Computer Simulations in Statistical Physics

lecture 1
2007/10/25

Notiztitel 28.10.2007

N. Blümer

I Statistical properties of time series

Motivation

Experimental measurements and simulations do not yield exact result, but a time series (i.e., an ordered list) of results (scalars, vectors, etc.)

Question/problem: How to characterize and interpret time series (here: scalars)? What can we learn about the exact result?

Example 1: A simulation/series of measurements yields the numbers 47.5, 47.7, 47.2, 46.8 – characterization?

- all values approximately 47
- average: 47.3 $\qquad \frac{1}{3}(0.2^2+0.4^2+0.1^2+0.5^2)$
- est. variance of individual measurement: $0.15 \approx 0.4^2$
- est. mean with standard deviation: 47.3 ± 0.2
- trend/transient: negative (later values smaller)
- histogram: (not enough values)
- autocorrelation: "

Example 2: Millikan's determination of the electronic charge (droplets elevated in electric field)

Millikan's results: $e = (4.774 \pm 0.009)\, 10^{-10}$ esu (1913)
$e = (4.774 \pm 0.005)\, 10^{-10}$ esu (1913)

correct value (2007): $e = 1.602 \cdot 10^{-19}\,C = 4.803 \cdot 10^{-10}$ esu

Result was very good, but error estimates were much too optimistic.

$\rightsquigarrow$ bias: first diffraction result for e also 1% too low, later data correct within error bars.

Examples: correlated/uncorrelated data:
trace, histogram, autocorrelation function...
DMFT-QMC data + extrapolation

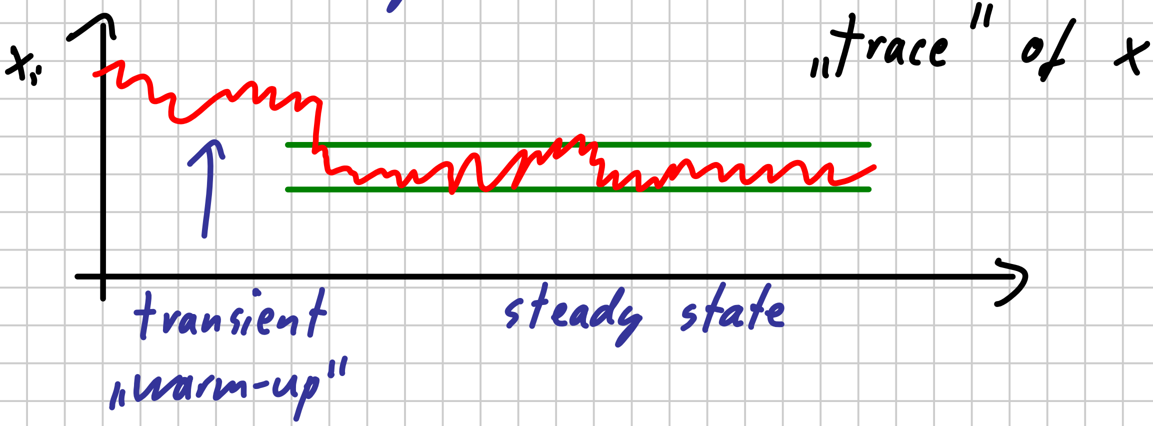
Fundamental question: how to get reliable results from simulations?

Result: estimate + error bars
- decide whether results are correct or incorrect
- essential for extrapolations (e.g. finite size)

Typical case in computer simulations:

simulations start with non-representative configurations and gradually approach steady state (e.g. thermal equilibrium)

$\rightsquigarrow$ both mean and variance of initial measurements can differ strongly from later ones and should be thrown away



Let us now assume that transient parts have been discarded. How to compute best estimate + error bars?

Mathematical excursion

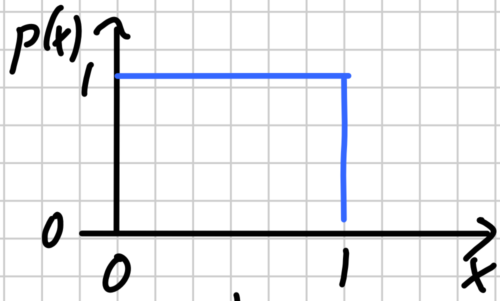
let X be a random variable with probability distribution p:

$$P(x \leq X \leq x + dx) = p(x)\,dx \qquad (p(x) \equiv p_X(x))$$

↑ probability of the statement in brackets to be fulfilled

example: uniform distribution

$$p(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases}$$



probability distribution is normalized: $\int_{-\infty}^{\infty} p(X)\,dX = 1$

mean (1st moment): $\langle X \rangle = \int_{-\infty}^{\infty} X\, p(X)\, dX$

2nd moment: $\langle X^2 \rangle = \int_{-\infty}^{\infty} X^2\, p(x)\, dx$

variance: $\sigma_X^2 = \langle (X - \langle X \rangle)^2 \rangle = \langle X^2 \rangle - \langle X \rangle^2$

$\sigma \equiv$ width of distribution

$\left(= \frac{1}{2}, \quad = \frac{1}{3}, \quad = \frac{1}{12}, \quad = \frac{1}{\sqrt{12}} \approx 0.29 \right)$ for uniform dist.

The mean is always additive; the variance only for uncorrelated random variables:

$$\langle X+Y \rangle = \int dX \int dY\, p(X,Y)\, (X+Y)$$

$$= \int dX\, X \underbrace{\int dY\, p(X,Y)}_{=p(x)} + \int dY\, Y \underbrace{\int dX\, p(X,Y)}_{=p(y)}$$

$$= \langle X \rangle + \langle Y \rangle$$

$$\sigma_{X+Y}^2 = \langle [(X+Y) - \langle X+Y \rangle]^2 \rangle = \langle [(X - \langle X \rangle) + (Y - \langle Y \rangle)]^2 \rangle$$

$$= \sigma_X^2 + \sigma_Y^2 + \underbrace{2 \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle}_{=0 \text{ for uncorrelated variables}}$$

Specifically for $\sigma_X = \sigma_Y \equiv \sigma$ in uncorrelated case:

$$\sigma\left[\frac{X+Y}{2}\right] = \frac{\sigma}{\sqrt{2}}$$

(and in fully correlated case $X = Y$: $\sigma\left[\frac{X+Y}{2}\right] = \sigma$)

and for arithmetic average of N uncorrelated random numbers with equal variance σ:

$$\sigma_{\bar{X}} = \sigma\left[\frac{1}{N} \sum_{i=1}^{N} X_i\right] = \frac{\sigma}{\sqrt{N}}$$

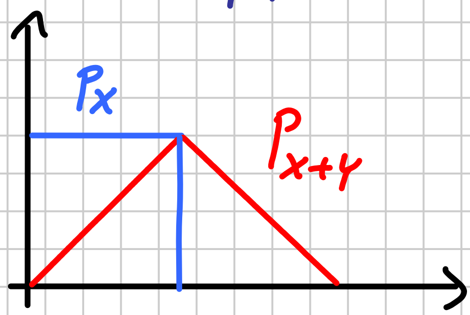
This is one formulation of the central limit theorem.

Full probability distribution of sum/average?

$$p_{X+Y}(z) = \int dX\, p_X(X) \int dY\, p_Y(Y)\; \delta(z-(X+Y))$$
$$= \int dX\; p_X(X)\; p_Y(z-X)$$

specifically for uniformly distributed X, Y:

$$p_{X+Y}(z) \begin{cases} z & \text{for} \quad 0 \le z \le 1 \\ 2-z & \text{"} \quad 1 \le z \le 2 \\ 0 & \text{else} \end{cases}$$



Look at higher order cumulants $\rightsquigarrow$ normal distribution

$$\boxed{p_{\bar{X}}(x) \xrightarrow{N\to\infty} \frac{1}{\sigma_{\bar{X}}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\bar{X}-\langle x\rangle}{\sigma_{\bar{X}}}\right)^2\right]; \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}}$$

central limit theorem (uncorrelated case!)

Now: application to simulation / measurements

again: random variables X_i with same distribution, non necessarily uncorrelated

Problem: probability distribution, mean, variance unknown. Best unbiased estimates?

true mean

easy: $\langle X \rangle \approx \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$

expectation value

$\bar{X}$ is (best) unbiased estimator for $\langle X \rangle$ since

$$\langle \bar{X} \rangle = \frac{1}{N} \sum_{i=1}^{N} \langle X_i \rangle = \langle X \rangle$$

(also unbiased: $\tilde{x} = \sum_{i=1}^{N} a_i x_i$ with $\sum_{i=1}^{N} a_i = 1$

also estimator: $\tilde{\tilde{x}} = \frac{1}{N-1} \sum_{i=1}^{N} x_i$ (since $\langle \tilde{\tilde{x}} \rangle = \frac{N}{N-1} \langle x \rangle \xrightarrow{N\to\infty} \langle x \rangle$))

difficult: best estimator for error of $\bar{x}$?

(i) unbiased estimator for σ_x?

$$\left\langle \sum_{i=1}^{N} (x_i - \bar{x})^2 \right\rangle = N \langle (x_i - \bar{x})^2 \rangle$$

$$= N \left\langle \left(x_i - \frac{1}{N} \sum_{j=1}^{N} x_j \right)^2 \right\rangle$$

$$= N \left\langle \left[(x_i - \langle x \rangle) - \frac{1}{N} \sum_{j=1}^{N} (x_j - \langle x \rangle) \right]^2 \right\rangle$$

$$= N \left\langle \left[\left(1 - \frac{1}{N}\right)(x_i - \langle x \rangle) - \frac{1}{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} (x_j - \langle x \rangle) \right]^2 \right\rangle$$

$$\overset{\text{no corr.}}{=} N \left[\left(1 - \frac{1}{N}\right)^2 \langle (x_i - \langle x \rangle)^2 \rangle + \frac{1}{N^2} \sum_{\substack{j=1 \\ j \neq i}}^{N} \langle (x_j - \langle x \rangle)^2 \rangle \right]$$

$$= \frac{(N-1)^2}{N} \sigma^2 + \frac{N-1}{N} \sigma^2$$

$$= \sigma^2 \frac{[(N-1)+1](N-1)}{N} = (N-1)\sigma^2$$

Thus, $\boxed{\sigma_{est}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$ is an unbiased estimator for the variance σ^2 <u>if</u> the data is uncorrelated.

Homework: Is this estimator also unbiased for data with finite autocorrelation?