nested_fit: developments and tests

Lune Maillard
Supervisor: Martino TRASSINELLI
Co-supervisor: Fabio FINOCCHI

31st May 2022

Table of contents

Motivation

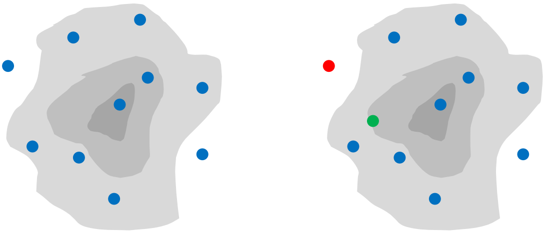
Long term: Make a quantum description of hydrogen (**electrons** + **nucleus**) in solid state matter

Quantum analysis of nucleus $\rightarrow$ increase of number of degrees of freedom

Necessary to reduce number of sampling points $\rightarrow$ Nested Sampling

Applications: analysis of experimental data, Lennard - Jones clusters

Nested Sampling: principle of exploration



Remarks

- Sampled volume at each iteration $\approx \left(\frac{K}{K+1}\right)^m$
- K: number of sampling points, m: iteration
- $E_{new} < E_m$

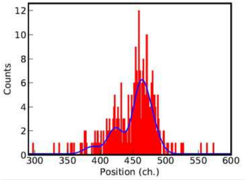
Nested Sampling: tool for fitting data

θ: parameters; $\mathcal{M}$: model; I: background information

$$\mathcal{E}(\mathcal{M}) = \mathbb{P}(\text{Data}|\mathcal{M}, I) = \int \cdots \int_{\Theta} L(\theta)\mathbb{P}(\theta|I)d\theta$$

Variable change: $X(\mathcal{L}) = \int \cdots \int_{L(\theta)>\mathcal{L}} \mathbb{P}(\theta|I)d\theta$

$\Rightarrow \mathcal{E}(\mathcal{M}) = \int_0^1 \mathcal{L}(X)dX$



Trassinelli, Ciccodicola, *Entropy*, 2020

Nested Sampling: tool for calculating the partition function

$\boldsymbol{x}$: positions; $\boldsymbol{p}$: momenta; $\beta = \frac{1}{k_b T}$

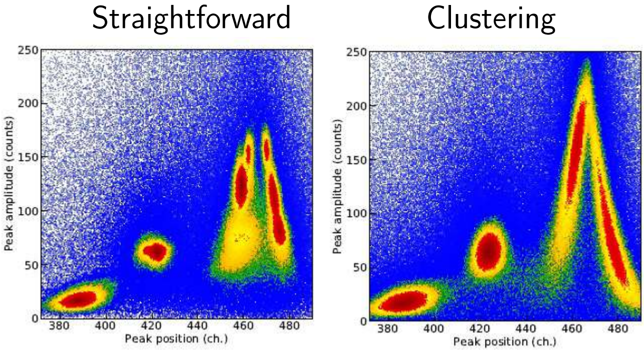
$$Z(\beta) = \int \exp(-\beta E(\boldsymbol{x}, \boldsymbol{p})) d\boldsymbol{x} d\boldsymbol{p}$$

Rewritten in function of E: $Z(\beta) = \int \rho'(E) \exp(-\beta E) dE$
$\rho'(E) \longrightarrow$ density of states between E and $E + dE$

$Z \approx \sum_m w_m \exp(-\beta E_m)$ ($w_m = \frac{1}{2}(\rho(E_{m-1}) - \rho(E_{m+1})))$

Internal energy: $U = -\frac{\partial \log(Z)}{\partial \beta}$
Heat capacity: $C_V = \frac{\partial U}{\partial T}$

Unsupervised learning: clustering data analysis



Trassinelli, Ciccodicola, *Entropy*, 2020

Computation time without clustering around eight times longer than with clustering

Without clustering: only one peak found for each run, multiple runs needed to find all peaks

Clustering methods

Mean Shift

- Two parameters: distance D + bandwidth l
- Iterative calculation of the mean of points within a given region

Density based spatial clustering for applications with noise (DBSCAN)

- Two parameters: radius ϵ + minimal number of neighbours m
- Three types of points: core, reachable, outliers

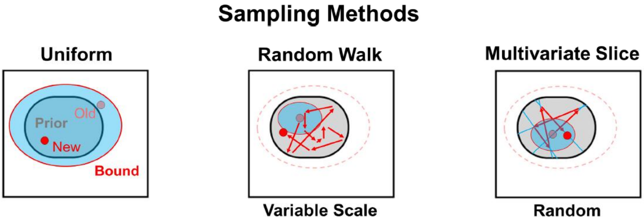
Agglomerative clustering with single linkage (Agglomerative)

- Parameter: threshold value α
- $\min_{a \in A, b \in B} d(a, b)$ (A, B: clusters; d: euclidean distance)

K nearest neighbours (KNN)

- No parameters
- Iterative on the number of neighbours k

Search methods

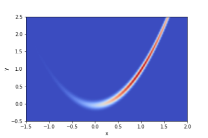


Speagle, *Montly Notices of the Royal Astronomical Society*, 2020

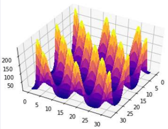
Four search methods in nested_fit: Random Walk, Uniform, Slice Sampling (size of the slice fixed) & Slice Sampling Adapt (size of the slice adaptable)

Examples studied

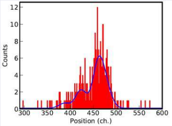
Rosenbrock 2D



Eggbox

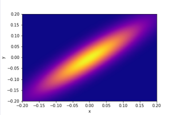


Four gaussian peaks likelihood

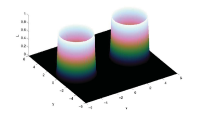


Trassinelli, Ciccodicola, *Entropy*, 2020

Gaussian with correlation



Gaussian shells

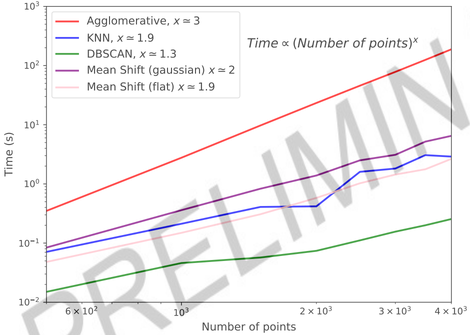


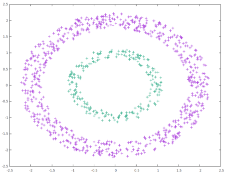
Graff, Feroz, Hobson, Lasenby, *Monthly Notices of the Royal Astronomical Society*, 2012

Other examples:

- Rosenbrock 4D
- Gaussian without correlation (5D)

Clustering methods: CPU time scaling

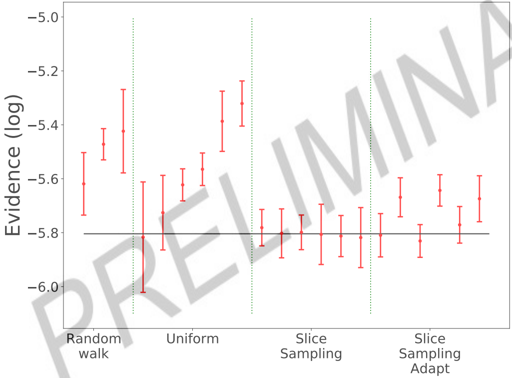


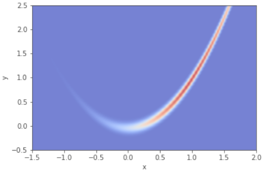


Remarks

- For a large number of points → DBSCAN preferable
- FORTRAN coded algorithms
- Tests carried out on a single processor

Search methods without clustering: Rosenbrock 2D



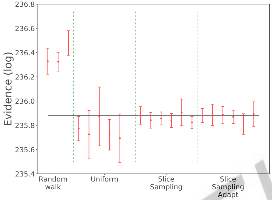


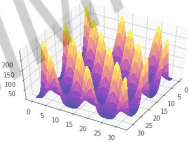
Remark

- Black line = expected value

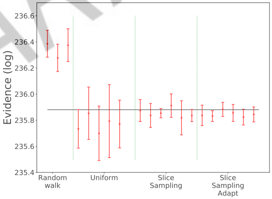
Search methods with clustering: Eggbox

Agglomerative

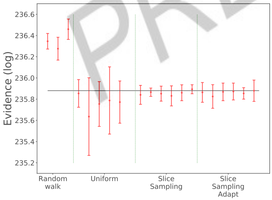




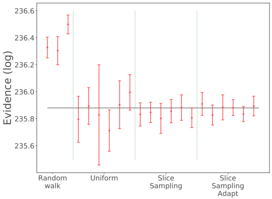
KNN



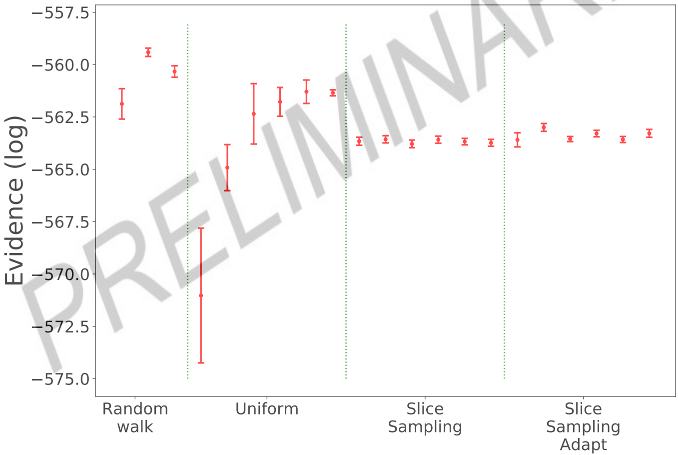
DBSCAN



Mean Shift

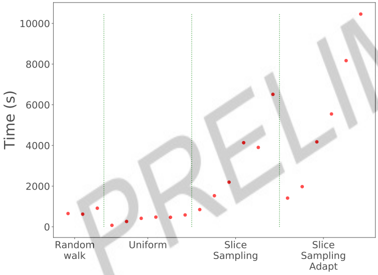


Four gaussian peaks

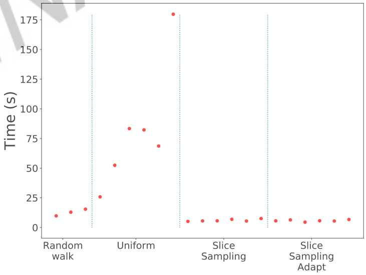


Search methods: time

Four gaussian peaks (10D)

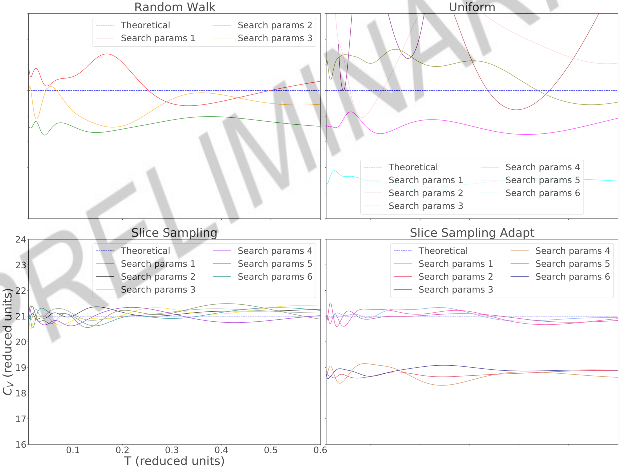


Rosenbrock 2D



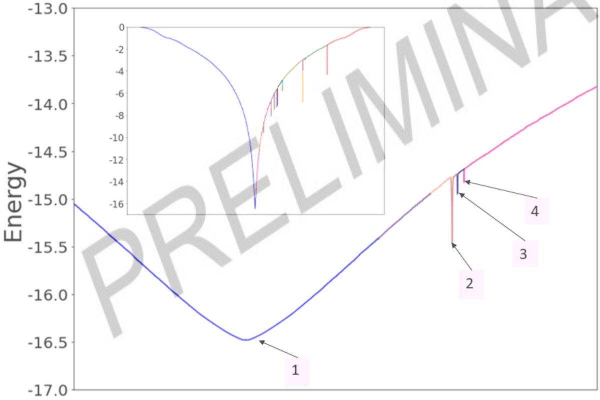
Harmonic potential

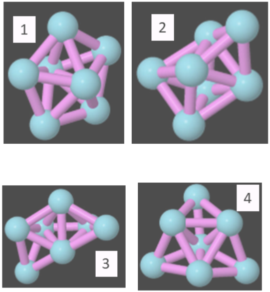
$$V_{harm}(x,y,z) = \frac{1}{2}m\omega^2(x^2+y^2+z^2),\ U = \frac{3}{2}Nk_bT,\ C_V = \frac{3}{2}Nk_b$$



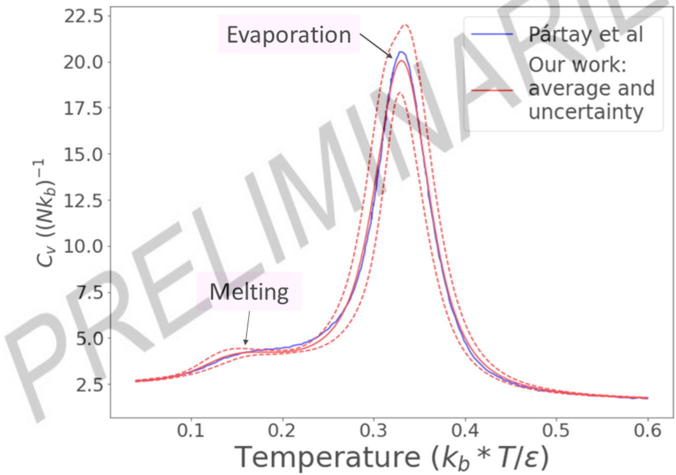
Lennard - Jones clusters I

Potential : $V_c(X) = \sum_{i<j} V_{LJ}(r_{ij})$ with $V_{LJ}(r_{ij}) = 4\epsilon \left[\left(\frac{r_0}{r_{ij}}\right)^{12} - \left(\frac{r_0}{r_{ij}}\right)^{6} \right]$, 7 atoms





Lennard - Jones clusters II



Pártay, Bartók, Csányi, *J. Phys. Chem. B*, 2010

Conclusion

Clustering

- Agglomerative: slowest method; DBSCAN: fastest method

Search methods

- Uniform search: least stable results
- Slice Sampling: best performances

Perspectives

- Extend to quantum case

- Parallelisation and better uncertainty estimation

- Neural networks with nested_fit points as input

- Reduce the numbers of degrees of freedom
  $\rightarrow$ "greedy" algorithm
  $\rightarrow$ effective Hamiltonians and Bayesian evidence calculation for their validation

Bibliography I

Cours III : Processus aléatoires et méthode de Monte Carlo, Master de physique, Parcours PAD, Sorbonne Université.

TP n°3 : Etude thermodynamique d'un cluster atomique à 2D par Monte Carlo, Master de physique, Physique numérique et projet, Sorbonne Université.

M. Ceriotti, M. Parrinello, T. E. Markland, and D. E. Manolopoulos.
Efficient stochastic thermostatting of path integral molecular dynamics.
The Journal of Chemical Physics, 133(12):124104, Sep 2010.

F. Feroz, M. P. Hobson, and M. Bridges.
MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics.
Monthly Notices of the Royal Astronomical Society, 398(4):1601–1614, 09 2009.

P. Graff, F. Feroz, M. P. Hobson, and A. Lasenby.
BAMBI: blind accelerated multimodal Bayesian inference.
Monthly Notices of the Royal Astronomical Society, 421(1):169–180, 03 2012.

W. J. Handley, M. P. Hobson, and A. N. Lasenby.
polychord: next-generation nested sampling.
Monthly Notices of the Royal Astronomical Society, 453(4):4385–4399, Sep 2015.

Maxime Sangnier.
Introduction to machine learning.

Bibliography II

I. Murray, D. MacKay, Z. Ghahramani, and J. Skilling.
Nested sampling for potts models.
In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2006.

R. M. Neal.
Slice sampling.
The Annals of Statistics, 31(3):705 – 767, 2003.

L. B. Pártay, G. Csányi, and N. Bernstein.
Nested sampling for materials.
Eur. Phys. J. B, 94(8):159, 2021.

F. Pietrucci and W. Andreoni.
Graph theory meets ab initio molecular dynamics: Atomic structures and transformations at the nanoscale.
Phys. Rev. Lett., 107:085504, Aug 2011.

L. B. Pártay, A. P. Bartók, and G. Csányi.
Efficient sampling of atomic configurational spaces.
The Journal of Physical Chemistry B, 114(32):10502–10512, 2010.
PMID: 20701382.

S. Schaack.
Nuclear quantum effects in hydrated nanocrystals.
Theses, Sorbonne Université, Sept. 2019.

Bibliography III

M. Trassinelli.
https://github.com/martinit18/nested_fit.

M. Trassinelli.
Bayesian data analysis tools for atomic physics.
Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms, 408:301–312, Oct 2017.

M. Trassinelli.
The nested_fit data analysis program.
Proceedings, 33(1), 2019.

M. Trassinelli and P. Ciccodicola.
Mean shift cluster recognition method implementation in the nested sampling algorithm.
Entropy, 22(2):185, Feb 2020.

M. Tuckerman.
Statistical Mechanics: Theory and Molecular Simulation.
Oxford Graduate Texts. OUP Oxford, 2010.