Molecular computer experiment

Also *pseudoexperiment*

REAL EXPERIMENT	**COMPUTER EXPERIMENT**
Record everything in a lab notebook	Record everything in a lab notebook
Choose method (device, assay)	Choose method (MD, MC, . . .)
Build the experimental apparatus (from parts)	Download/buy/write a computer program (blocks of code)
Purchase chemicals, synthetise if not available	Get a force field, fit/calculate parameters if not available (e.g., partial charges)
Prepare the experiment	Prepare initial configurations, etc.
Perform the experiment, watch what's going on	Run the code, observe time development, control quantities, etc.
Analyse and calculate	Calculate mean values (with error estimates)
Clean the laboratory, store samples	Make backups, erase temporary files

MD or MC?

Often, MC and MD can be applied to similar systems.

MD

- realistic models, complex molecules (bonds, angles. . .)
- condensed matter in general (fluids, solutions; biochemistry)
- kinetic quantities (diffusivity, viscosity. . .)
- more packages available, better parallelization (domain-decomposition for large systems)

MC

- simple qualitative models (lattice, hard-sphere-like)
- dilute systems
- critical phenomena
- fluid equilibria
- overcoming barriers, exchange of molecules, etc. is easier with MC
- fewer packages available, less efficient parallelization (prefetching)

Is it correct?

Systematic errors:

- inaccurate molecular model (force field)
- neglected quantum effects, neglected many-body forces . . .
- small sample (finite-size effects)
- insufficient time scale (long correlations, bottleneck problems)
- method problems: integration errors (too long timestep), inappropriate thermostat/barostat, not equilibrated enough, inaccurate treatment of Coulomb forces. . .

Random (stochastic, statistical) errors* are essential in stochastic methods

- time-correlated
- can be decreased by long calculations

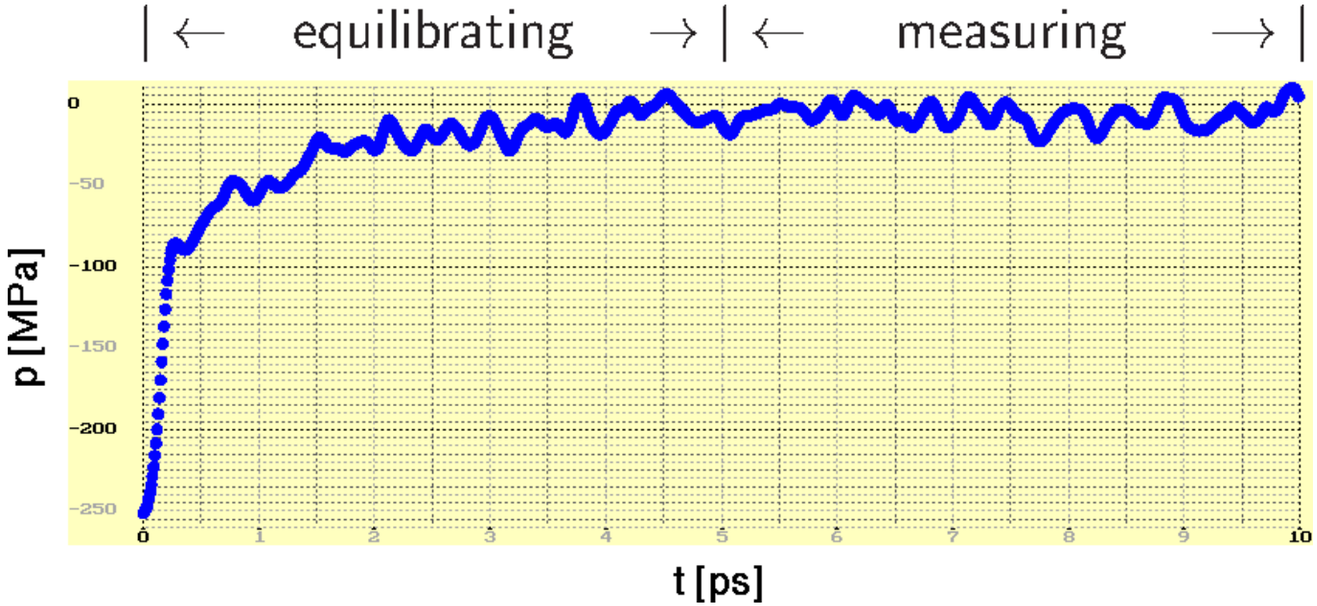
Uncertainty (in metrology) includes critical assessment of both the systematic and random errors*

---

* different terminology in different fields (mathematical statistics, metrology, physics, chemistry)

Simulation methodology

- Start (initial configuration):
  - experimental structure (biomolecules)
  - crystal → liquid (melt), gas → liquid (shrink); Packmol
  - random configuration (overlaps of molecules = problem in MD)
    problem for ill-defined models (exp-6, TIP4P etc.)
  - pre-optimize to remove overlaps etc. (conjugate gradients etc.)
  - lattice models: crystal/chaos
  - MD: velocities = Maxwell–Boltzmann (approximation enough)
- Equilibration → watch graphically (convergence/time profile)
- Measuring the quantities of interest incl. estimates of errors



Calculations

Example. We simulate an argon droplet in a periodic cubic simulation cell. Let us have $N = 1000$ atoms and temperature $T = 85$ K. The distance between surfaces of periodic images of droplets should be equal to the droplet diameter. Calculate the size of the box in Å. Argon density is $\rho = 1.4\,\mathrm{g\,cm^{-3}}$, molar mass $M(\mathrm{Ar}) = 40$ g/mol.

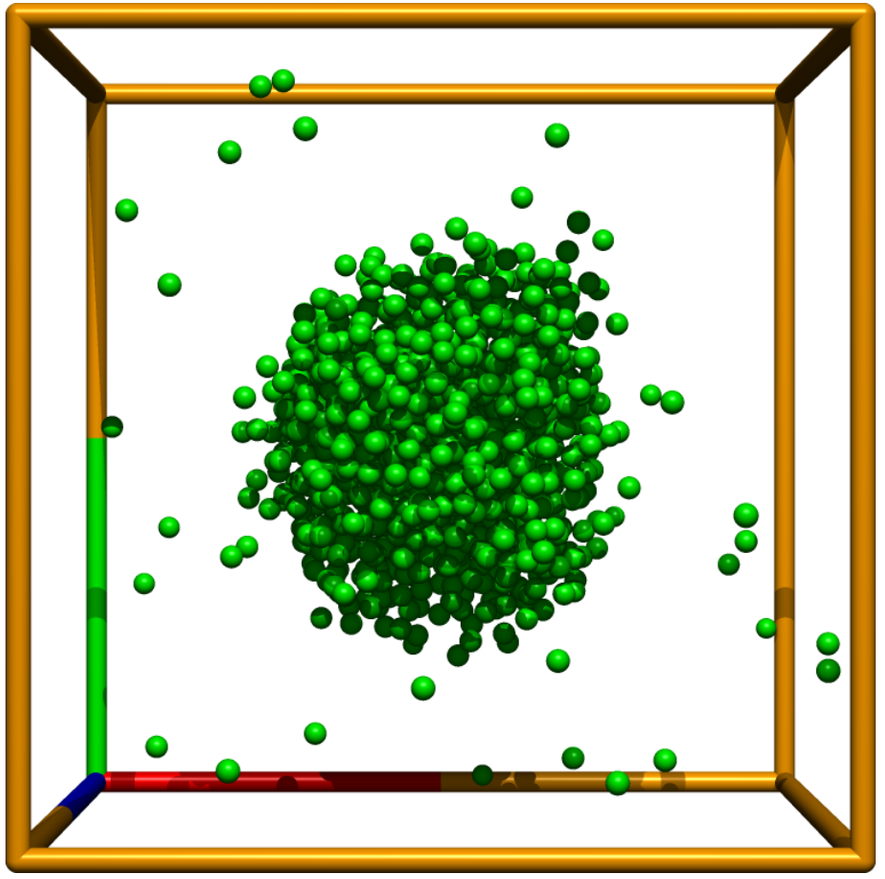
molar volume: $V_\mathrm{m} = M/\rho$

volume per 1 atom: $V_1 = V_\mathrm{m}/N_\mathrm{A}$

volume of N atoms: $V = NV_1 = NM/\rho N_\mathrm{A}$
$= 1000 \cdot 0.040\,\mathrm{kg\,mol^{-1}}/(1400\,\mathrm{kg\,m^{-3}} \cdot 6.022{\cdot}10^{23}\,\mathrm{mol^{-1}})$
$= 4.744{\cdot}10^{-27}\,\mathrm{m^3}$

droplet radius: $\frac{4}{3}\pi R^3 = V \Rightarrow R = 2.24{\cdot}10^{-9}$ m

box size: $\underline{L = 90\ \text{Å}}$

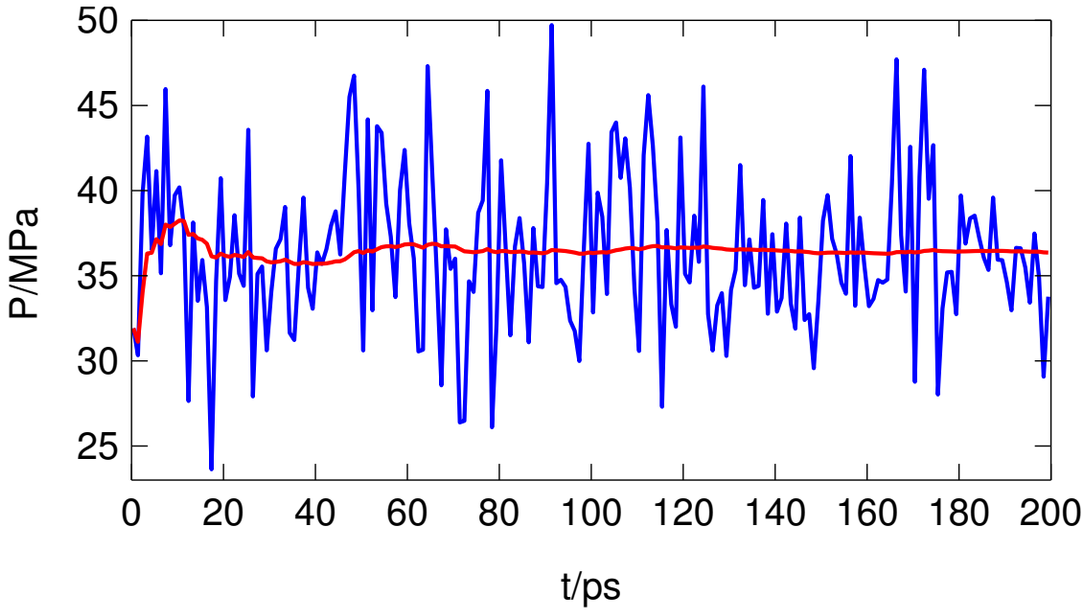


Measurements

Trajectory = sequence of configurations (MD: in time)

Convergence profile:

- time development of a quantity (time profile, —)
  problems better seen
- cumulative (running average, —)
  can estimate the inaccuracy



Type of statistical treatment:

- averaged values (← ergodic hypothesis)
- less often fluctuations

Type of quantity:

- mechanical (temperature, pressure, internal energy, order parameters. . .)
- entropic (S, F, μ,. . .)
- structure (correlation functions, number of neighbors, analysis of clusters. . .)
- auxiliary or control quantities (order parameters, integrals of motion in MD)

Random errors

quantity = (estimate of the mean value) ± (estimate of the error)

Arithmetic average (example of a **statistic**, also *statistical functional, estimator*, in metrology *measurement function*):

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$$

statistic = estimator
statistics = field of mathematics

Standard error = standard deviation of the statistic, usually denoted as σ

$$\sigma_X = \sqrt{\left\langle \left(\overline{X} - \langle X \rangle\right)^2 \right\rangle}$$

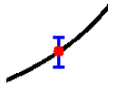
For **uncorrelated** (independent) X_i and large m, $\overline{X}$ has Gaussian normal distribution

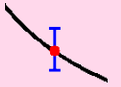
The **estimate** of the standard error of the arithmetic average of uncorrelated data:

$$\sigma_X^{\text{estim}} = \sqrt{\frac{\sum_{i=1}^{m} \Delta X_i^2}{m(m-1)}}, \qquad \text{where } \Delta X_i = X_i - \overline{X}$$

Customs and bad habits

How the uncertainty of measured quantities are expressed in different fields:

- **Physics:** $Q = 123.4 \pm 0.5 \equiv 123.4(5) \equiv 123.4_5$ 

  asymmetry: $Q = 1.23^{+0.67}_{-0.45}$ 

  $0.5 = \sigma(Q)$ = (estimated) standard error/uncertainty of statistic Q (e.g., $Q = \overline{X}$), also: standard deviation (meaning of the average or other statistic)
  loosely: (estimated) error/uncertainty, standard deviation, error margin, error bar,... In case normal distribution, it holds $\langle Q \rangle \in 123.4 \pm 0.5$ with probability 68 %

- **Biology, economy, politology, engineering, pharmacology:** $Q = 123.4 \pm 1.0$
  $\pm 1.0 = \pm 2\sigma(Q)$ = confidence interval at (confidence) level 95 %
  loosely: ± 1.0 = confidence interval, 1.0 = error/uncertainty, ...

  In case normal distribution, it holds $\langle Q \rangle \in 123.4 \pm 1.0$ with probability 95 %

- **Bonferroni**: The confidence interval chosen so that **all** calculated parameters are in the interval given.

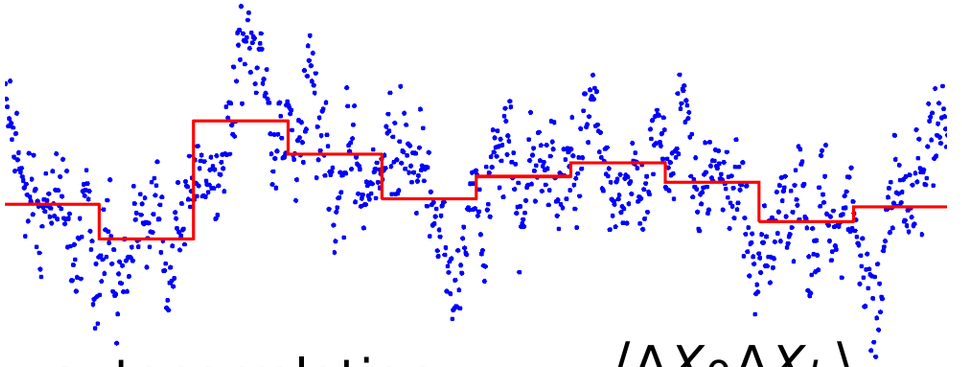
- **„Physical certainty“** starts at $\pm 5\sigma_X$ (confidence level 0.999 999 43)

The type of error/uncertainty must be always specified

α = significance level, often 5 %
$1 - \alpha$ = confidence level, often 95 %

Analysis of time series and error estimation

Problem: correlations



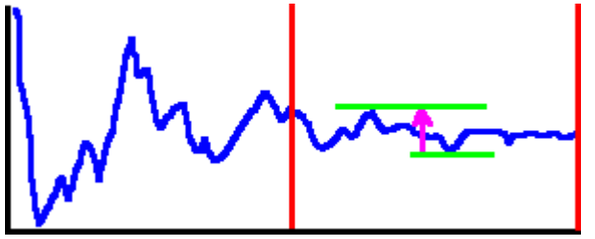
- block method: $\overline{X}_j = \frac{1}{B}\sum_{i=1}^{B} X_{i+(j-1)B}$
- analysis of correlations ⇒

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{m} \Delta X_i^2}{m(m-1)}(1+2\tau)}$$

correlation time/length: $\tau = \sum_{k=1}^{\infty} c_k$

autocorrelation coefficient: $c_k = \frac{\langle \Delta X_0 \Delta X_k \rangle}{\langle (\Delta X)^2 \rangle}$

MC: c_k is monotonously decreasing [ex.: $c_k = \sum_{\lambda \neq 1} c_\lambda \lambda^k$, $\lambda \in (-1, 1)$]
MD: $c_k \to c(t)$ (time autocorrelation function): damped oscillations



- even better = both approaches combined:
  first to block a bit, then $\tau \approx c_1$
- from running average (roughly ≈ 10 blocks):

$$\sigma_X^{\text{estim}} \approx 0.6[\max_{\text{2nd half}}(X) - \min_{\text{2nd half}}(X)]$$

or to be on the safe side (this formula is approximate):

$$\text{err}_X \approx \max_{\text{2nd half}}(X) - \min_{\text{2nd half}}(X)$$

⇒ $\langle X \rangle \in (\overline{X} - \text{err}_X, \overline{X} + \text{err}_X)$ with probability ≈ 85% (for long enough time series)

Exercise/Example

- Generate random correlated data (1st order process):

$$X_{k+1} = qX_k + u$$

  where $u = u_{[0,1)}$ or u_{Gauss} etc., and $|q| < 1$.

- Calculate the arithmetic average incl. error by different methods

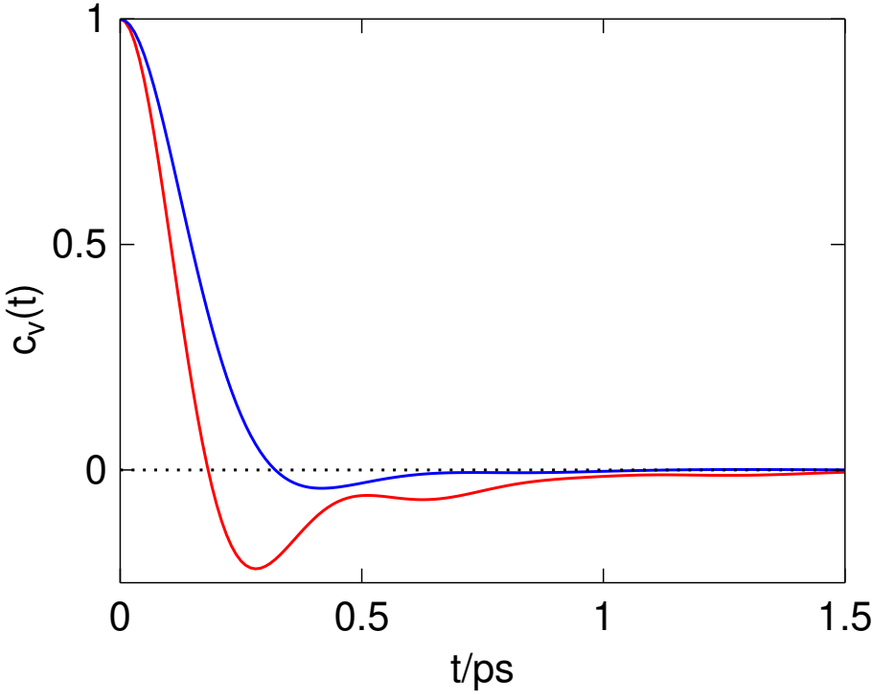
Note: it is known analytically,

$$\sigma_X = \sqrt{\frac{1+q}{1-q}}\sqrt{\frac{\operatorname{Var} X}{m}} = \frac{1}{1-q}\sqrt{\frac{\operatorname{Var} u}{m}}$$

where the variance, or fluctuation, is defined by $\operatorname{Var} X = \langle (X - \overline{X})^2 \rangle$

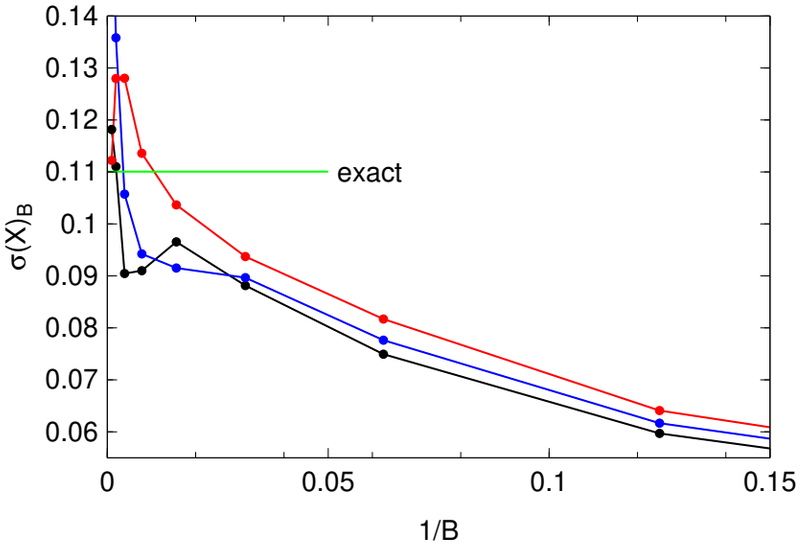
Time autocorrelation function

Velocity-velocity autocorrelation function of liquid argon:
— 150 K, 1344 $\mathrm{kg\,m^{-3}}$,
— 120 K, 1680 $\mathrm{kg\,m^{-3}}$.
Results from a 100 ps trajectory for 216 Lennard-Jones particles



Typical behavior (MC + MD):

- fluid: $\lim_{t\to\infty} c(t) = \mathrm{const}\, t^{-3/2}$ (hydrodynamic tail)
  blocking method: $\sigma(X)_B$ is linear in $1/\sqrt{B}$ for $B \to \infty$
- jumps between states (high barrier): $c(t) \propto \lambda^t$ (λ just below 1)
  blocking method: $\sigma(X)_B$ is linear in $1/B$ pro $B \to \infty$



Error analysis – addition and subtraction

Sum of independent measurements: squares of standard deviations are additive

Example. Let us integrate numerically $I = \int_0^1 f(x)\mathrm{d}x$ by the Simpson's formula:

$$I = \int_0^1 f(x)\mathrm{d}x \approx \frac{1}{6}[f(0) + 4f(0.5) + f(1)]$$

For $f(x)$ we have measured the following data with standard errors:

x	0	0.5	1
$f(x)$	1.34(5)	1.57(3)	1.77(6)

Calculate I including the error estimate.

$$I = \frac{1}{6}[1.34 + 4 \times 1.57 + 1.77] = 1.565$$

$$\sigma(I)^2 = (0.05/6)^2 + (0.03 \times 4/6)^2 + (0.06/6)^2 = 0.000569 \quad \Rightarrow \quad \sigma(I) = 0.024$$

$$\underline{I = 1.565(24)}$$

Error analysis – division and multiplication

Squares of relative errors are additive

Example. Calculate $3.46(7)/0.934(13)$.

fraction: $3.46/0.934 = 3.704$

rel. error $= \sqrt{\left(\frac{0.07}{3.46}\right)^2 + \left(\frac{0.013}{0.934}\right)^2} = 0.0246$

abs. error $= 3.704 \times 0.0246 = 0.091$

$3.46(7)/0.934(13) = 3.70(9)$ (or rounded up: $3.70(10)$)

Error analysis

Error of function f of a variable with error is (linearized; i.e., for small σ):

$$f(x \pm \sigma_x) = f(x) \pm f'(x)\sigma_x$$

$$\ln(x \pm \sigma_x) = \ln x \pm \frac{\sigma_x}{x}, \qquad \exp(x \pm \sigma_x) = \exp x \pm \sigma_x \exp x, \qquad \frac{1}{x \pm \sigma_x} = \frac{1}{x} \pm \frac{\sigma_x}{|x|^2}$$

Example. Calculate the activity of H^+ from pH = 2.125(5).

activity:

$$a_{H^+} = 10^{-2.125} = \exp(-2.125 \times \ln 10) = 0.00750$$

error Method 1:

$$\sigma = 0.005 \times \ln 10 \times a = 0.000086$$

error Method 2:

$$\sigma = |10^{-2.125} - 10^{-2.125-0.005}| = 0.000087$$

activity with error (uncertainty) estimate:

$$\underline{a_{H^+} = 0.00750(9)}$$