is said to be transient if, given that we start in state i, there is a non-zero probability that we will never return to i. State i is recurrent if it is not transient. State i is **positive-recurrent** if the expected return time (to the state) is positive (this feature is important for infinite graphs).

A state is **ergodic** if the state is aperiodic and positive-recurrent. If all states in an irreducible MC are ergodic then the MC is said to be ergodic. A MC is ergodic if there is a finite number k_* such that any state can be reached from any other state in exactly k_* steps. For the example #1 $k_* = 2$. Note, that there are other (alternative) descriptions of ergodicity. Thus most intuitive one is: the MC is ergodic if it is irreducable and aperiodic. In this course we will not dwell much on the rich mathematical formalities and details, largely considering generic ergodic MC.

Practical consequence of ergodicity is that the steady state is unique and universal. Universality means that the steady state does not depend on the initial condition.

3. *Steady State Analysis*

Component-wise positive, normalized, π^*, is called stationary distribution (invariant measure) if

$$\pi^* = p\pi^* \tag{II.7}$$

An irreducible MC has a stationary distribution iff all of its states are positive recurrent. Solving Eq. (II.7) for the example # 1 of Eq. (II.5) one finds

$$\pi^* = \begin{pmatrix} 0.625 \\ 0.375 \end{pmatrix}, \tag{II.8}$$

which is naturally consistent with Eq. (II.5). In general,

$$\pi^* = \frac{e}{\sum_i e_i}, \tag{II.9}$$

where e is the eigenvector with the eigenvalue 1. And how about other eigenvalues of the transition matrix?

4. *Spectrum of the Transition Matrix & Speed of Convergence to the Stationary Distribution*

Assume that p is diagonalizable (has $n = |p|$ linearly independent eigenvectors) then we can decompose p according to the following eigen-decomposition

$$p = U^{-1}\Sigma U \tag{II.10}$$

where $\Sigma = \text{diag}(\lambda_1, \cdots, \lambda_n)$, $1 = |\lambda_1| > \lambda_2 \geq |\lambda_3| \geq \cdots |\lambda_n|$ and U is the matrix of eigenvectors (each normalized to having an l_2 norm equal to 1) where each raw is a right eigenvector of p. Then

$$\pi^{(k)} = p^k\pi = (U^{-1}\Sigma U)^k\pi_0 = U^{-1}\Sigma^k U\pi_0. \tag{II.11}$$

Let us represent p_0 as an expansion over the normalized eigenvectors, $u_i, \cdots i = 1, \cdots, n$:

$$\pi = \sum_{i=1}^{n} a_i u_i. \tag{II.12}$$

Taking into account orthonormality of the eigenvectors one derives

$$\pi^{(k)} = \lambda_1 \left(a_1 u_1 + a_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k u_2 + \cdots + a_n \left(\frac{\lambda_n}{\lambda_1}\right)^k u_n \right) \tag{II.13}$$

Since $\pi^{(k)}_{k\to\infty} \to \pi^* = u_1$, the second term on the rhs of Eq. (II.13) describes the rate of convergence of $\pi^{(k)}$ to the steady state. The convergence is exponential in $\log(\lambda_1/\lambda_2$.

5. *Reversible & Irreversible Markov Chains.*

MC is called **reversible** if there exists π s.t.

$$\forall \{i,j\} \in \mathcal{E}: \quad p_{ji}\pi_i^* = p_{ij}\pi_j^*, \tag{II.14}$$

where $\{i,j\}$ is our notation for the undirected edge, assuming that both directed edges $(i \leftarrow j)$ and $(j \leftarrow i)$ are elements of the set $\mathcal{E}$. In physics this property is also called **Detailed Balance** (DB). If one introduces the so-called ergodicity matrix

$$Q \doteq (Q_{ji} = p_{ji}\pi_i^* | (j \leftarrow i) \in \mathcal{E}), \tag{II.15}$$

then DB translates into the statement that Q is symmetric, $Q = Q^T$. The MC for which the property does not hold is called **irreversible**. $Q - Q^T$ is nonzero, i.e. Q is asymmetric for reversible MC. An asymmetric component of Q is the matrix built from currents/flows (of probability). Thus for the case #1 shown in Fig. (5)

$$Q = \begin{pmatrix} 0.7 * 0.625 & 0.5 * 0.375 \\ 0.3 * 0.625 & 0.5 * 0.375 \end{pmatrix} = \begin{pmatrix} 0.4375 & 0.1875 \\ 0.1875 & 0.1875 \end{pmatrix} \tag{II.16}$$

Q is symmetric, i.e. even though $p_{12} \neq p_{21}$, there is still no flow of probability from 1 to 2 as the "population" of the two states, π_1^* and π_2^* respectively are different, $Q_{12} - Q_{21} = 0$. In fact, one observes that in the two node situation the steady state of the MC is always in DB.

6. *Detailed Balance vs Global Balance. Adding cycles to accelerate mixing.*

Note that if a steady distribution, π^*, satisfy the DB condition (II.14) for a MC, $(\mathcal{V}, \mathcal{E}, p)$, it will also be a steady state of another MC, $(\mathcal{V}, \mathcal{E}, \tilde{p})$, satisfying the more general Balance (or global balance) B-condition

$$\sum_{j:(j\leftarrow i)\in\mathcal{E}} \tilde{p}_{ji}\pi_i^* = \sum_{j:(i\leftarrow j)\in\mathcal{E}} \tilde{p}_{ij}\pi_j^*. \tag{II.17}$$

This suggests that many different MC (many different dynamics) may result in the same steady state. Obviously DB is a particular case of the B-condition (II.17).

The difference between DB- and B- can be nicely interpreted in terms of flows (think water) in the state space. From the hydrodynamic point of view reversible MCMC corresponds to irrotational probability flows, while irreversibility relates to nonzero rotational part, e.g. correspondent to vortices contained in the flow. Putting it formally, in the irreversible case antisymmetric part of the ergodic flow matrix, $Q = (\tilde{p}_{ij}\pi_j^* | (i \leftarrow j))$, is nonzero and it actually allows the following cycle decomposition,

$$Q_{ij} - Q_{ji} = \sum_\alpha J_\alpha \left(C_{ij}^\alpha - C_{ji}^\alpha \right) \tag{II.18}$$

where index enumerates cycles on the graph of states with the adjacency matrices C^α. Then, J_α stands for the magnitude of the probability flux flowing over cycle .

One can use the cycle decomposition to modify MC such that the steady distribution stay the same (invariant). Of course, cycles should be added with care, e.g. to make sure that all the transition probabilities in the resulting $\tilde{p}$, are positive (stochasticity of the matrix will be guaranteed by construction). The procedure of "adding cycles" along with some additional tricks (e.g. the so-called lifting/replication) may help to improve mixing, i.e. speed up convergence to the steady state — which is a very desirable property for sampling π^* efficiently. This and other features of MC will be discussed in details on a three node example at the recitation linked to the lecture.