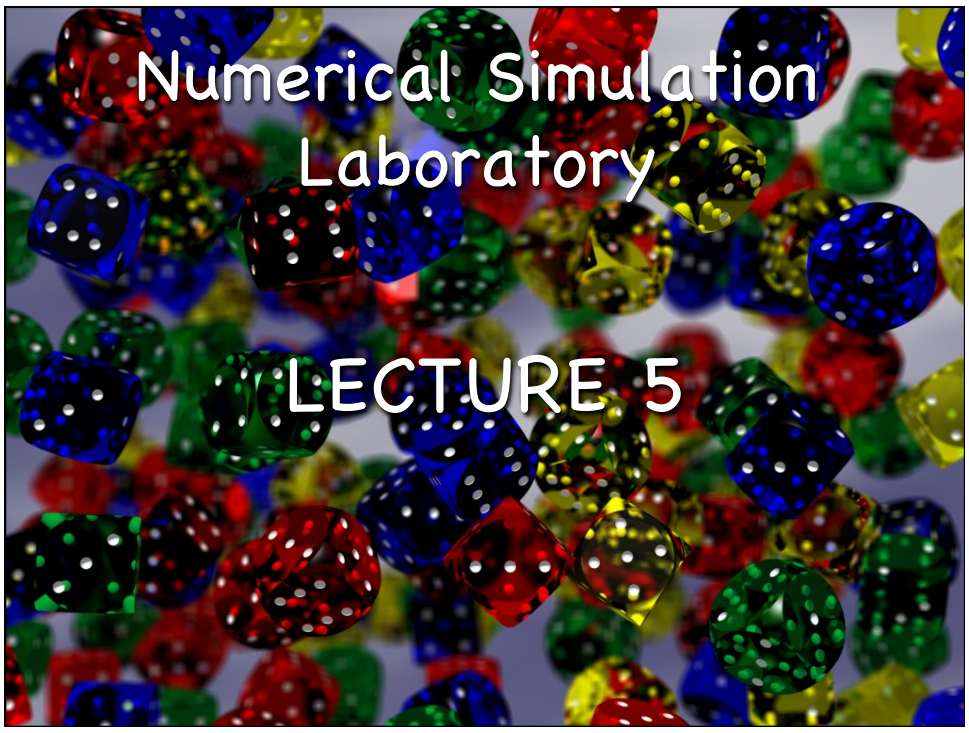


Lecture 5 - outline

- Memory, power laws & scaling
- Power spectral density
- White, Brownian & Pink noise
- Time series models: AR(p), ARCH, GARCH, MA(q), ARMA(p,q)
- Metropolis algorithm

Memory

- Memory is a term that often appears in discussions about complex dynamics. What is it meant when a scientist says that a certain complex system exhibits memory? Memory in a complex system is the ability of past events to act as an influence on future dynamics, that extends in time in a scale-invariant fashion
- As introduced before, the covariance between two random processes, $x_i(t)$ and $x_j(t)$, is defined as:

$$cov_{[x_i,x_j]} := \left\langle x_i(t_1)x_j(t_2)\right\rangle - \left\langle x_i(t_1)\right\rangle\left\langle x_j(t_2)\right\rangle$$

- The autocorrelation function of a stationary random process, x, is in essence the statistical covariance of the process with a time-delayed copy of itself. It is defined, for any time lag $\tau > 0$, as:

$$Ac_{[x]}(\tau) := \frac{cov_{[x,x]}(t,t+\tau)}{\sqrt{cov_{[x,x]}(t,t)}\sqrt{cov_{[x,x]}(t+\tau,t+\tau)}} = \frac{cov_{[x,x]}(t,t+\tau)}{cov_{[x,x]}(t,t)} = \frac{\left\langle x(t)x(t+\tau)\right\rangle - \left\langle x(t)\right\rangle\left\langle x(t+\tau)\right\rangle}{\sigma_x^2}$$

- The denominator reduces to the variance of the process at time t. Thanks to the assumed stationarity of the process, the autocorrelation is independent of t.

Power laws and scaling

- You'll often hear that power laws are scale-free, scale-invariant. What does this statement mean?
- To understand this point, let us compare the functions:

$$f_1(x) = e^{x/x_0} \qquad f_2(x) = \left(x/x_0\right)^p$$

- The ratio of the largest value to the smallest value of f_1 in the ranges $0.5x_0 \le x \le 2x_0$, $5x_0 \le x \le 20x_0$, and $50x_0 \le x \le 200x_0$ are, respectively

$$f_1(2x_0)/f_1(0.5x_0) = e^2/e^{1/2} = e^{3/2} \qquad f_1(20x_0)/f_1(5x_0) = e^{15} \qquad f_1(200x_0)/f_1(50x_0) = e^{150}$$

- In contrast, the ratio of the largest value to the smallest value of f_2 in the same three ranges are identical and each equal to $(4)^p$
- Consequently, the three graphs of $f_2(x)$ over these three ranges can be superimposed by a simple change of scale but the same superposition cannot be achieved in the case of $f_1(x)$. In other words, power-laws look the same no matter on what scale one looks at them

Memory and Statistical Dependence

- The autocorrelation is useful to investigate the presence of memory or, more precisely, statistical dependence in a process because it inherits from the covariance the following properties:

1. If $Ac_{[x]}(\tau)=0$ for all $\tau > 0$, it means that all the values of the process are independent of each other. That is, the process lacks any kind of memory.
2. If $Ac_{[x]}(\tau)=0$ for $\tau > \tau_c$, it implies that the process becomes independent of its past history only after a lapse of time τ_c , or that the process has no memory about itself beyond a lapse τ_c , that provides a characteristic memory timescale.
3. If $Ac_{[x]}(\tau) \approx \tau^{-a}$ $a > 0$; $\tau >> 0$, one might suspect that memory remains present in the system for all times in a self-similar manner, as made apparent by the power-law dependence, thus lacking any memory timescale.

- In practice, things are never this clear-cut. For instance, the autocorrelation is almost never exactly zero. Therefore, one has to decide which threshold value is a reasonable one, below which one can safely consider that autocorrelation as negligible

Typical Autocorrelation Tails

- The exact form of the tail of $Ac_{[x]}(\tau)$ will depend on the specifics of the process at hand. However, there are a couple of shapes that illustrate well what is often found in many practical situations:

- Exponential tails: Let's consider first the model autocorrelation function

$$Ac_{[x]}(\tau) = \exp\left(-\left|\tau\right|/\tau_c\right)$$

This function is always positive, thus corresponding to a process that presents positive correlations for all lags. The important quantity here is τ_c , that provides an estimate for how long the process remembers about itself

- Power-law tail: Another type of tail that is often encountered is the decaying power-law. As shown before power-laws are indicative of scale invariance.

$$Ac_{[x]}(\tau) = B \mid \tau \mid^{-a} \qquad a > 0 \qquad \tau \to \infty$$

The Power Spectrum/Power spectral density

- The power spectrum or power spectral density (PSD) of an stationary random process x(t) is defined as the Fourier transform of its autocorrelation function $S_{[x]}(\omega) := \int_{-\infty}^{\infty} d\tau\, e^{i\omega\tau} Ac_{[x]}(\tau)$
- The PSD can be estimated directly from the process x(t) computing its Fourier transform $X_T(\omega)$ on the finite interval [0,T], one can show that the following is equivalent to the former definition:

$$S_{[x]}(\omega) = \lim_{T\to\infty} \left|X_T(\omega)\right|^2 \Big/ T$$

- $S_{[x]}(\omega)$ of a stochastic process x(t) thus describes the distribution of power into frequency components composing that process.
- $S_{[x]}(\omega)$ can be generalized to discrete time variables x_n. As above we can consider a finite window of $1 \le n \le N$, up to $T = N\Delta t$. Then a single estimate of the PSD can be obtained through:

$$S_{[x]}\left(\omega_j = 2\pi j/N\right) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x_n e^{-i\omega_j n}\right|^2 \qquad j = 0, \cdots, N-1$$

- Typically one would average this $S_{[x]}(\omega)$ over many trials to obtain a more accurate estimate

7

Typical Power Spectra Shapes

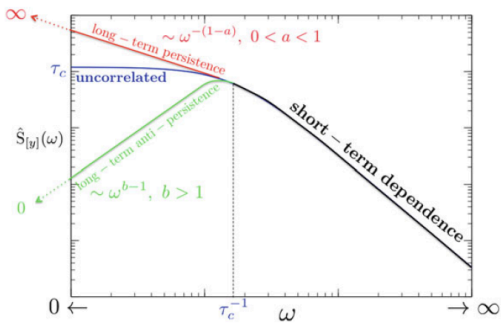
8

- Exponential tail: In the case of the exact exponential, its power spectrum can be computed analytically. The result is:

$$S_{[x]}(\omega) := \int_{-\infty}^{\infty} d\tau\, e^{i\omega\tau} e^{-|\tau|/\tau_c} = \frac{2\tau_c}{1+\tau_c^2\omega^2}$$

that decays as ω^{-2} for large frequencies (i.e., $\omega >> \tau_c^{-1}$) and becomes flat for smaller frequencies (i.e., $\omega << \tau_c^{-1}$).

The break-point between the two regions happens at $\omega \simeq \tau_c^{-1}$ (see in Fig. the blue curve), that coincides with the inverse of the memory timescale of the process

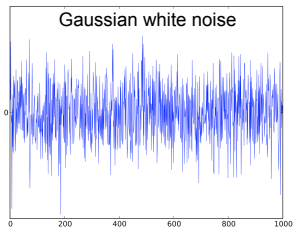


- The PSD will become eventually flat for frequencies $\omega << \tau_c^{-1}$ also for any other autocorrelation function that is not exactly exponential but that decays sufficiently fast at long times.

White noise

- The association of flatness and lack of correlations also holds for any other ω interval, even if they do not extend for arbitrarily low frequencies, but are instead restricted to a limited frequency range
- In the limit $\tau_c \to 0$, we have a random process that is uncorrelated for all lags, the flat region naturally extends from $\omega \to 0$ towards $\omega \to \infty$, thus filling the whole spectrum. White noise is a random signal having equal intensity at different frequencies, giving it a constant PSD
- Any distribution of values with zero mean is possible. Even a binary signal which can only take on the values 1 or −1 will be white if the sequence is statistically uncorrelated. Starting from a Bernulli process, it is easy to generate a binary white noise:

$$W_n^b = 2(x_n - 0.5) \qquad x_n = \begin{cases} 1 & \text{with prob. } 1/2 \\ 0 & \text{with prob. } 1/2 \end{cases}$$

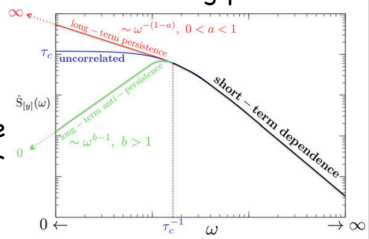


- Noise having a continuous distribution can of course be white. A Gaussian white noise is obtained with

$$W_n^G = x_n \qquad x_n \sim N(0,1)$$

- The behavior of the PSD for $\omega >> \tau_c^{-1}$ will depend on the details of the autocorrelation function for $\tau << \tau_c$. If it is not exactly exponential, the $S_{[x]}(\omega)$ could decay either faster or slower than ω^{-2}. Sometimes, it may even exhibit several scaling regions, depending on how complicated the $Ac_{[x]}(\tau)$ structure is.
- Power-law tail: In the case of an $Ac_{[x]}(\tau)$ that decays with a power law tail, with exponent in (0,1], one can invoke a property of Fourier transforms to find out the asymptotic behavior of $S_{[x]}(\omega)$. Namely, that if $f(t) \approx t^{-a}$ with $0 < a \leq 1$ when $t \to \infty$, it then follows that its Fourier transform $F(\omega) \approx \omega^{-(1-a)}$ for $\omega \to 0$. Therefore, the PSD we are looking for asymptotically decays according to the decreasing power law,

$$S_{[x]}(\omega) \approx \frac{1}{\omega^{1-a}} \qquad \omega \to 0$$



The spectrum now diverges for $\omega \to 0$. A finite memory time does not exist as is expected for processes with long-term persistence.

- To conclude, any region in $S_{[x]}(\omega)$ that behaves like a decreasing power-law $1/\omega^c$ with $c<2$ can be related to the action of positive memory-like correlations over those frequencies.

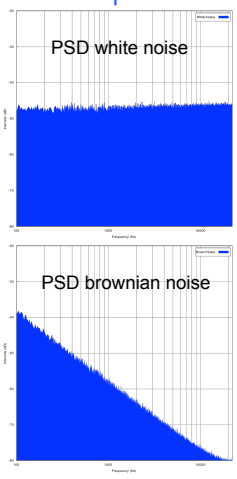
Brownian (brown/red) noise

- Consider in the following a fixed time step dt, from its definition it is clear that the increments of a Wiener process sample a Gaussian white noise process

$$\Delta W = W(t_{i+1}) - W(t_i) \approx \mathcal{N}(0,dt) \qquad dt = t_{i+1} - t_i$$

- Taking the limit dt→0 one can formally obtain the Wiener process as the integral of a white noise signal:

$$W(t) = \int_0^t \frac{dW(\tau)}{d\tau} d\tau$$



  - Brownian noise is the kind of signal noise produced by Brownian motion (which is not a stationary process!), hence its alternative name of random walk noise
  - It is possible to compute the PSD of the Brownian noise, which turns out to be a power law decay with exponent -2 in the whole spectrum:

$$S_{[W]}(\omega) = S_0 / \omega^2$$

being S_0 the constant PSD of the white noise.

Pink noise

- Pink noise or 1/f noise is a 1D stochastic process (with memory) with a frequency spectrum such that the power spectral density is inversely proportional to the frequency of the signal. The name arises from the pink appearance of visible light with this power spectrum. This is in contrast with white noise which has equal intensity per frequency interval.

- Within the scientific literature the term pink noise is sometimes used a little more loosely to refer to any noise with a PSD of the form

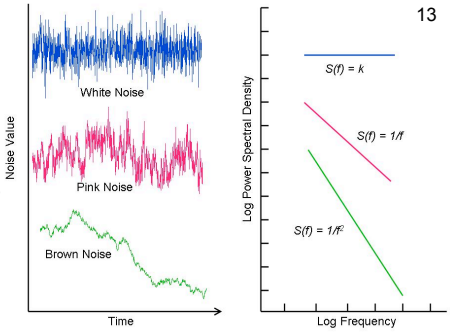
$$S_{[P]}(\omega) \propto \frac{1}{\omega^a} \qquad 0 < a < 2$$

- These pink-like noises occur widely in nature and are a source of considerable interest in many fields. In the past years, pink noise has been discovered in the statistical fluctuations of an extraordinarily diverse number of physical and biological complex systems.

- We have seen that brownian (brown/red) noise, or random walk noise,

$$S_{[P]}(\omega) \propto \frac{1}{\omega^2}$$

- One may ask why this should be less interesting than a spectrum with an exponent equal to 1. In itself, a genuine exponent equal to 2 is certainly interesting.



- The point, though, is that in most cases these spectra are expected to be the high-frequency limit of some stochastic process with a spectrum like

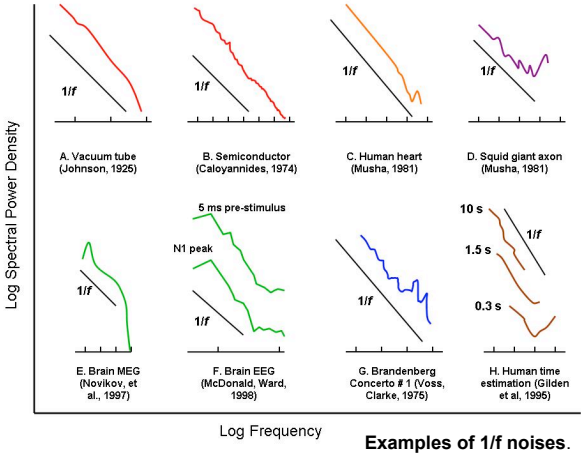
$$S_{[P]}(\omega) \propto \frac{1}{1+\left(\omega/\omega_c\right)^2}$$

as we have seen, this is in fact the case when no correlations exist in the long-time/low-frequency limit, due to an autocorrelation function that decays exponentially in time.

- This is the reason for the prejudice that considers $1/\omega^2$ spectra to be less exciting than $1/\omega$ spectra. But this is only a fair judgment when the $1/\omega^2$ behavior persists down to very small frequencies

- One should expect that the $1/\omega^a$ behavior must eventually be cut-off for frequencies below a certain frequency ω_{cut} ...

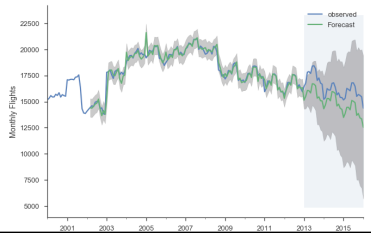
... but the interesting point is that this frequency in pink noise processes often corresponds to time scales many orders of magnitude longer than the time scale of the individual microscopic processes of the system. This is an indication that such behavior should originate in complex systems as a specific emerging phenomena (like in the so-called self-organized criticality)

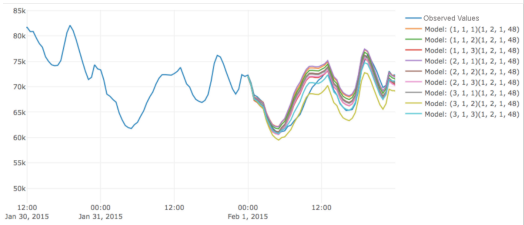


Examples of 1/f noises. Curves are illustrative based on data from the indicated sources. Adjacent pairs of tick marks on the horizontal axis beneath each figure indicate one decade of frequency.

- There are many theories of the origin of pink noise. Some theories attempt to be universal and remain a matter of current research interest.

- The main aim of time series modeling is to collect and study the past observations of a time series to develop an appropriate model which describes the inherent structure of the series. This model is then used to generate future values for the series, i.e. to make forecasts.
- Due to the indispensable importance of time series forecasting in numerous practical fields such as economics, finance, science and engineering, proper care should be taken to fit an adequate model to the underlying time series. I will give just a brief introduction to models in this field in order to show more examples of stochastic processes.





- **Definition**: A time series is a sequential set of data points, measured typically over successive times. It is mathematically defined as a set of values $X(t)$, $t \in \mathbb{N}, \mathbb{R}^+ (t>0)$ where t represents the time elapsed.
- The variable $X(t)$ is a random variable. The sequence of observations of the series is actually a sample realization of the stochastic process that produced it
- The measurements taken during an event in a time series are arranged in a proper chronological order. A time series can thus be continuous or discrete. For example temperature readings, rainfalls etc. can (in principle!) be recorded as a continuous time series.
- On the other hand population of a particular city, production of a company, exchange rates between two different currencies may represent discrete time series. Usually in a discrete time series the consecutive observations are recorded at equally spaced time intervals such as hourly, daily, weekly, monthly or yearly time separations.
- Time series data, however, are always finite records of the underlying stochastic process and thus can be easily transformed to a discrete time series by merging data together over a specified time interval.

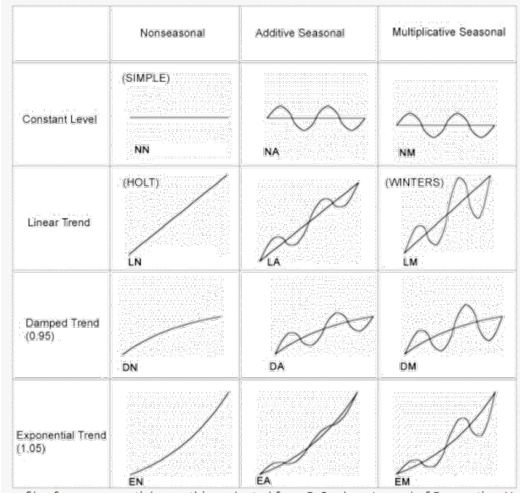
- A time series containing records of a single variable is termed as univariate. But if records of more than one variable are considered, it is termed as multivariate.
- A time series in general is supposed to be affected by some main components, which can be separated from the observed data. These components are: Trend, Cyclical/Seasonal and Irregular components.
- The general tendency of a time series to increase, decrease or stagnate over a long period of time is termed Trend: T(t)
- The cyclical/seasonal variation, C(t), in a time series describes the medium-term changes in the series, caused by circumstances, which repeat in cycles
- Irregular or random variations, R(t), in a time series are caused by unpredictable influences, which are not regular and also do not repeat in a particular pattern. These variations are caused by incidences such as war, strike, earthquake, flood, revolution, etc.
- Considering the effects of these components, two different class of models are generally used for a time series: multiplicative and additive models

Multiplicative models: $X(t) = T(t) \cdot C(t) \cdot R(t)$

Additive models: $X(t) = T(t) + C(t) + R(t)$

where X(t) are the observations

- Multiplicative model is based on the assumption that the components of a time series are not necessarily independent and they can affect one another; whereas in the additive model it is assumed that the components are independent of each other.
- These models contain parameters and in practice a suitable model is fitted to a given time series and the corresponding parameters are estimated using the known data values.



Stationarity

- There are two types of stationary processes, the first type was already encountered:
- **Definition**: A process {x(t), t=0,1,2,…} is strongly/strictly stationary if the joint probability distribution function of {x(t), …, x(t+s)} is independent of t for all s.
- However in practical applications, with a finite data set, one can never check if a process is strong stationary and so a somewhat weaker form is considered:
- **Definition**: A stochastic process is said to be weakly stationary of order k if the statistical moments of the process up to that order depend only on time differences and not upon the time of occurrences of the data being used to estimate the moments.
- For example a stochastic process {x(t), t=0,1,2,…} is second order stationary if it has time independent mean and variance and the covariance values, Cov[x(t),x(t-s)], depend only on s. It is important to note that neither strong nor weak stationarity implies the other (in fact, strong stationarity does not assure the existence of the statistical moments of order k)

- The concept of stationarity is a mathematical idea constructed to simplify the theoretical modeling of stochastic processes. To design a proper model, adequate for future forecasting, the underlying time series is expected to be stationary. Unfortunately it is not always the case.
- The greater the time span of historical observations, the greater is the chance that the time series will exhibit non-stationary characteristics. Usually time series, showing trend or seasonal patterns are non-stationary in nature. In such cases, preliminary transformations are often used to remove the trend and to make the series stationary.
- While building a proper time series model we have to consider the Occam's razor principle when face with a number of competing and adequate explanations, pick the most simple one: the more complicated the model, the more possibilities will arise for departure from the actual model assumptions. With the increase of model parameters, the risk of overfitting also subsequently increases. An over fitted time series model may describe the training data very well, but it may not be suitable for future forecasting. I will briefly discuss now the simplest models used in time series modeling.

Autoregressive models AR(p)

- Definition: It is called autoregressive model of order p, or model AR(p), the linear equation

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p} + b + \varepsilon_t$$

Memory; constant; white noise

- AR models are additive and thus they are in the form of a stochastic difference equation.
- The simplest AR process is AR(0), which has no dependence between the terms. Only the noise term contributes to the output of the process, so AR(0) corresponds to white noise.
- AR(p) processes with p>0 have memory: a one-time shock affects values of the evolving variable infinitely far into the future. For example, consider the AR(1) model, a non-zero value for $\varepsilon_{t=1}$ affects X_1 by its amount. Then it affects X_2 by the amount $a_1^2\varepsilon_1$, X_3 by the amount $a_1^3\varepsilon_1$. Thus the effect of ε_1 never ends, although if the process is stationary then the effect diminishes toward zero in the limit.

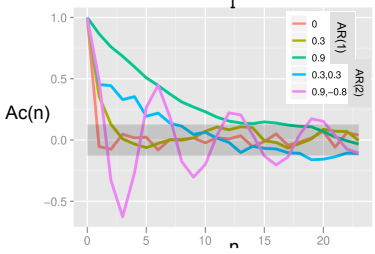
- Some parameter constraints are necessary for AR models to be stationary. For example, AR(1) models with $a_1 \geq 1$ are not stationary and, in fact, an AR(1) with $a_1=1$ and b=0 is a random walk

- For an AR(1) process with a positive a_1, only the previous term in the process and the noise term contribute to the output. If a_1 is close to 0, then the process still looks like white noise, but as a_1 approaches 1, the output gets a larger contribution from the previous term relative to the noise. This results in an "integration" of the output, similar to a low pass filter
- For an AR(2) process if both a_1 and a_2 are positive, the output will resemble a low pass filter, with the high frequency part of the noise decreased. If a_1 is positive while a_2 is negative, then the process favors changes in sign between terms of the process. The output oscillates.

- For an AR(1) stationary process, i.e. assuming $a_1<1$, the mean $E[X_t]$ is identical for all values of t by the very definition of stationarity. If the mean is denoted by μ

$$E\left[X_t\right] = a_1 E\left[X_{t-1}\right] + E\left[b\right] + E\left[\varepsilon_t\right] \qquad \mu = a_1\mu + b + 0 \Rightarrow \mu = \frac{b}{1-a_1}$$

in particular, if b=0 then μ=0



- The variance is

$$Var\left[X_t\right] = E\left[X_t^2\right] - \mu^2 = \frac{\sigma_\varepsilon^2}{1-a_1^2}$$

where σ_ε is the standard deviation of ε

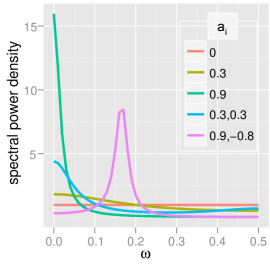
- The autocorrelation is given by

$$Ac(n) = \frac{E\left[X_{t+n}X_t\right] - \mu^2}{\mathrm{var}\left[X_t\right]} = a_1^n$$

$$\Rightarrow Ac(n) = a_1^n = e^{\ln a_1^n} = e^{n\ln a_1} = e^{-n/\tau_c}$$

$$\tau_c = -1/\ln a_1$$

Ac(n) thus decays with a decay time τ_c



ARCH/GARCH models

- An ARCH (autoregressive conditionally heteroscedastic) model is a model for the variance of a time series. ARCH models are used to describe a changing, possibly volatile (changing) variance.
- Assuming that the series has zero mean, the ARCH(1) model for the variance of model Y_t is

$$X_t = \varepsilon_t\sqrt{\alpha_0 + \alpha_1 X_{t-1}^2}$$

Gaussian white noise: N(0,1)

- We impose the constraints $\alpha_0 \geq 0$ and $\alpha_1 \geq 0$. Note that the variance at time t is connected to the value of the series at time t−1. A relatively large value of the series at time t−1 gives a relatively large value of the variance at time t.
- An ARCH(m) process is one for which the variance at time t is conditional on observations at the previous m times
- A GARCH (generalized autoregressive conditionally heteroscedastic) model uses values of the past squared observations and past variances to model the variance at time t. As an example, a GARCH(1,1) is

$$X_t = \varepsilon_t\sqrt{\alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1\sigma_{t-1}^2}$$

Other models

- Definition: It is called moving-average model of order q, or model MA(q), the linear equation

$$X_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}$$

constant — White noise — White noise memory

- in the additive MA model a shock affects X values only for the current period and q periods into the future; MA models are always stationary
- For a MA(1) model Ac(n) = 0 for n≥2 and:

$$E\left[X_t\right] = \mu \qquad E\left[X_t^2\right] - \mu^2 = \sigma_\varepsilon^2(1-\theta_1^2)$$

- The notation ARMA(p,q) refers to the models with p autoregressive terms and q moving-average terms:

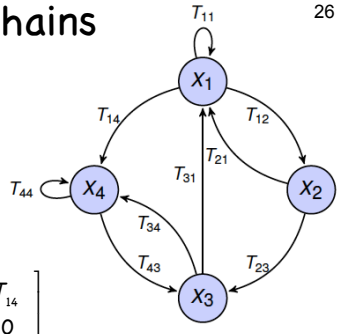
$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} a_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

- ARMA is appropriate when a system is a function of a series of unobserved shocks (the MA part) as well as its own behavior (the AR part). For example, Stock prices may be shocked by fundamental information as well as exhibiting effects due to market participants

Discrete Markov chains

- A discrete Markov chain can be represented via a graph. The nodes in the graph represent the possible states, x_i , of the system, the arrows represent the possible transitions with the relative transition probabilities, T_{ij}
- Of course, we can associate a transition matrix, which in our example turns out to be:

$$T = \begin{bmatrix} T_{11} & T_{12} & 0 & T_{14} \\ T_{21} & 0 & T_{23} & 0 \\ T_{31} & 0 & 0 & T_{34} \\ 0 & 0 & T_{43} & T_{44} \end{bmatrix}$$



T_{ij} = transition prob. for: $x_i \rightarrow x_j$

- Note that the sum of the transition probabilities from one state to all the others must sum to one: $\sum_j T_{ij} = 1$
- We can try to follow the probabilistic evolution of a discrete Markov chain by considering the probability that the random variable x would be in the state x_i at step k via the following vector $\pi^{(k)}$:

$$\pi^{(k)} = \left[\pi_1^{(k)}, \pi_2^{(k)}, \cdots, \pi_n^{(k)}\right]$$

27

- The probability that the Markov chain would evolve in the state x_i at the step k+1 can be obtained from

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \pi_j^{(k)} T_{ji}$$

which can be written in matrix form: $\pi^{(k+1)} = \pi^{(k)}\mathbf{T}$

- From the previous equation one can simply infer that:

$$\pi^{(k)} = \pi^{(k-1)}\mathbf{T} = \pi^{(k-2)}\mathbf{T}^2$$

and thus that: $\pi^{(k)} = \pi^{(0)}\mathbf{T}^k$

- By increasing the number of steps, the Markov chain will follow a probability distribution obtained from

$$p_i = \lim_{k\to\infty} \pi_i^{(k)}$$

- Definition: a probability distribution is said to be invariant iff: $\pi = \pi\mathbf{T}$
- Definition: a transition matrix is said to be irreducible if it is possible to get to any state from any state
- Definition: a transition matrix **T** is said to be regular if there exists a number m>0 such that $(\mathbf{T}^m)_{ij}>0$ for all states i,j in the system (regular→irreducible)

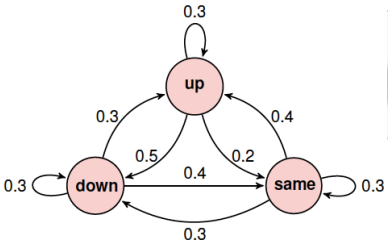
28

- **Markov theorem**: If a transition matrix **T** is regular, it admits a unique invariant distribution p* and for all initial probability distributions $\pi^{(0)}$:

$$p_i^* = \lim_{k\to\infty} \left(\pi^{(0)}T^k\right)_i$$

- For example suppose that the price of a particular asset today is increasing (up), i.e. $\pi^{(0)}$: [1, 0, 0] and let's compute how the probabilities [up, same, down] evolve in the next days given a fixed transition matrix:

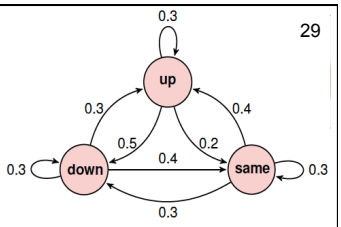
$$T = \begin{bmatrix} 3/10 & 2/10 & 5/10 \\ 4/10 & 3/10 & 3/10 \\ 3/10 & 4/10 & 3/10 \end{bmatrix}$$



- We have:

$$\pi^{(1)} = \pi^{(0)}T = \begin{bmatrix}1,0,0\end{bmatrix}\begin{bmatrix} 3/10 & 2/10 & 5/10 \\ 4/10 & 3/10 & 3/10 \\ 3/10 & 4/10 & 3/10 \end{bmatrix} = \begin{bmatrix}3/10, 2/10, 5/10\end{bmatrix}$$

$$\pi^{(1)} = \pi^{(0)}T = \left[1,0,0\right]\begin{bmatrix} 3/10 & 2/10 & 5/10 \\ 4/10 & 3/10 & 3/10 \\ 3/10 & 4/10 & 3/10 \end{bmatrix} = \left[3/10, 2/10, 5/10\right]$$



- By going on with the probabilistic evolution of our model we obtain

$$\pi^{(2)} = \pi^{(0)}T^2 = \left[0.32,\ 0.32,\ 0.36\right] \qquad \pi^{(3)} = \pi^{(0)}T^3 = \left[0.332,\ 0.304,\ 0.364\right]$$

And thus, by solving for $\pi = \pi\mathbf{T}$ (with the normalization condition $\pi_1+\pi_2+\pi_3=1$)

$$\pi = \lim_{k\to\infty} \pi^{(0)}T^k = \left[\frac{37}{112}, \frac{34}{112}, \frac{41}{112}\right] \approx \left[0.33035714,\ 0.30357143,\ 0.36607143\right]$$

- Question: it is possible to exploit a Markov chain to be able to sample a target probability distribution p* ?
- To do this, given a probability distribution p* you whish to sample, you should be able to build one regular transition matrix **T** that has exactly p* as its invariant probability distribution in order to exploit the Markov theorem

$$P_i^* = \lim_{k\to\infty}\left(\pi^{(0)}T^k\right)_i$$

- It seems not an easy task ...

The Metropolis algorithm

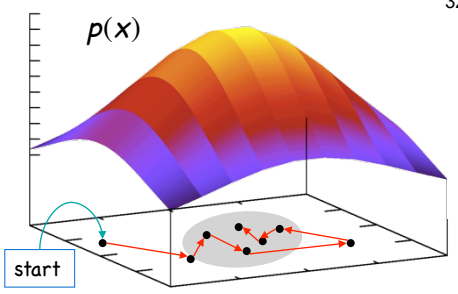
- The last sampling method we shall discuss is an advanced sampling technique first described in a paper by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller [Metropolis algorithm, Markov chain Monte Carlo, Metropolis-Hasting algorithm, M(RT)^2 algorithm!].
- The method is related to rejection techniques since it involves explicitly proposing a tentative value which may be rejected and because the normalization of the sampled function is irrelevant, we need never know it.
- The method is also related to Markov stochastic processes since the sampling of a particular probability density is obtained building a related Markov process and simulating it.
- The M(RT)^2 algorithm is of very great simplicity and power; it can be used to sample essentially any density function regardless of analytic complexity in any number of dimensions $\Rightarrow$ Monte Carlo $\approx$ M(RT)^2
- Complementary disadvantages are that sampling is correct only asymptotically and that successive variables produced are correlated, often very strongly. This means that the evaluation of integrals normally produces positive correlations in the values of the integrand with consequent increase in variance for a fixed number of steps as compared with independent samples.

31

- The method was motivated by an analogy with the behaviour of systems in statistical mechanics that approach an equilibrium whose statistical properties are independent of the kinetics of the system.
- By kinetics we mean a stochastic transfer matrix K(x|y) that governs the evolution of the system; it is nothing more than a conditional probability distribution, and it is also called a stochastic kernel or transition probability.
- In treating a physical system, one usually assumes that K(x|y) is known, and one has the task of finding p(x) . The M(RT)^2 algorithm reverses this: one has the task of finding a convenient and correct kinetics (stochastic dynamics) that will equilibrate the system so that the given p(x) turns out to be the chance of observing the system near x.
- This turns out to be extremely easy given the elegant device suggested by M(RT)^2 . Transitions are proposed from, say, y to x using essentially "any" distribution T(x|y). Then on comparing p(x) with p(y) and taking into account T as well, the system is either moved to x (move "accepted") or returned to y (move "rejected").

32

- Given a probability distribution function p(x), that we wish to sample and where x is a many-dimensional vector, the M(RT)^2 technique establishes a random walk whose steps are designed so that when repeated again and again, the asymptotic distribution of x's is p(x).



- Suppose that x_1, x_2, ... are the steps in a random walk. Each of the x's is a random variable and has an associated probability $d_1(x_1)$, $d_2(x_2)$, ... where d_n can be any distribution for x. We whish that the $d_n(x)$ have the property that asymptotically

$$\lim_{n\to\infty} d_n(x) = p(x)$$

- To build such random walk we consider a stationary Markov process:

$$p_1(x,t+\Delta t) = p_1(x,t) = p(x)$$

$$p_{1|1}(x_2,t_2|x_1,t_1) = p_t(x_2|x_1) \quad with \quad t = t_2 - t_1$$

33

- Thus the (marginal) probability density related to the single random variable x is time independent and equal to p(x) and the conditional probability $p_{1|1}$ is a function of the difference $t=t_2-t_1$
- Let us consider discrete times: $t=t_n$ $n \in \mathbb{N}$ and $t_{n+1}-t_n=\text{cost}.$ $\forall n$ it follows that the conditional probability (stochastic transfer matrix or transition probability) related to two nearby events in the random walk or stochastic process, x_n and x_{n+1}, depends only on them:

$$p_{1|1}(x_2,t_2|x_1,t_1) = p_t(x_2|x_1) = K(x_2|x_1)$$

- With these elements we can write the second equation which characterize a stochastic Markov process

$$p(x) = \int K(x|y)\, p(y)\, dy$$

this integral equation show that a stationary Markov process has, at least, one probability distribution which is invariant with respect to its defining transition probability $p_{1|1}=K$.

- By considering a different probability distribution, say $d_1(x)$, which is not invariant with respect to the stochastic nucleus K one will build a non-stationary Markov process characterized by the sequence of marginal distributions:

$$d_{n+1}(x) = \int K(x|y)\, d_n(y)\, dy \quad n = 1,2,\cdots$$

- By fixing the transition probability K(x|y) there could be more than one invariant probability density; for example with $K(x|y)=\delta(x-y)$

$$\int K(x|y)\, p(y)\, dy = \int \delta(x-y)\, p(y)\, dy = p(x) \quad \forall p$$

- Definition: a stochastic nucleus K(x|y) is ergodic if and only if one strictly positive probability density p(x) exists such that

$$p(x) = \lim_{n\to\infty} d_{n+1}(x) = \lim_{n\to\infty} \int K(x|y)\, d_n(y)\, dy \quad \forall d_1(x)$$

- Definition: a sequence of function d_n is equi-continuous when

$$\forall \varepsilon > 0 \quad \exists \delta \geq 0 \quad \ni \quad |x_1 - x_2| < \delta \quad \Rightarrow \quad |d_n(x_1) - d_n(x_2)| < \varepsilon \quad \forall n$$

- Definition: a stochastic nucleus K(x|y) is regular if and only if it gives rise to a sequence of equi-continuous probability densities $d_n(x)$ when $d_1(x)$ is uniformly continuous. In other words the probability distributions d_n "remains good" if the starting distribution d_1 is "good" and if the nucleus K(x|y) is regular.
- Theorem: A nucleus K(x|y) strictly positive and regular is ergodic if and only if it has an invariant strictly positive probability density p(x)

34

- Thus if we find a (1) regular transition probability K(x|y) such that the probability distribution p(x) (that we wish to sample) is (2) invariant for K, i.e.,

$$p(x) = \int K(x|y)\, p(y)\, dy$$

the previous theorem guarantees us that

$$p(x) = \lim_{n\to\infty} d_{n+1}(x) = \lim_{n\to\infty} \int K(x|y)\, d_n(y)\, dy \quad \forall d_1(x)$$

- Therefore the idea which underlies the M(RT)^2 algorithm is to find one (there could be infinite… but one in enough!) particular and regular transition probability in order to exploit the properties of stochastic nucleus of stationary Markov processes and sample a desired probability distribution
- Thus given a probability distribution p(x) that we wish to sample, what is the regular stochastic nucleus K(x|y) that has p(x) as invariant probability density?

  Given our p(x), suppose that K(x|y) is regular stochastic nucleus; to be ergodic (previous theorem) K(x|y) must be such that

$$p(x) = \int K(x|y)\, p(y)\, dy$$

35

- This is however an integral equation which is difficult to handle with… let's use a stronger (point-like) property, it is the detailed balance:

$$K(x|y)\, p(y) = K(y|x)\, p(x)$$

- The detailed balance implies the previous integral equation as it can be easily seen:

$$\int K(x|y)\, p(y)\, dy \overset{\text{detailed balance}}{=} \int K(y|x)\, p(x)\, dy = p(x) \int K(y|x)\, dy = p(x)$$

where $\int K(y|x)\, dy \to 1$ (Probability to go somewhere)

- Thus detailed balance on K(x|y) and p(x) ⇒ that p(x) is invariant for K
- In order to find one regular K(x|y) which fulfils detailed balance M(RT^2) decomposed the transition probability K into the product of a trial transition probability T(x|y) time a probability to accept a particular proposed move y→x, A(x|y):

$$K(x|y) = T(x|y) \times A(x|y)$$

36

- Given T(x|y) one possible choice [M(RT)²] is:

$$A(x|y) = \min\left[1, q(x|y)\right] \quad where \quad q(x|y) = \frac{T(y|x)p(x)}{T(x|y)p(y)} \geq 0$$

it is enough to show that detailed balance is fulfilled:

$$\begin{aligned} K(x|y)p(y) &= T(x|y)A(x|y)p(y) \\ &= T(x|y)p(y)\min\left[1, \frac{T(y|x)p(x)}{T(x|y)p(y)}\right] \\ &= \min\left[T(x|y)p(y), T(y|x)p(x)\right] \\ &= \min\left[\frac{T(x|y)p(y)}{T(y|x)p(x)}, 1\right]T(y|x)p(x) \\ &= T(y|x)A(y|x)p(x) = K(y|x)p(x) \end{aligned}$$

- It follows that p(x) is invariant for such K(x|y)=T(x|y)A(x|y) and that $d_{n+1}(x)=\int K(x|y)d_n(y)dy \rightarrow p(x)\ \forall d_1(x)$
- This result is valid ∀ "good" transition probability T(x|y) is used to propose the move y→x which has to be checked with the acceptance probability A(x|y).

37

- Obviously, the choice of T(x|y) has a great influence on the efficiency of the Metropolis algorithm.
- In the cases where T(x|y)=T(y|x) the acceptance reduces to $A(x|y) = \min\left[1, \frac{p(x)}{p(y)}\right]$
- Other choices (not Metropolis) are possible, but the M(RT)² is in general the most efficient; for example:

$$A(x|y) = \frac{p(x)}{p(x)+p(y)} \quad \left(with \quad T(x|y) = T(y|x)\right)$$

detailed balance:

$$K(x|y)p(y) = T(x|y)\frac{p(x)}{p(x)+p(y)}p(y) = T(y|x)\frac{p(y)}{p(x)+p(y)}p(x) = K(y|x)p(x)$$

another choice is

$$A(x|y) = \frac{q(x|y)}{1+q(x|y)} \quad \left(with \quad q(x|y) = \frac{T(y|x)p(x)}{T(x|y)p(y)}\right)$$

detailed balance:

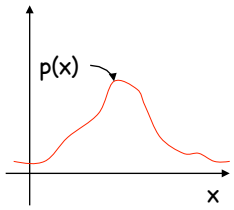
$$K(x|y)p(y) = T(x|y)\frac{q(x|y)}{1+q(x|y)}p(y) = T(x|y)\frac{\frac{T(y|x)p(x)}{T(x|y)p(y)}}{1+q(x|y)}p(y) = \frac{T(y|x)p(x)}{1+q(x|y)} =$$

$$= T(y|x)\frac{q^{-1}(x|y)}{q^{-1}(x|y)+1}p(x) = T(y|x)\frac{q(y|x)}{q(y|x)+1}p(x) = K(y|x)p(x)$$

38

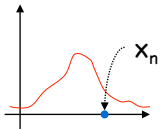
How M(RT)^2 works?

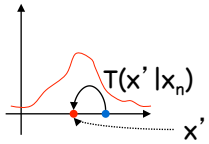
- Let p(x) be the probability we wish to sample; a random walk of values x_n will be generated that, for the properties of the transition probability K=T·A are distributed with probability $d_n(x)$ such that $d_n(x) \to p(x)$:



Step n:

- $x=x_n$
- Generate x' from $T(x'|x_n)$



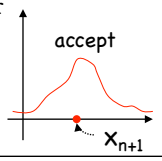


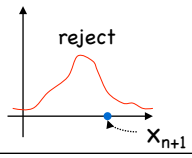
- Evaluate $A(x'|x_n)=\min\{1,[T(x_n|x')p(x')]/[T(x'|x_n)p(x_n)]\}=\alpha$
- Accept with probability α the move $x_n \to x'$

  $r \in [0,1]$ uniform random number

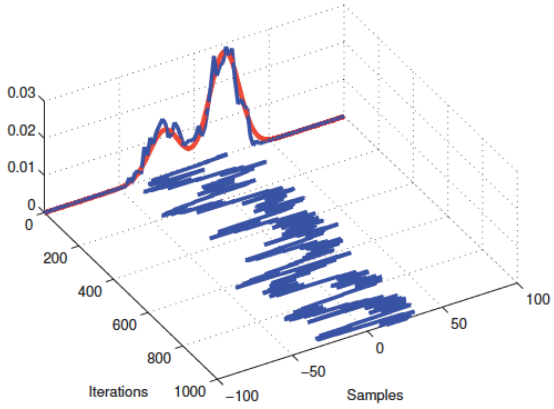
  if $r \le \alpha$ accept: $x_{n+1}=x'$

  else reject: $x_{n+1}=x_n$
- repeat





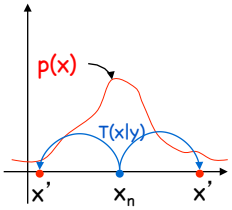
- By repeating *ad libitum* the simulation of the Markov process we obtain the sampling of the target probability distribution



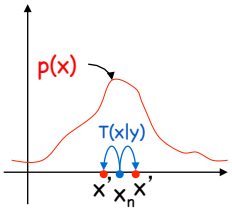
- Thus the basis on which the stochastic process is built is the transition probability T(x|y) on which the trial move is sampled. Its choice is fundamental not only for the efficiency of the Metropolis algorithm but also for its convergence. For convenience, in the following we suppose that T(x|y)=T(y|x) in order to have A(x|y)= min[1,p(x)/p(y)]

The 50% empirical "rule"

- The Metropolis algorithm samples $d_n(x)$ which converge to p(x) after what is called the equilibration time (that we will discuss in the next lecture). After the equilibration time, x_n in general will be found in regions of the sample space where p(x) is large. If $T(x'|x_n)$ is such that "large" moves are sampled, many sampled trial moves, $x_n \to x'$, will be discharged because on average $p(x')/p(x_n) \ll 1$ $\Rightarrow$ low efficency



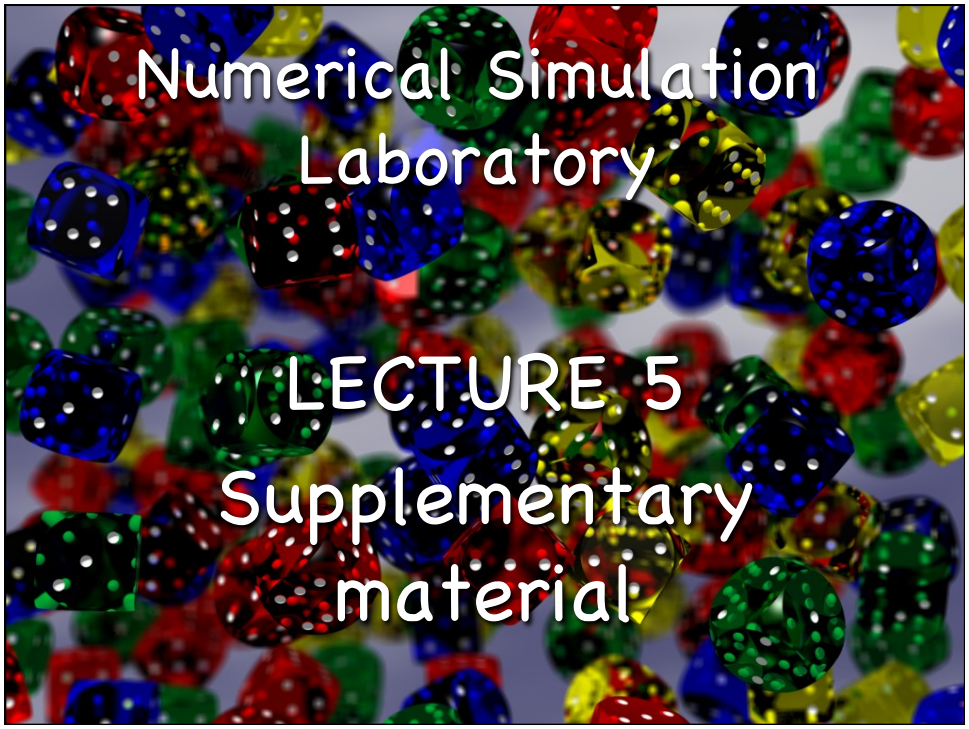
- On the contrary if $T(x'|x_n)$ is such that "small" moves are sampled, in general many sampled trial moves, $x_n \to x'$, will be accepted because $x' \approx x_n$ and therefore $A(x'|x_n)=\min[1,p(x')/p(x_n)] \approx 1$ $\Rightarrow$ high correlation



- Empirical rule: the functional form of T(x|y) and its parameters should be fixed in order to have on average A(x|y)=50% (there is a certain degree of freedom on this choice depending on p(x) and its dimensionality)

Lecture 5: Suggested books

- R. Sánchez & D. Newman, *A Primer on Complex Systems*, Springer 2018
- D.K.C. MacDonald, *Noise and fluctuations: an introduction*, Dover 2001
- Peter J. Brockwell & Richard A. Davis, *Introduction to Time Series and Forecasting*, Springer 2016
- Kalos, Whitlock *Monte Carlo Methods*, Wiley 1986
- Newman, Barkema *Monte Carlo Methods in Statistical Physics*, Oxford 2001
- E. Vitali, M. Motta, D.E. Galli, *Theory and Simulation of Random Phenomena*, Springer (2018)



The convergence of M(RT)²

- One can convince himself that the sequence of $d_n(x)$ generated by the Metropolis random walk at least does not go away from p(x) in the following way: a possible measure of the discrepancy between two distribution is $D=\int|p_A(x)-p_B(x)|dx$
- Let us call $D_n=\int|d_n(x)-p(x)|dx$; we have

$$D_{n+1}=\int|d_{n+1}(x)-p(x)|dx=\int\left|\int K(x|y)d_n(y)dy-p(x)\right|dx=$$

$$=\int\left|\int\left[K(x|y)d_n(y)-K(y|x)p(x)\right]dy\right|dx=$$

$$=\int\left|\int\left[K(x|y)d_n(y)-K(x|y)p(y)\right]dy\right|dx=$$

$$=\int\left|\int K(x|y)\left[d_n(y)-p(y)\right]dy\right|dx\le\int\int K(x|y)|d_n(y)-p(y)|dydx=$$

$$=\int|d_n(y)-p(y)|\underbrace{\int K(x|y)dx}_{1}dy=\int|d_n(y)-p(y)|dy=D_n$$

thus $D_{n+1}\le D_n$