Wang-Landau Algorithm: an adapted random walk to boost convergence

Augustin Chevallier, Frédéric Cazals

Wang-Landau Algorithm: an adapted random walk to boost convergence

Augustin Chevallier and Frédéric Cazals

RESEARCH REPORT
N° 9223
November 2018
Project-Team Algorithms-Biology-Structure

ISSN 0249-6399   ISRN INRIA/RR--9223--FR+ENG

Wang-Landau Algorithm: an adapted random walk to boost convergence

Augustin Chevallier and Frédéric Cazals

Project-Team Algorithms-Biology-Structure

Research Report n° 9223 — version 2 — initial version November 2018 —
revised version November 2019 — 43 pages

Abstract: The Wang-Landau (WL) algorithm is a recently developed stochastic algorithm computing densities of states of a physical system. Since its inception, it has been used on a variety of (bio-)physical systems, and in selected cases, its convergence has been proved. The convergence speed of the algorithm is tightly tied to the connectivity properties of the underlying random walk.

As such, we propose an efficient random walk that uses geometrical information to circumvent the following inherent difficulties: avoiding overstepping strata, toning down concentration phenomena in high-dimensional spaces, and accommodating multidimensional distribution.

Experiments on various models stress the importance of these improvements to make WL effective in challenging cases. Altogether, these improvements make it possible to compute density of states for regions of the phase space of small biomolecules.

Key-words: MCMC, Wang-Landau, statistical physics, random walk, high dimension, sampling, importance sampling

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Wang-Landau Algorithm: an adapted random walk to boost convergence

Résumé : L'algorithme de Wang-Landau est un algorithme stochastique récemment développé calculant la densité d'états pour des systèmes physiques. Depuis sa création, il a été utilisé sur des systèmes (bio-)physiques. Dans certain cas, sa convergence a été prouvée. La vitesse de convergence de l'algorithme est intimement liée aux propriétés de connectivité de la marche aléatoire sous-jacente.

Nous proposons ici une marche aléatoire efficace utilisant des informations géométriques pour prévenir les difficultés suivantes: passer par dessus des strates, atténuer les phénomènes de concentration de la mesure en grande dimension, et gérer les distributions multimodales.

Les expériences numériques sur différents modèles démontrent l'importance de ces améliorations pour rendre WL efficace dans des cas complexes. In fine, ces améliorations rendent possible le calcul de densité d'état pour des régions de l'espace des phases de petite bio-molécules.

Mots-clés : MCMC, Wang-Landau, physique statistique, marche aléatoire, grande dimension, échantillonnage, importance sampling

Contents

1 Introduction

1.1 The Wang-Landau algorithm

The Wang-Landau algorithm for density of states calculations. The derivation of observable properties of (bio-)molecular systems at thermodynamic equilibrium relies on statistical physics, with the formalism of stochastic ensembles playing a pivotal role[1, 2, 3, 4]. Amidst the various algorithms available, the Wang-Landau (WL) algorithm [5, 6] is now well known and widely used despite its recent inception, in particular due to its simplicity and genericity. The WL algorithm estimates the density of states (DOS) of a system, namely the number of states (volume in phase space) of the system available at each energy level (whence the term density), which is especially useful to compute partition functions in statistical physics, and more generally observables–e.g. the average energy or the heat capacity. Estimating the DOS is especially challenging in presence of broken ergodicity; in that case, the presence of multiple energy wells prevents the system to efficiently sample the potential energy landscape (PEL), as it remains confined in selected wells [7, 8].

To review previous work, it is important to recall that the WL algorithm falls in the realm of adaptive Monte Carlo Markov Chain (MCMC) sampling algorithms. In a nutshell, WL returns an estimation of the DOS in terms of histogram. The bins of the histogram correspond to a partitioning of the energy range of the system. The algorithm resorts to importance sampling, using a biasing function derived from the current estimation of the DOS. Since the limit distribution sought is defined by the density of states, the random walk is built from the Metropolis-Hastings algorithm (M-H) [9], using the current DOS estimate in the rejection rate. (We note in passing that since the DOS values used to define transition probability depend on the history, WL is not a Markov process.) Additionally, a so-called flat histogram rule may be used to count the visits in each energy stratum and update the learning rate when all strata have been evenly visited. These main ingredients recalled, one may observe that numerous improvements were made to the original algorithm [5], both in terms of design and analysis of performances. The first key improvement has been the 1/t algorithm which solved the so-called error saturation problem [10, 11], in which a constant error on DOS estimates was incurred, due to a too fast reduction of the learning rate. Another key initiative has been to tune the random walk / proposal and the energy discretization [12], as large bins may hinder convergence by keeping the system trapped. To avoid this pitfall, a dynamic maintenance of bins has been proposed, in order to maintain a proper balance of samples across a stratum. Concomitantly, a proposal defined from a mixture of Gaussians has been introduced, in order to attempt moves of the proper size. In a different vein, it has been proposed to speed up convergence resorting to parallelism via multiple walkers [6]. However, this approach should be taken with care, as problems arise when a large number of walkers are used [13].

On the mathematical side, for the WL algorithm variant using the flat histogram, the importance of the analytical form of the DOS update rule was established [14]. For WL with a deterministic adaptation of the learning rate, to which the 1/t variant belongs, the correctness of the DOS estimates was proved, regardless of the particular analytical expression of the update rule [15].

Applications. Application-wise, WL has been used on a variety of physical systems, and more recently to biomolecules. Thermodynamics properties of RNA secondary structures were estimated using the WL algorithm [16]. Properties of clusters and peptides (up to 8 a.a.) were studied in [17]. Likewise, the thermodynamics properties of misfolded (containing a helix structure rather than a β-sheet) proteins, such as those involved in mad cow and Creutzfeldt-Jakob diseases, were studied by feeding a coarse grain protein to the 1/t WL algorithm variant [18]. In a similar spirit, a modified flat rule histogram was used in [19] to study properties of polymers on a lattice, in the HP model. However, processing continuous models of protein of significant size has remained out of reach so far [20].

1.2 Contributions

The proposal used to generate candidate conformations and the energy discretization influence one another: the average step size of the proposal should be dependent on the size energy bins. For large energy strata, the step size should be large, and small for narrow energy bins. Thus the proposal and bin sizes should not be independent, and the step size of the random walk should depend on local information. Such intricacies have precluded the development of effective WL algorithms to to handle systems as complex as bio-molecules. We make a step into this direction, as this paper focuses on the design of proposals improving the convergence speed of the algorithm, especially in high dimensional settings.

More specifically, we make three contributions. First, we design a proposal to avoid overstepping strata (section 3.3), using information on the level set surfaces bounding the strata. Second, we design a proposal to avoid congestion–i.e. remaining stuck in a stratum, a difficulty faced to move downward towards a local minimum or upward towards a local maximum. This proposal uses a heuristic based on cones to fight measure concentration problems inherent to high dimensional spaces (section 3.4). Finally, we introduce a darting move for multimodal distributions (section 3.5).

2 The Wang-Landau algorithm

2.1 Problem statement

We first introduce some classical notations [15]. Consider a bounded subset $\mathcal{E} \subset \mathbb{R}^n$ endowed with the Lebesgue measure λ as a reference. Consider also a probability distribution π with density $\pi(x)$ with respect to the Lebesgue measure. Denoting $U : \mathcal{E} \to \mathbb{R}^D$ a real valued function, consider a discretization $U_0 < U_1 < \cdots < U_d$ of its range, with possibly $U_0 = -\infty$ and $U_d = +\infty$. Also consider the partition of $\mathcal{E}$ into so-called strata $\{\mathcal{E}_1, \ldots, \mathcal{E}_d\}$, defined as the pre-images of the potential energy, namely the strata $\mathcal{E}_i$ are

$$\mathcal{E}_i = \{x \in \mathcal{E} | U(x) \in [U_{i-1}, U_i)\}.$$

Our problem is to estimate

$$\theta_i^* \overset{Def}{=} \int_{\mathcal{E}_i} \pi(x)\lambda(dx), \forall i = 1, \ldots, d. \tag{1}$$

Note that Eq. (1) is actually $\pi(\mathcal{E}_i)$, the probability of $\mathcal{E}_i$ with respect to π. This problem arises in many areas of science and engineering, two of which are discussed below.

Statistical physics: partition function. In this setting, the function U is the potential energy of a physical system. The distribution π stands for Boltzmann's distribution, that is, with $\beta = 1/(k_B T)$ the inverse temperature (and k_B Boltzmann's constant):

$$\pi(x) = Z_\beta^{-1} exp(-\beta U(x)), \text{ with } Z_\beta = \int_{\mathcal{E}} exp(-\beta U(x))dx. \tag{2}$$

Note that Z_β is the so-called partition function of the system. In this context, Eq. (1) reads as $\theta_i^* = Z_\beta^{-1} \int_{\mathcal{E}_i} exp(-\beta U(x))dx$, and one has $\sum_i \theta_i^* = 1$. The WL algorithm computes estimates θ_i for θ_i^*, which also satisfy $\sum_i \theta_i = 1$. The individual quantities θ_i are of interest since their values provide the relative weights of the strata. However, they do not give access to the partition function Z_β itself, whose calculation requires a re-normalization.

It should also be noticed that incorporating Boltzmann's factor π into Eq. (1) results in quantities θ_i which depend on the particular temperature used.

Statistical physics: density of states. To avoid the aforementioned temperature dependence, we use

$$\pi(x) = 1/\lambda(\mathcal{E}), \tag{3}$$

so that Eq. (1) yields

$$\theta_i^* = \lambda(\mathcal{E}_i)/\lambda(\mathcal{E}). \tag{4}$$

The relative volume of the i-th stratum can then be used to estimate the partition function at any temperature, using a calculation akin to numerical integration. Practically, Eq. (4) is the target of our experiments.

Numerical integration. A closely related problem is the calculation of a D-dimensional integral

$$I_D = \int_{\mathcal{E}} f(x)dx. \tag{5}$$

Assume that the range $Y_f = [y_{\min}, y_{\max}]$ spanned by f in the domain I is known, and that this interval has been split into n interval $[y_i, y_i + dy]$. Consider the estimates from Eq. (1) with $\pi(x) = 1/\lambda(\mathcal{E})$. Assume that WL has been run, and denote $\overline{y_i}$ the average value of function f computed over all points such that $f(x) \in [y_i, y_i + dy]$. The integral can be estimated as [21]

$$I \approx \sum_{i=1,\ldots,n} \theta_i^{Norm}\ \overline{y_i}, \text{ with } \theta_i^{Norm} = \lambda(\mathcal{E})\theta_i. \tag{6}$$

2.2 Algorithm

Sampling from a probability distribution μ. Assume we wish to sample a target distribution μ. We assume the existence of a *proposal* q on $\mathcal{E}$ with probability density $q(x, \cdot)$ – a transition starting from x. Using the Metropolis-Hastings transition kernel for general state spaces, and denoting δ_x a point mass at x, we introduce a random walk P_μ whose limiting distribution is μ – see [22, Section 3.4] and [23]:

$$\begin{cases} P_\mu(x, dy) & = q(x,y)\alpha(x,y)dy + \delta_x(dy) \int_{\mathcal{E}} (1 - \alpha(x,z))q(x,dz), \\ \alpha(x,y) & = \min(1, \frac{\mu(y)q(y,x)}{\mu(x)q(x,y)}). \end{cases} \tag{7}$$

with $\alpha(x,y)$ the acceptance probability of the new state y. Note that the correction factor ${q(y,x)}/{q(x,y)}$ allows q to be non-symmetric, a feature used extensively thereafter.

Algorithm. Consider the strata $\mathcal{E}_i$ defined above, as well as the mapping $J : \mathcal{E} \to \{1, \dots, d\}$ returning the index $J(x)$ of the stratum containing x.

The WL algorithm iteratively construct a sequence $\theta(t) = (\theta_1(t), ..., \theta_d(t))$ of estimates for the unknown vector $\theta^* = (\theta_1^*, ..., \theta_d^*)$ defined from Eq. (1). (NB: for the sake of conciseness, we drop the index t, as the time dependency is implicit.) For an estimate θ, we introduce the piecewise continuous probability density

$$\pi_\theta(x) = \left(\sum_{i=1}^{d} \frac{\theta_i^*}{\theta_i} \right)^{-1} \frac{\pi(x)}{\theta_{J(x)}} \tag{8}$$

The weight of each $\mathcal{E}_i$ under π_θ is proportional to θ_i^* / θ_i; In particular, all energy levels have the same weight $1/d$ under π_{θ^*}. The algorithm is then an importance sampling-like strategy, using π_θ as the bias. Observe that points sampled according to π_θ fall on average more in bins with underestimated density. Then the algorithm iterates the following two main steps:

- First, at each step, a point x is sampled according to π_θ. More precisely, using $\mu = \pi_\theta$ in Eq. (7) yields a Markov kernel P_{π_θ} denoted P_θ for the sake of conciseness; this kernel is used to sample its invariant distribution π_θ.
- Second, multiply $\theta_{J(x)}$ by an increment $\gamma > 1$ called the **learning rate**, and finally decrease the learning rate by a small fraction.

The way the learning rate γ is decreased calls for a short discussion. Historically, the Flat Histogram criterion was used [5]. Let $\nu_t(i)$ be the number of samples up to iteration t falling into bin $\mathcal{E}_i$. The vector $\{\nu_t(i)\}$ is said to verify the **flat histogram** (FH) criterion provided that, given a constant c:

$$\max_{i=1,\dots,d} \mid \frac{\nu_t(i)}{t} - 1/d \mid < 1 - c. \tag{9}$$

If this criterion holds, γ is decreased using $\gamma \leftarrow \sqrt{\gamma}$. Since $\log \gamma \leftarrow (1/2) \log \gamma$, this regime is called the *exponential regime* in the sequel. Unfortunately, this too fast rate yields an error known as the saturation error [10, 11]. Also, the flat histogram variant is sensitive to the particular analytical form of the update rule [14]. To circumvent this difficulty, the $\gamma_t = exp(1/t)$ rule was proposed [10]. Practically, one combines the two update strategies by starting with the flat histogram and the exponential regime, switching to the $1/t$ rule as soon as γ is smaller than $exp(1/t)$ [10].

The complete algorithm (Algo. 1) depends on the following parameters which influence its convergence speed: (i) the constant c for the flat histogram, (ii) the value of γ_0, (iii) the energy discretisation, (iv) the proposal q.

2.3 Theoretical convergence

A number of methods (auto-correlation as s function of the lag time, total variation distance, statistical tests) have been developed to assess the convergence of MCMC methods in particular and random processes in general [24, 25].

The convergence of WL is more challenging, since the algorithm uses a sequence of Markov kernels P_θ. It has been studied in [15], using suitable assumptions on (i) the equilibrium measure, (ii) the Metropolis-Hastings kernel, and (iii) the sequence of learning rates. Under these assumptions, the authors proved a central-limit like theorem giving a theoretical convergence speed of $O(1/\sqrt{n})$ with n the number of steps.

Algorithm 1 Wang Landau

```
Set $\theta = (1/d, ..., 1/d)$
Set exponential regime = True
Set $\gamma = \gamma_0$ with $\gamma_0 > 1$
while $t < t_{max}$ do
    Sample $x_{t+1} \sim P_\theta(x_t, .)$
    Set $\theta_{J(x_{t+1})} = \gamma\ \theta_{J(x_{t+1})}$
    Renormalise $\theta$
    if Exponential regime then
        if Flat histogram then
            $\gamma = \sqrt{\gamma}$
        if $\gamma < \exp(\frac{1}{t+1})$ then
            Set exponential regime = False
            Set $\gamma = \exp(\frac{1}{t+1})$
    else
        $\gamma = \exp(\frac{1}{t+1})$
```

2.4 Convergence rate: further insights

The convergence speed of the WL is tightly coupled to the mixing times of the Markov chains P_θ, which depends on the proposal q. In [12], a refinement rule for the discretisation is provided as well as a rule to find suitable parameters for a multi-modal Gaussian proposal. The paper does not establish any explicit link between the proposal and the discretisation. However, a symmetric proposal is used. Such proposal will sample the space uniformly. Hence, to obtain a high transition probability between two energy levels, the ratio of their respective volumes must be controlled: should this ratio be too small (or to high), the probability of proposing a move going from the smallest energy level to the biggest one vanishes. This obstruction vanishes for a non symmetric proposal, a strategy we will be using.

For multi-modal distributions, the difficulty to switch from one mode to another can also be a bottleneck for the mixing time. In [26], a strategy called darting is proposed. It consists in attempting long range jumps between regions associated to precomputed modes. The knowledge of the volume of the targeted regions allows one to guarantee detailed balance [27] whence a procedure sampling the desired distribution. Note that for molecular systems, where Boltzmann distribution yields one mode for each local minimum of the potential energy, local minima can be obtained by gradient descents and associated search methods such as basin hopping and variants [28, 29].

2.5 MCMC and adaptivity

For general MCMC algorithm, it has been shown that an adaptive Markov chain can lead to erroneous results [30]. Practically, for a given probability π, there might exist a sequence P_i of Markov kernels with limiting distribution π for all i such that for a given X_0, the sequence of random variables defined by $X_i \sim P_i(X_{i-1}, .)$ does not converge to the limiting distribution π. This does not affect the Wang-Landau algorithm itself. However, any adaptivity must be stopped before the end of the algorithm. The choice we make is to stop any adaptivity once the flat histogram has been met a given number of times denoted N_{FHE} in the sequel.

3 Methods: improving convergence speed

3.1 Rationale and difficulties targeted

Diffusivity across strata. As outlined in section 2, the convergence of the Wang-Landau algorithm largely relies on a suitable choice for the proposal q and the energy discretisation. In particular, a suitable proposal q and energy discretisation should leave no energy stratum poorly connected. The problem is stringent for energy strata near local minima and maxima where two difficulties are faced: accessing the strata, and leaving them.

For the particular case of local minima which is of interest in the sequel, accessing a low energy stratum is especially hard in high dimensional spaces: the volume of such strata being (in general) exponentially small in the dimension, they may be visited very seldom, resulting in an underestimation of the DoS θ_i.

In turn, this underestimation makes it difficult to leave the stratum. To see why, assume that the stratum i has been reached, and that the proposal q attempts a move from say $x \in \mathcal{E}_i$ to $y \in \mathcal{E}_{i+1}$. By the M-H criterion Eq. (7) applied to the target distribution Eq. (8), this move is accepted with probability proportional to

$$\frac{\pi(y)}{\pi(x)} \frac{\theta_{J(x)}}{\theta_{J(y)}} \frac{q(y,x)}{q(x,y)}. \tag{10}$$

Assuming that the ratio between $q(y,x)$ and $q(x,y)$ is bounded, if $\theta_{J(x)}$ is severely underestimated, the move upward will be accepted with very low probability.

As suggested by this discussion, we aim for a *ladder-like* random walk connecting each energy level with the one bellow and the one on top with an as high as possible probability (Fig. 1). Such a random walk is termed *diffusive.* To this end, the following difficulties must be overcome by the proposal q:

- Difficulty 1: **no overstep** – section 3.3. To avoid overstepping strata, we introduce a topography adapted proposal which exploits the geometry of level set surfaces of the landscape.

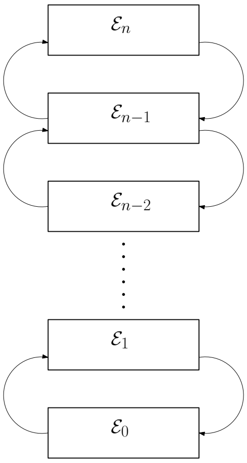
- Difficulty 2: **no congestion** – section 3.4. We develop a high dimensional geometry aware biasing strategy to promote the effective move across strata. This strategy learns the right aperture angle of cones encompassing the level set surfaces of strata below/above. In high dimensions, this is a difficult endeavor since proposals exploring uniformly the entire space face difficulties to hit these strata.

- Difficulty 3: **multimodal distributions** – section 3.5. To deal with the case of multimodal distributions, we resort to darting, a strategy meant to connect parts of the energy landscape which are separated by energetic or entropic barriers.

- Difficulty 4: **energy range discretization** – section 3.6. Slow convergence may be due to an inappropriate energy discretization. We resort to a refinement strategy to fix such problems.

Remark 1. *It may be noticed that Eq. (10) simplifies in two settings: the first ratio is equal to one with π is uniform – see also Eq. 3; the last one is also equal to one when the proposal is symmetric. In that case, one is left with $\theta_{J(x)}/\theta_{J(y)}$.*

Figure 1 Diffusivity between strata in WL. We consider a potential energy function U whose strata or energy levels are denoted $\mathcal{E}_i = \{x \in \mathcal{E} | U(x) \in [U_{i-1}, U_i)\}$. Energy levels may be seen as the nodes of a graph and may be connected in a variety of ways. In this work, we exploit a proposal q, which, via the Metropolis-Hastings criterion, connects strata in the WL algorithm in a ladder-like fashion. This strategy is especially well suited to sample a basin of arbitrary geometry.



3.2 Metrics to assess the diffusivity

We now introduce criteria to be correlated with the convergence of WL. We actually assess the notion of diffusivity just introduced in two ways: (i) using the moves proposed by q to study the diffusivity of the proposal, and (ii) using the moves proposed by q and processed by the WL M-H criterion to study the diffusivity of WL. To present the criteria based on these two types of moves, recall that there are d strata denoted $\mathcal{E}_1, \dots, \mathcal{E}_d$.

Descending times. As a global metric targeting moves accepted by WL, we resort to so-called descending times:

Definition. 1. *A* descent *across d strata is defined by two times t_0 and t_1 such that*

- $x_{t_0} \in \mathcal{E}_d$ *and* $x_{t_0 - 1} \notin \mathcal{E}_d$.
- $x_{t_1} \in \mathcal{E}_1$ *and* $\forall t \in [t_0, t_1[$, $x_t \notin \mathcal{E}_1$.

The descending time *is then $t_1 - t_0$.*

Aggregated transition matrices (ATM). To further our analysis, we also define:

Definition. 2. *Consider an execution of WL. The* aggregated transition matrix *for q, denoted ATM_q, is the $d \times (d+1)$ row stochastic matrix providing the frequencies of transitions between strata corresponding to the moves proposed by q along the execution. (Nb: column $d+1$ corresponds to moves ending outside the bounded region $\mathcal{E}$.)*

The aggregated transition matrix *for WL, denoted ATM_{WL}, is the $d \times d$ row stochastic matrix providing the frequencies of transitions within WL.*

These definitions call for the following comments:

- The entries of the matrices are frequencies observed over an execution run of WL, rather than probabilities.
- For the sake of coherence, we represent matrices ATM_q and $\mathrm{ATM}_{\mathrm{WL}}$ as $d \times (d+1)$ matrices, even though the last column differs: it may be populated for ATM_q, but is never so for $\mathrm{ATM}_{\mathrm{WL}}$. Indeed, in the course of the WL algorithm, moves leaving the strata are rejected.
- Matrix $\mathrm{ATM}_{\mathrm{WL}}$ is not the matrix of a discrete Markov chain whose state space is the set of strata. Indeed, transitions are recorded for the evolving kernels P_θ used in the course of the WL algorithm.
- Finally and most importantly, these matrices are meant to provide a concise encoding of the diffusivity of the random walks. In short, for a given row of either matrix, consider the connected components of the populated cells–that is the strata which are accessible from a given stratum. Ideally, one expects one connected component of width at least three, meaning that from a given stratum, one stays in place or visits the strata above and below in energy.

Remark 2. *Descending times and aggregated transition matrices target different purposes. On the one hand, descending times only use the extreme strata, while transition matrices encode information involving all strata. On the other hand, descending times are easily localized. Consider a situation where the potential energy landscapes has two regions characterized by* low *and* large *descending times. Spatializing descending times, e.g. by clustering, is easier than defining localized aggregated transition matrices.*

3.3 No overstepping across strata

3.3.1 Problem

Strata of small *thickness* tend to be stepped over. This typically happens when the landscape is *steep* or the discretization is fine. It is thus important to adapt the travel distance of the proposal in such regions.

A Gaussian mixture identical for all strata has been used [12]. However, the mixture is symmetric (see section 2.4) and does not exploit the geometry of the landscape.

3.3.2 Solution

Rationale. We estimate the local *shape* of the energy function using a Taylor expansion of the energy. We use an order two expansion since the gradient vanishes near local minima.

We first sample a vector u uniformly at random in the unit sphere $\mathbb{S}^{n-1}$. We then compute the Taylor expansion in the direction of u with $h \in \mathbb{R}$:

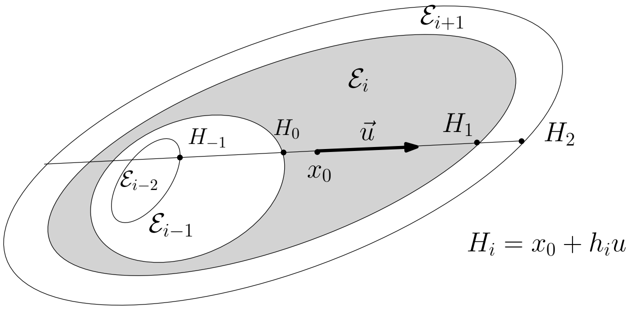
$$U(x + hu) = U(x) + h(\nabla U \cdot u) + \frac{1}{2}h^2(u^T \mathrm{Hess}\, u). \tag{11}$$

Assuming that $x \in \mathcal{E}_i$, using the Taylor expansion, we compute the interval $[h_0, h_1]$ such that for $h \in [h_0, h_1]$ (Fig. 2)

$$x + hu \in \mathcal{E}_i \tag{12}$$

Doing the same for $\mathcal{E}_{i-1}$ and $\mathcal{E}_{i+1}$ yields $[h_{-1}, h_0]$ and $[h_1, h_2]$. The last steps are to pick any of these 3 intervals with probability 1/3 and to sample h uniformly in the chosen interval. Doing so effectively adapts the proposal to the local shape of the energy landscape, allowing multiple scales. Even better, it also changes and adapts to the chosen direction.

Figure 2 Proposal exploiting the geometry of the landscape to avoid overstepping strata: a move from $x_0 \in \mathcal{E}_i$ should either stay in $\mathcal{E}_i$, move to $\mathcal{E}_{i-1}$, or move to $\mathcal{E}_{i+1}$. The intersection between a random line through x_0 with the level set surfaces of a quadratic approximation of the potential yields points $\{H_i\}$ defining three segments $[H_{-1}, H_0], [H_0, H_1], [H_1, H_2]$. Each such segment is chosen at random with probability 1/3, and a point is generated uniformly at random within the chosen segment. Note that the same random line can be obtained for two opposite vectors u and $-u$.



Probability for the proposal. We now set up the formulae. Assume that the vector u is sampled uniformly at random in the unit sphere $\mathbb{S}^{n-1}$. Also compute the h_i from from the Taylor expansion at x in the direction of vector $u = \frac{y-x}{\|y-x\|}$. Then, the probability density of going from x to y for any x and y with respect to the Lebesgue measure on $\{x + \mathbb{R}u\}$ is

$$P_{\text{no-overstep}}(x, y \mid u) = \mathbf{1}_{\{x+\mathbb{R}u\}}(y) \sum_{i=0}^{2} \mathbf{1}_{[h_{i-1},h_i]}(< y - x, u >) \frac{1}{3|h_i - h_{i-1}|} \tag{13}$$

Consider also the uniform probability density on S^{n-1}:

$$P_{dir}^{unif}(u) = 1/Area(S^{n-1}) = \frac{\Gamma(n/2)}{2\pi^{n/2}} \tag{14}$$

Taking into account the fact that vectors u and $-u$ yield the same intervals, and that we consider the sphere of radius $\|y - x\|$, the final proposal reads as

$$q_{\text{no-overstep}}^{unif}(x, y) = 2 \frac{1}{\|y - x\|^{n-1}} P_{dir}^{unif}\left(\frac{y - x}{\|y - x\|}\right) P_{\text{no-overstep}}\left(x, y \mid u = \frac{y - x}{\|y - x\|}\right). \tag{15}$$

Remark 3. *Equation (11) explicitly uses the Hessian. In practice, the second order directional term is estimated numerically using the gradient.*

Remark 4. *To understand the interest of the previous calculation, the analysis of an overly simplified case is of interest. Assume that the domain $\mathcal{E}$ is a cube, and the level set surfaces bounding the strata are hyperplanes parallel to one face. We call the width $W(\mathcal{E}_i)$ of a stratum $\mathcal{E}_i$ the distance between its two bounding planes. Assume the strata $\mathcal{E}_{i_x} \ni x$ and $\mathcal{E}_{i_y} \ni y$ are consecutive. It is easy to show using Thales theorem that if the boundary of the domain is not hit:*

$$\frac{q_{no\text{-}overstep}^{unif}(y, x)}{q_{no\text{-}overstep}^{unif}(x, y)} = \frac{W(\mathcal{E}_{i_y})}{W(\mathcal{E}_{i_x})}.$$

Furthermore, in the cube:

$$\frac{W(\mathcal{E}_{i_x})}{W(\mathcal{E}_{i_y})} = \frac{Volume(\mathcal{E}_{i_x})}{Volume(\mathcal{E}_{i_y})} = \frac{\theta_{i_x}^*}{\theta_{i_y}^*}.$$

Therefore, from Eq. (10), the acceptance ratio in Metropolis-Hastings (eq. 10) in Wang-Landau when the vector θ is close to θ^* is close to 1. It effectively brings the asymptotic rejection rate of the Wang-Landau random walk to 0.

3.4 No congestion: dealing with high dimensionality and concentration

3.4.1 Problem

As noticed in section 3.1, we wish to design a proposal diffusing from $\mathcal{E}_i$ to $\mathcal{E}_{i-1}$ and $\mathcal{E}_{i+1}$. In the sequel, we present a strategy meeting this goal, based on a random direction in a suitable cone, in conjunction with the no overstep strategy from section 3.3. Our presentation focuses on moves downward which is helpful near local minima, yet readily generalizes to moves upward for the case of local maxima.

Consider the problem of lowering the energy by moving from stratum $\mathcal{E}_i$ to $\mathcal{E}_{i-1}$ (Fig. 3), starting at point x_0. To do so, a direction in a cone of angle $\alpha(x_0, \mathcal{E}_{i-1})$ must be chosen because directions outside this cone do not intersect $\mathcal{E}_{i-1}$. As the dimension increases, the probability of sampling a point in a cone of fixed aperture decreases exponentially with the dimension (Fig. 4), preventing the proposal to reach the stratum $\mathcal{E}_{i-1}$.

Figure 3 Reaching region $\mathcal{E}_{i-1}$ from $\mathcal{E}_i$ near a local minimum: cone of suitable directions. Note that a similar problem is faced to move upward near a local maximum.

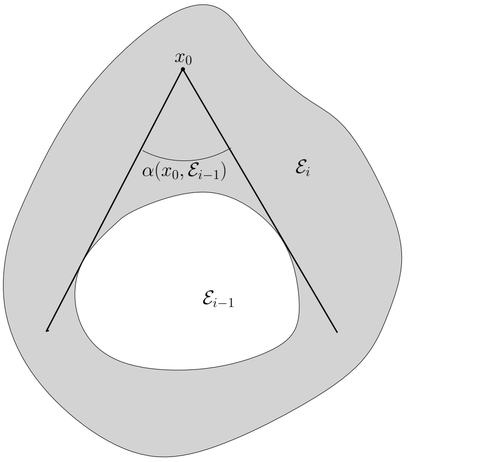
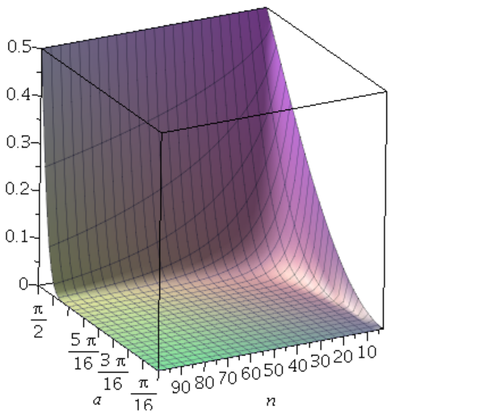


Figure 4 Ratio between the area of the spherical cap subtained by an angle θ and that of the whole n-dimensional hemisphere $S^{n-1}/2$ – see Eq. (36 in Appendix. Ranges explored: dimension $n \in [3, 100]$, and angle $\theta \in [0, \pi/2]$.



3.4.2 Solution

Rationale. The most straightforward way to overcome this problem is to decrease the bin size. Indeed doing so makes $\alpha(x_0, \mathcal{E}_{i-1})$ closer to $\pi/2$. In practice, the bin splitting strategy from [12] achieves this goal. However, the number of required strata increases with dimension, making this strategy less effective.

Let $C_{\text{down}}(x_0, \mathcal{E}_{i-1}) \subset S^{n-1}$ the subset of directions which allow moving into $\mathcal{E}_{i-1}$ from a point $x_0 \in \mathcal{E}_i$. Ideally, we could avoid splitting bins if we could bias the choice of direction towards this subset. An approximation of $C_{\text{down}}(x_0, \mathcal{E}_{i-1})$ could be found using a full second order Taylor expansion (thus requiring the full Hessian matrix). However this would be computationally very costly and the sampling procedure on such an approximation is unknown. Therefore, we limit ourselves to isotropic estimations of $C_{\text{down}}(x_0, \mathcal{E}_{i-1})$, i.e. $C_{\text{down}}(x_0, \mathcal{E}_{i-1})$ is estimated by an isotropic cone of angle $\alpha(x_0, \mathcal{E}_{i-1})$.

At a given point, consider the gradient of U. We take ∇U as the cone axis since it is the most natural candidate in the absence of additional information. In contrast, the aperture angle $\alpha(x_0, \mathcal{E}_{i-1})$ depends on the geometry of the level set bounding $\mathcal{E}_{i-1}$ – a global information unavailable. We bypass this problem by assuming that a single aperture $\alpha(\mathcal{E}_i, \mathcal{E}_{i-1})$ is suitable for the majority of points in $\mathcal{E}_i$. This assumption holds if for example the curvature and width of stratum do not vary too much. The aperture can now be estimated during runtime. However, as noted in 2.5, adaptivity in MCMC algorithms can prevent convergence. Therefore, if the flat histogram criterion is used, we stop the learning procedure when the flat histogram criterion has been reached N_{FHE} times .

Estimating the aperture angle. For a stratum $\mathcal{E}_i$, we wish to select an angle amidst a predefined set $\{\alpha_i^{(0)}, ..., \alpha_i^{(k)}\}$. We apply the following procedure–which is independent from the generation of x_{t+1}. For a given point $x_t \in \mathcal{E}_i$ sampled by Wang-Landau, consider the cones of apex x_t, axis $\nabla U(x_t)$, and aperture angles $\alpha_i^{(j)}$. We sample M vectors uniformly in each cone, and check for each of them whether the stratum $\mathcal{E}_{i-1}$ can be reached. Assume m points x_t has been sampled. We compute for each cone the probability to reach $\mathcal{E}_{i-1}$ over the $m \times M$ vectors sampled. Since this probability is expected to be large for cones of small aperture, we pick the largest cone whose probability is larger than a user specified threshold p_c. If no such angle exists, we use the fallback bin split strategy of section 3.6.

Probability for the proposal. Assume one knows how to sample uniformly at random a point in a cone, whence a random direction defined by the apex of the cone and that point – SI Section 6. With $\mathbf{1}_{cone}(\cdot)$ the indicator function of the cone, the corresponding probability reads as

$$P_{dir}^{cone}(u) = \mathbf{1}_{cone}(u)/Area(S^{n-1} \cap cone). \tag{16}$$

We introduce the following proposal based upon the no overstep proposal in the direction just chosen:

$$q_{\text{no-overstep}}^{cone}(x,y) = 2\frac{1}{\|y-x\|^{n-1}} P_{dir}^{cone}\left(\frac{y-x}{\|y-x\|}\right) P_{\text{no-overstep}}\left(x, y \mid u = \frac{y-x}{\|y-x\|}\right) \tag{17}$$

Remark 5. *Similarly to Eq. (15), Eq. (17) uses both vectors u and $-u$ indistinguishably, even though our focus in on moving downwards. This is a design choice meant to make the implementation easier. In the worst case, a factor of two is lost to move downward.*

Remark 6. *The previous strategy, which uses a dictionary of cone apertures, calls for the following comments:*

- *The cone strategy combines local and non local pieces of information. The local information resides in the gradient. The non local information resides in the geometry of level set surfaces bounding the strata; this geometry is indeed accounted for indirectly by learning the right aperture angle. Using non local information makes it different from Hamiltonian Monte Carlo (HMC) [31] or Metropolis-adjusted Langevin algorithm (MALA) [32].*

- *The aperture angle of the cone is actually critical. Consider the set of directions delimited on S^{n-1} by a given cone. When the dimension increases, the mass of this set of directions concentrates on the boundary of the cone. Therefore, if the cone aperture is overestimated, sampled directions will end up with high probability in this region, and the corresponding proposal will miss the targeted stratum.*
- *Since the used cones are isotropic, one expects the method to be less efficient to handle highly non isotropic cases. However, moderately non isotropic cases are handled as well – see Experiments.*
- *Even if a suitable cone is found by the previous procedure, the Metropolis-Hasting acceptance rate might be low – for instance if strata are too wide or if the curvatures of level sets non constant.*

3.5 Handling multimodal distributions via darting

3.5.1 Problem

Classical proposals face difficulties to connect basins which are separated by energetic or entropic barriers (regions with high density of states but possibly no local minimum) [33, 34], inducing long mixing times.

3.5.2 Solution

Assuming one has a *a priori* knowledge of the positions $m_1, .., m_K$ of theses minima, it is natural to introduce another type of move allowing jumps from one minima to another. To implement this, we use a darting strategy – see [26, 27].

Darting in its simplest form defines a radius ρ, then adds transitions between the balls $B(m_i, \rho)$. In practice, if $x_t \in B(m_i, \rho)$, one picks a ball at random–let j be its index, and proposes the following move: $x_{t+1} \sim Unif(B(m_j, \rho))$. However the balls $B(m_i, \rho)$ do not match the level set surfaces of their respective basins. Hence $U(x_{t+1}) - U(m_j)$ might be much larger than $U(x_t) - U(m_i)$, leading to poor acceptance rates in the Wang-Landau algorithm.

Choice of the candidate point. Denoting m_{k_t} the local minimum whose basin contains x_t, define $\Delta U_t = U(x_t) - U(m_{k_t})$. Our rationale to optimize the acceptance ratio in Wang-Landau is to control both $\Delta U_t - \Delta U_{t+1}$ and the ratio $q(y, x)/q(x, y)$. For the former, we proceed as follows in two steps. First, we chose a target energy. For a given x_t, let k be the index of the minimum chosen at random. For some $\beta > 0$, we choose a target energy

$$T_U \sim Unif(U(m_k) + \Delta U_t - \beta, U(m_k) + \Delta U_t + \beta). \tag{18}$$

Second, we propose a point x_{t+1} such that $U(x_{t+1}) = T_U$ (Fig. 6). To this end, we sample a vector u uniformly in the ellipsoid defined by $v^T H(k)v = 1$ where $H(k)$ is the Hessian of U at m_k. Then we do a line search to find the intersection between the target energy T_U and the half line $m_k + \mathbb{R}^+ u$ (Fig. 5).

When to jump. The strategy just described requires a line search which is expensive if the target point is far from the local minimum. Furthermore, under some assumptions which are true if jumps are only allowed close to the local minima, the expression of the transition kernel can be simplified. Hence we introduce S a user defined parameter, and the darting proposal is only used if $\Delta U_t \leq S$.

Figure 5 Darting: reaching a prescribed energy level set via line search in the direction of vector u. The line search starts from minimum m_k with target energy $Target$.

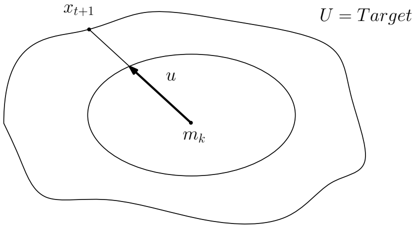
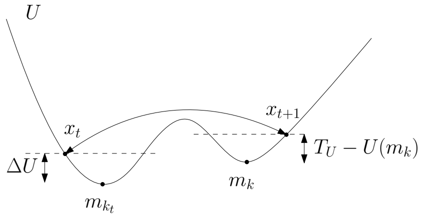


Figure 6 Handling multimodal distributions via darting: jumping between two local minima. While darting, the difference of energy with the local minima is controlled to monitor the acceptance rate. Note that $\Delta U = U(x_t) - U(m_{k_t})$ and $T_U \sim Unif(U(m_k) + \Delta U_t - \beta, U(m_k) + \Delta U_t + \beta)$



Transition kernel. Computing the transition kernel of the darting move is non trivial, but can be done using a suitable change of variables. The full computation is detailed in the appendix – section 7, however for the sake of conciseness, we only give here the final result. For any x and y in $\mathbb{R}^n$, let k be the closest minima to y, $I_k = [U(m_k) + \Delta U - \beta, U(m_k) + \Delta U + \beta]$, λ_i and e_i the eigenvalues and eigenvectors of the Hessian of U at m_k. Finally, denoting $< u, v >$ the dot product, let

$$l = \sqrt{\sum_i \lambda_i < y - m_k, e_i >^2}. \tag{19}$$

Using the latter and denoting $\mathbf{1}_{I_k}(u)$ the indicator functor for the interval I_k, the transition probability is given by:

$$q_{\text{darting}}(x, y) = \frac{1}{K} \frac{\Gamma\left(\frac{n}{2}\right) 2\beta}{2\pi^{\frac{n}{2}}} \mathbf{1}_{I_k}(U(y)) \frac{1}{l^n} \nabla U(y)^T (y - m_k) \prod_{i \leq n} \sqrt{\lambda_i}. \tag{20}$$

3.6 Splitting energy bins

Splitting an energy bin may prompt two scenarios: it can improve the mixing time of the random walk, and thus improve both the precision of the estimated DoS and the complexity to get it; or it has no effect on the mixing time, resulting an improvement of the precision at the expense of the computational burden. The goal is to refine bins in the former case only. Thus we only split bins when the cone strategy fails, as a fallback. We monitor the failure of the cone strategy by computing the proportion of steps in which the random walk has the *possibility* to go up or down in energy (see 3.3) and the success rate of the Metropolis-Hasting criterion when going up or down. If either of these statistics are below a user defined threshold for a given bin, the bin is halved.

3.7 The final combined proposal

The final proposal combines the building blocks introduced in sections 3.3, 3.4 and 3.5, plus a Gaussian proposal with transition q_{gauss}. Let $p(x) = (p_{\text{no-overstep}}(x), p_{\text{gauss}}(x), p_{\text{darting}}(x))$ such that $p_{\text{no-overstep}}(x) + p_{\text{gauss}}(x) + p_{\text{darting}}(x) = 1$ for all $x \in \mathbb{R}^n$. In addition, let $p_{\text{cone}}(x) < 1$ be the probability of modifying the no overstep strategy by using the cone. The final proposal is

- choose one of the proposals at random with probability vector $p(x_t)$,
- if the no overstep strategy is selected, use the cone with probability $p_{\text{cone}}(x)$ to choose the direction along which the proposal acts,
- sample the point x_{t+1} according to the chosen proposal.

The combined proposal therefore has the following transition probability:

$$\begin{aligned} q(x,y) = {} & p_{\text{no-overstep}}(x)\left[(1-p_{\text{cone}}(x))q^{unif}_{\text{no-overstep}}(x,y) + p_{\text{cone}}(x)q^{cone}_{\text{no-overstep}}(x,y)\right] \\ & + p_{\text{darting}}(x)q_{\text{darting}}(x,y) + p_{\text{gauss}}(x)q_{\text{gauss}}(x,y) \end{aligned} \tag{21}$$

Remark 7. *Eq. 21 calls for two comments:*

- *Since the proposal q is used in the Metropolis-Hastings algorithm, it is crucial to be able to compute $q(x,y)$ for any x and y.*
- *The cone based strategy restricts the set of directions explored. The constraint $p_{cone}(x) < 1$ leaves open the probability to pick any direction on the unit sphere.*

4 Experiments

4.1 Setup

4.1.1 Implementation and parameters

Our implementation of a fully generic Wang-Landau algorithm is described in the companion paper [35]. The corresponding C++ generic code, which allows tuning the physical system, as well as the main ingredients of the algorithm is currently being integrated to the Structural Biology Library – see [36] and `http://sbl.inria.fr`.

In the tests presented below, the initial number of strata d was specified along with each example.

Otherwise, the following hyper-parameters were set as follows:

- Parameter from section 2.2:
  - Flat histogram threshold: $c = 0.1$ for synthetic models, $c = 0.05$ for the molecule.
- Parameters from section 3.4:
  - $k = 10$ cones, with the associated aperture angles sampled uniformly in $[0.2\pi/2, 0.8\pi/2]$,
  - $M = 1$ direction sampled in a cone,
  - $p_c = 0.4$ as probability threshold to select the largest possible cone.

- Parameters from section 3.5
    - Parameter $S = \infty$ and $\beta = 0.001$ used for darting.
- Parameters from section 3.7:
    - Relative weights of the three components of the proposal, i.e. $p_{\text{no-overstep}}(x), p_{\text{cone}}(x), p_{\text{darting}}(x)$, p_{gauss}. $p_{\text{cone}}(x)$ and $p_{\text{darting}}(x)$ are set to zero to turn off these strategies. Otherwise, the values are specified along with the examples.

4.1.2 Statistics of interest

Our experiments target three points: correctness, mixing time, and ability to handle complex systems. For the sake of conciseness, time t refers to t steps of Wang-Landau.

Correctness, stability, and convergence. When an analytical solution for θ^* is known, we simply resort to the relative error for estimates at time t, defined by:

$$\text{error}(t) = \sum_i \frac{|\theta_i^* - \theta_i(t)|}{\theta_i^*}. \tag{22}$$

We plot this function along time.

When no analytical solution is known–see the dialanine molecular model below, we assess convergence in two ways, based on several runs ($N = 60$). First, we plot an observable along time, akin to the partition function, at a fixed temperature:

$$\frac{Z}{\lambda(\mathcal{E})} = \frac{1}{\lambda(\mathcal{E})} \int_{\mathcal{E}} \exp\left(-U(x)/kT\right) \approx \sum_{\text{Energy levels U}} \theta_i^* \exp\left(-U/kT\right). \tag{23}$$

Second, we provide box plots on a per bin basis. We also resort to violin plots when more details are required in terms of distribution modes. (Recall that a violin plot displays a kernel density estimate of the data points processed.)

Mixing time. We use the descending times and the aggregated transition matrices–Section 3.1.

Remark 8. *Note that we do not normalize the descending times by the number of strata. Indeed, as seen in section 3.6, increasing the number of strata can decrease the mixing time. Hence the number of strata is a parameter tuned for convergence speed and therefore should not be taken into account when measuring mixing time.*

4.1.3 Contenders

As a yardstick, we compare our proposal (section 3.7) against the isotropic Gaussian proposals. However, while our proposal is used with its default parameters as specified above, the Gaussian variance needs to be tuned for a fair comparison [25]. Hence we compare with 3 Gaussian proposals (resp. high, adequate and low variances).

4.1.4 Models

Analytical models. We study three analytical models. The first model is the isotropic harmonic potential. The second is a non isotropic harmonic potential, to ensure that the algorithm behaves correctly in non trivial settings (Remark 6). The last model is a potential with two local minimum designed to study darting.

Molecular model. Finally, our last system is a classical toy molecular system, namely the blocked alanine peptide Ace-Ala-Nme (Fig 8), referred to as dialanine for the sake of conciseness. This system underwent extensive thermodynamic studies, using techniques as diverse as molecular dynamics [37], Monte Carlo and energy landscape based methods [38], or dimensionality reduction methods [39].

We use the amber99-sb force field in vacuum and aim to compute the density of state between -21 kcal/mol and 4 kcal/mol associated to one local minima – by enforcing the simulation to remain inside the basin of this local minimum ($\phi = 59.8862, \psi = -35.5193$).

Figure 7 Dual well function in dimension three. Plot of Eq. (24) for $x_i \in [-1, 1], x_i \in [-1/2, 1/2]$.

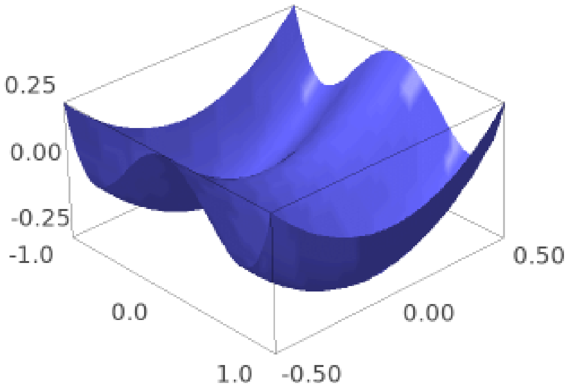
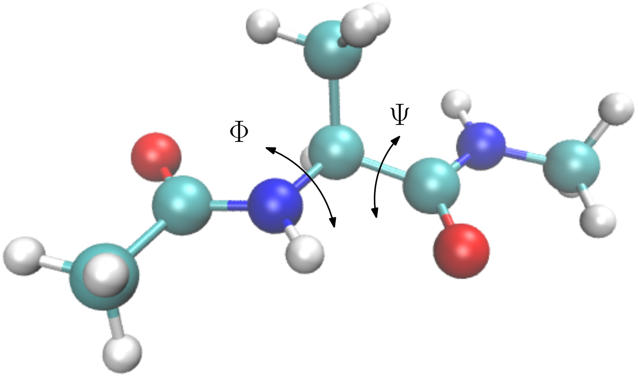


Figure 8 Dialanine (Ace-Ala-Nme) and the two dihedral angles Φ and Ψ



4.2 Results

4.2.1 Single well potential: isotropic

We study here the simple harmonic model

$$U(x) = \sum x_i^2$$

with state space the unit ball in dimension $n = 25$. The energy discretisation is $[a_i, a_{i+1}]$ with $a_i = i/10$ for $i \in \{0..9\}$. The improved proposal is defined with $p_{\text{no-overstep}} = 1$, $p_{\text{gauss}} = 0$, $p_{\text{darting}} = 0$, and $p_{\text{cone}} = 0.5$. The flat histogram threshold is set to $c = 0.1$. Since the exact result is known, we plot the exact error.

Let us first focus on the error (Fig. 9(Left)). For the Gaussian proposals, the best scale is $\sigma = 0.1$, as smaller and larger values (resp. $\sigma = 0.01$ and $\sigma = 1$) yield large errors. Our proposal based on the no overstep alone is on par with the best Gaussian. Moreover, adding the cone yields the best results. This radical improvement owes to a descending time orders of magnitude smaller for the cone strategy (Fig. 9(Right)). These observations are coherent with the ATM matrices (Fig. 10), which show that (i) the proposals based on Gaussians face a high diversity in terms of strata reached, (ii) our proposal has an ideal diffusivity encoded in a *tridiagonal* matrix, (iii) the cone improves the probability to move downward. (Compare e.g. the probabilities to move from the energy slice $[0.1, 0.2]$ to $[0.0, 0.1]$. Equivalently, one sees a gradient of colors when moving upward and to the left in the ATM matrices from the third and fourth row of Fig. 10.)

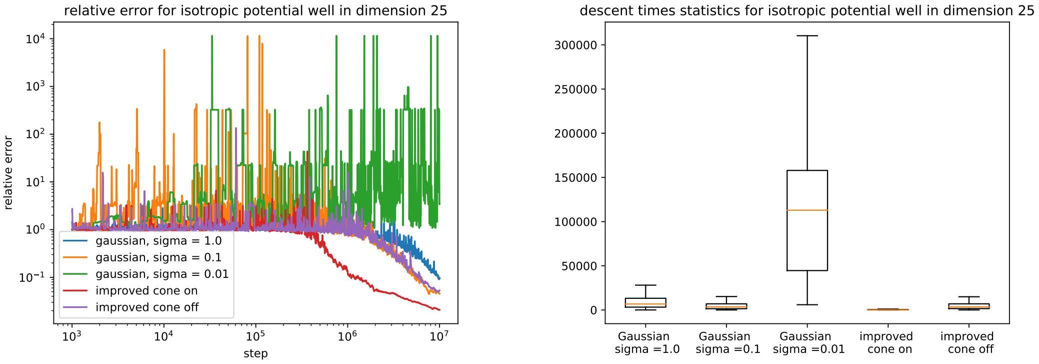
Inspection of Fig. 10 also yields three interesting points:

- The WL Metropolis-Hastings criterion also balances the probabilities to enter and leave a stratum, as shown by the comparable colors apart from the diagonal for ATM_{WL} – as opposed to ATM_q.

- While doing so, it refuses moves, which results in populating the diagonal.

- Matrix ATM_q shows it is easier to climb strata than to go downhill.

Figure 9 Isotropic single well in dimension 25: comparison of the five proposals. Values have been averaged over 30 runs. The five proposals used are the three Gaussian based proposals, plus our combined proposal with and without the cone improvement. **(Left) Comparison of the evolution of relative error – Eq. (22) (Right) Box plot of the descending times.**



4.2.2 Single well potential: non isotropic

To challenge all methods with a non-isotropic case (Remark 6), we use the following non isotropic potential energy:

$$U(x) = \sum_{i=1}^{n} i x_i^2$$

Also in dimension $n = 25$, the energy discretisation is $[a_i, a_{i+1}]$ with $a_i = i/10$ for $i \in \{0..9\}$. The improved proposal is defined with $p_{\text{no-overstep}} = 1$, $p_{\text{gauss}} = 0$, $p_{\text{darting}} = 0$, and $p_{\text{cone}} = 0.5$. The flat histogram threshold is set to $c = 0.1$. The results are on par with the isotropic case (SI Figs. 18 and 19), showing in particular the effectiveness of the cone strategy in this anisotropic setting.

Figure 10 Isotropic single well, $n = 25$, : aggregated transition matrices. One line or row corresponds to one stratum. For a given matrix line, the color coding indicates the probabilities (log scale) to move from a stratum to the remaining ones. Note that the matrices of our proposal are tridiagonal.

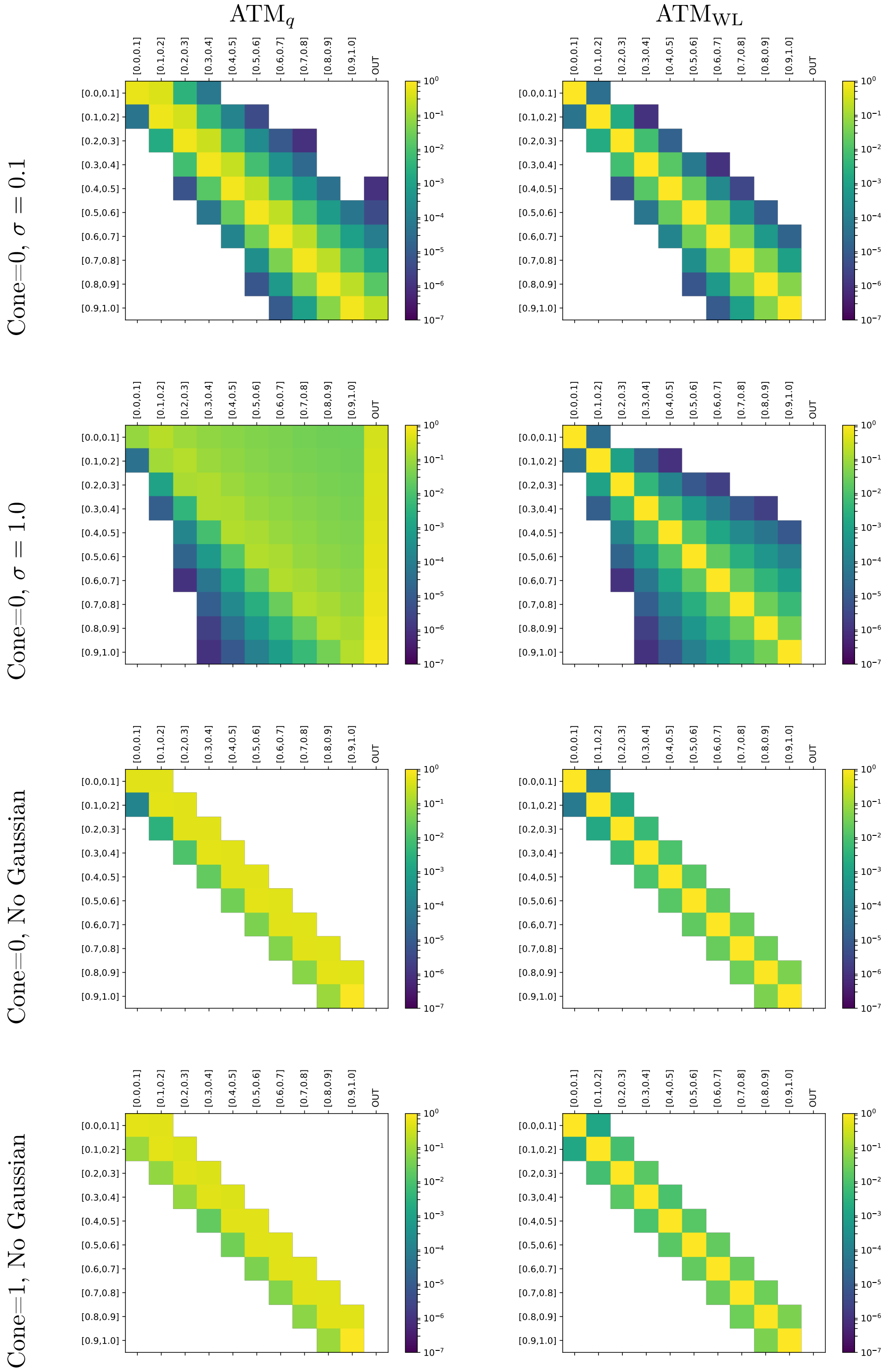
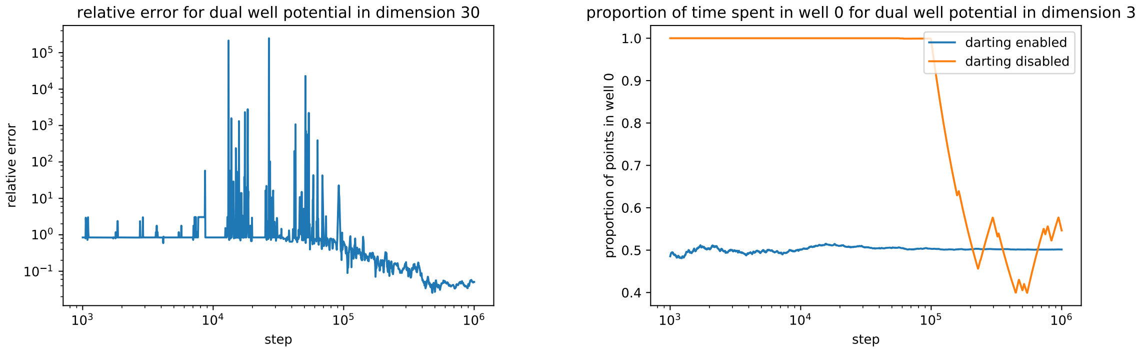


Figure 11 Dual well potential in dimension 30: analysis of darting. (Left) Evolution of relative error of Eq. (22) when using darting. (Right) Comparison of time spent in the first well with and without darting.



4.2.3 Dual wells potential: darting

To challenge darting, we use the usual one-dimensional dual well potential energy function $x^4 - x^2$, and add a quadratic potential in other dimensions:

$$U(x) = x_1^4 - x_1^2 + \sum_{i=2}^{n} x_i^2. \tag{24}$$

This potential energy has 2 local minimum at $(-\frac{1}{\sqrt{2}}, 0, ..., 0)$ and $(\frac{1}{\sqrt{2}}, 0, ..., 0)$ (Fig. 7). The additional coordinates makes it harder to travel from one minimum to the other by making it hard to choose suitable directions. More specifically, direction x_1 corresponds to an energetic barrier, while the remaining directions add volume in phase space, which corresponds to an entropic barrier [33].

We setup the darting proposal with these two minima, and compare our proposal with and without darting. The energy discretisation is $-0.25 < 0 < 0.2 < 0.4 < 0.6 < 0.8 < 1$. The proposal is set up with $p_{\text{darting}} = 0.5$, $p_{\text{no-overstep}} = 0.45$, $p_{\text{gauss}} = 0.05, p_{\text{cone}} = 0.5$ when darting is enabled and $p_{\text{darting}} = 0$, $p_{\text{no-overstep}} = 0.9$, $p_{gauss} = 0.1, p_{\text{cone}} = 0.5$ when darting is disabled. The Gaussian move variance is set to $\sigma = 1$. The flat histogram threshold is set to $c = 0.1$.

Both methods yield correct values (Fig. 11(Left)). (Data not shown for darting disabled: since the potential energy function is symmetric for the first coordinate, the algorithm computes the correct value even if it never crosses the energy barrier.)

It appears that the proposal allows crossing the energy barrier almost instantly, while with darting disabled the first jump appears after 10^5 samples (Fig. 11(Right)). This induces a large difference in the mixing time.

4.2.4 Dialanine

On this system, the results reported thereafter were obtained using the cone improvement, as convergence could not be obtained without it.

To compute the value defined by Eq. (23) restricted to the basin of the local minimum with torsion angles, we enforce the simulation to remain within this basin. Checking whether a point is in a given minima basin of attraction requires a minimization of the potential energy. Since this is costly operation, we check this condition every N (=100) steps. If the random walk has escaped, we roll back to the latest point in this basin. (Note that this requires downgrading all statistics and random number generators [35].)

Remark 9. *The roll back strategy just described may introduce some bias, as an excursion outside the basin may not be detected. The effectiveness of this strategy relies on a bounded proportion of roll backs, see [35] for details.*

We perform 60 runs, each with 10^7 steps. The energy discretisation can be found in Fig. 12(Top Right)). The proposal was set up with no darting and no Gaussian moves: $p_{\text{no-overstep}} = 1$ and $p_{\text{cone}} = 0.6$. The flat histogram threshold is set to $c = 0.05$.

To analyze convergence, we plot the time evolution of the observable defined from the partition function at $T = 300K$ (Eq. (23), Fig. 12(Top Left)). Since all simulations use the same number of bins, we also provide a box plot for each bin Fig. 12(Top Right)). Finally, to check whether the observable is unimodal or not, we perform a violin plot at three different time frames along the course of the simulation (Fig. 12(Bottom)).

We note that the convergence was reached to a different extent as a function of the volume of strata: the smaller the volume of a stratum, the higher the variance of estimates. This is expected as sampling rare events is always more challenging. It appears, though, that the observable converges (Fig. 12(Bottom)), since values concentrate along a single mode.

To get further insights, we study the contribution of individual strata reweighted by Boltzmann's factor (Fig. 13). We first note that the energy range used is sufficiently large since contributions of the last stratum is one order of magnitude smaller than the highest one (Fig. 13(Left)). The more detailed violin plots–not in log scale, also shows that distributions within strata are unimodal.

The analysis of ATM reveals two interesting facts. First, we first observe that without cone, moving downward is more challenging whatever the energy level (Fig. 14 bottom left, lower diagonal of ATM_q). The cone improves this state of affairs, even though the PEL is not isotropic – a priori. Second, we note that the moves proposed using the cone are accepted to a significant extent (Fig. 14, comparison of lower diagonals of ATM_q and ATM_{WL}.)

Figure 12 Analysis of convergence for dialanine, using amber-69sb force field in vacuum. Results averaged over 60 independent simulations. **(Top Left) evolution of the partition function. (Top Right) Box plot of the estimation θ_i for each bin i. (Bottom) Violin plot of the partition function at $T = 300K$, at three different time frames along the course of the simulation.**

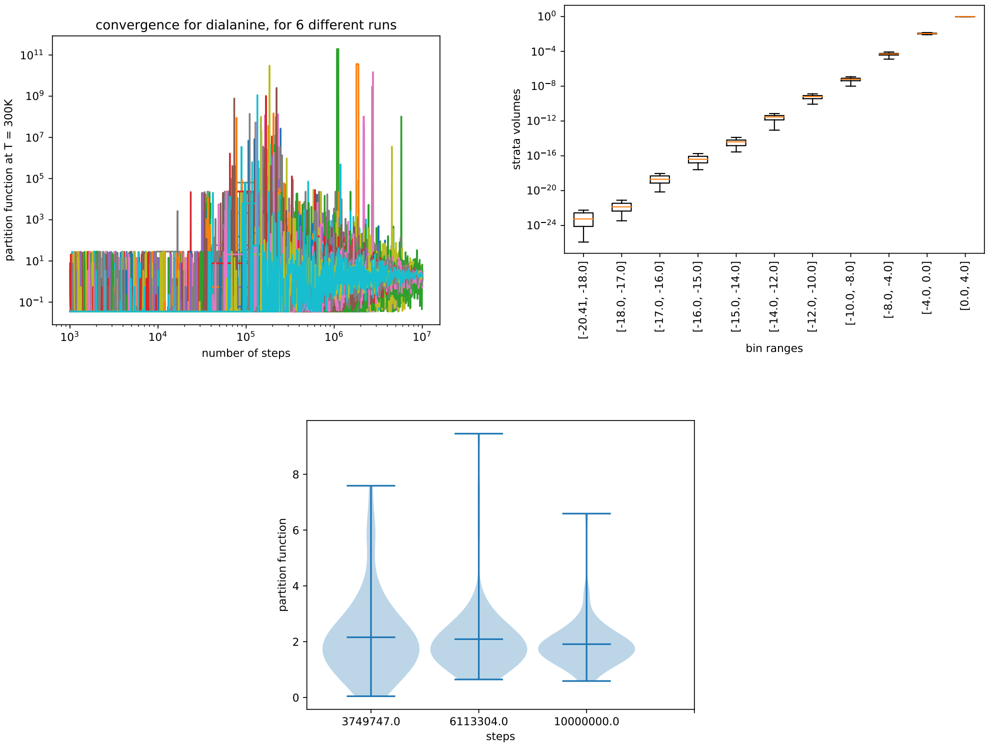


Figure 13 Dialanine with amber-69sb force field in vacuum: analysis of convergence. Results averaged over 60 independent simulations. **(Left) Box plot of the final bins volume with respect to the Boltzmann distribution at $T = 300K$. Log scale. (Right) Violin plot of the final bins volume with respect to the Boltzmann distribution at $T = 300K$**

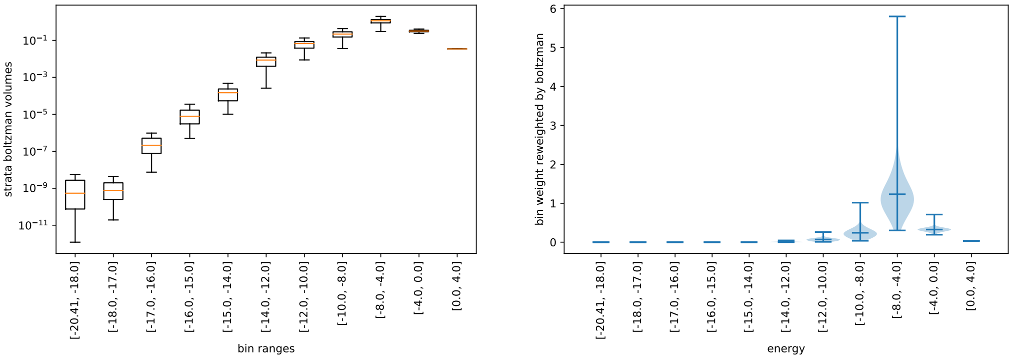
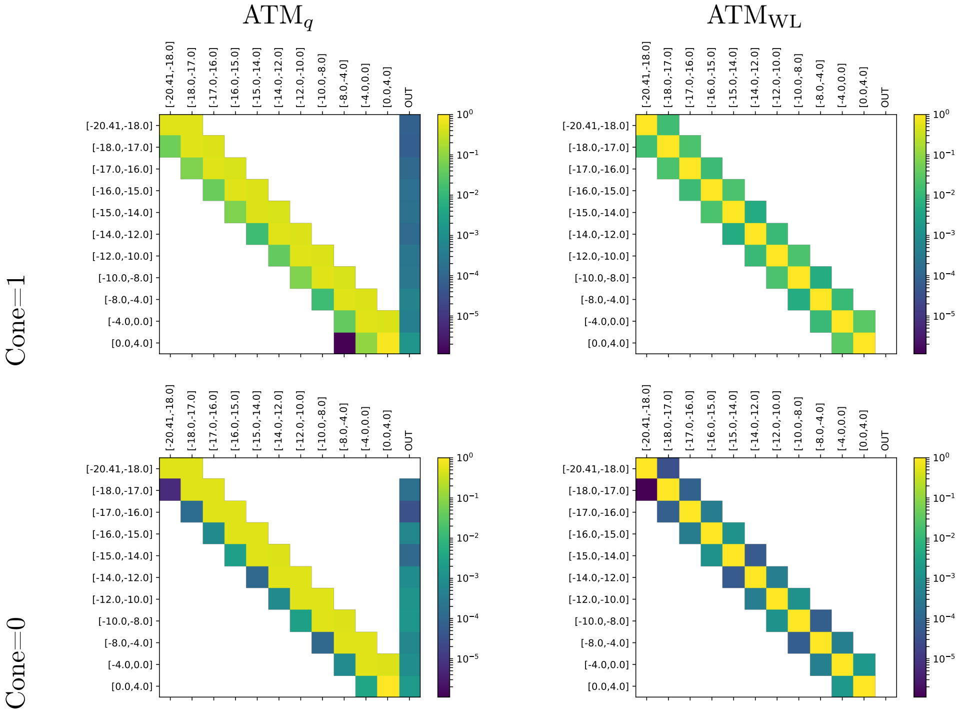


Figure 14 Dialanine with amber-69sb force field in vacuum: aggregated transition matrices. One line or row corresponds to one stratum. For a given matrix line, the color coding indicates the probabilities (log scale) to move from a stratum to the remaining ones. Note that the matrices of our proposal are tridiagonal.



5 Conclusion and outlook

Contributions. Given a physical system characterized by an energy, the Wang-Landau (WL) algorithm is a stochastic method returning an estimation of the density of states in terms of histogram. WL is an adaptive Monte-Carlo method inherently using a discrete set of values whose pre-images in phase space define regions called strata. To compute the volume of these strata, a core component of WL is the proposal used to generate candidate conformations and navigate between the strata. In this work, we make an explicit link between the convergence of the Wang-Landau algorithm and the underlying proposal, developing of a novel composite proposal targeting the following three difficulties: avoiding overstepping strata, avoiding congestion – remaining trapped within strata, and accommodating multimodal distributions. The performances of our proposal are assessed by measuring so-called descending times which quantify the diffusivity across strata, and so-called aggregated transition matrices which encode the diffusivity across strata. All in all, the resulting Wang-Landau algorithm is effective to compute observables for small biomolecules, within hours on a laptop computer.

We believe that a key feature of our proposal is its ability to exploit non local geometric information in the following sense: the no overstep strategy exploits the geometry of level set surfaces bounding the strata; the cone strategy implicitly combines information on the gradient and the level set surfaces too; finally, darting exploits the a priori knowledge on the location of local minima. We note in passing that the WL algorithm based upon a proposal using geometric information beyond the gradient makes it different in spirit from methods such as HMC or MALA which only use local geometric information. In addition, WL is based on an energy discretization yielding a piecewise continuous target probability distribution, while HMC and MALA primarily target continuous models. In fact, the combination of the original WL with a geometry aware proposal bears similarities with the multi-phase Monte Carlo sampling methods used to compute the volume of polytopes in high-dimensional spaces, as in both cases, the algorithm exploits a discretization based on strata – in a sequential rather than global way for polytope volume calculations.

Further work. We foresee stimulating questions in three complementary directions: design, analysis, and applications. On the design front, the cone based strategy is rooted in the exploitation of the geometry of level set surfaces near one local minimum or maximum. Therefore, designing a similar strategy to handle the descent into multiple basins would open new perspectives, possibly in conjunction with topological persistence and darting. On the analysis side, since level set surfaces bounding the strata are used, it would be particularly important to understand how the choice of strata affects the convergence of WL, and see whether error bounds (rather than asymptotic estimates) can be obtained for DoS estimates. Along the way, an analysis connecting aggregated transition matrices and the convergence of Wang-Landau would be of high interest. Finally, on the application side, while our proposal operates in Cartesian coordinates, switching to internal coordinates is an appealing strategy to handle biomolecules whose conformational changes are best described by valence and torsion angles.

References

[1] D. J. Wales. *Energy Landscapes.* Cambridge University Press, 2003.

[2] T. Lelièvre, G. Stoltz, and M. Rousset. *Free energy computations: A mathematical perspective.* World Scientific, 2010.

[3] W. Janke. Monte Carlo simulations in statistical physics: From basic principles to advanced applications. *Order, Disorder and Criticality: Advanced Problems of Phase Transition Theory*, 3:93–166, 2012.

[4] D. Landau and K. Binder. *A guide to Monte Carlo simulations in statistical physics.* Cambridge university press, 2014.

[5] F. Wang and D.P. Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical review letters*, 86(10):2050, 2001.

[6] D.P Landau, S-H. Tsai, and M. Exler. A new approach to Monte Carlo simulations in statistical physics: Wang-Landau sampling. *American Journal of Physics*, 72(10):1294–1302, 2004.

[7] RG Palmer. Broken ergodicity. *Advances in Physics*, 31(6):669–735, 1982.

[8] D. Wales and P. Salamon. Observation time scale, free-energy landscapes, and molecular symmetry. *Proceedings of the National Academy of Sciences*, 111(2):617–622, 2014.

[9] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[10] R.E. Belardinelli and V.D. Pereyra. Fast algorithm to calculate density of states. *Physical Review E*, 75(4):046701, 2007.

[11] R.E. Belardinelli and V.D. Pereyra. Wang–Landau algorithm: A theoretical analysis of the saturation of the error. *The Journal of chemical physics*, 127(18):184105, 2007.

[12] L. Bornn, P. Jacob, P. Del Moral, and A. Doucet. An adaptive interacting Wang–Landau algorithm for automatic density exploration. *Journal of Computational and Graphical Statistics*, 22(3):749–773, 2013.

[13] R. Belardinelli and V. Pereyra. Nonconvergence of the Wang-Landau algorithms with multiple random walkers. *Physical Review E*, 93(5):053306, 2016.

[14] P. Jacob and R. Ryder. The Wang–Landau algorithm reaches the flat histogram criterion in finite time. *The Annals of Applied Probability*, 24(1):34–53, 2014.

[15] G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre, and G. Stoltz. Convergence of the Wang-Landau algorithm. *Mathematics of Computation*, 84(295):2297–2327, 2015.

[16] F. Lou and P. Clote. Thermodynamics of RNA structures by wang–landau sampling. *Bioinformatics*, 26(12):i278–i286, 2010.

[17] P. Poulain, F. Calvo, R. Antoine, M. Broyer, and P. Dugourd. Performances of Wang-Landau algorithms for continuous systems. *Physical Review E*, 73(5):056704, 2006.

[18] Pedro Ojeda-May and Martin E Garcia. Electric field-driven disruption of a native β-sheet protein conformation and generation of a helix-structure. *Biophysical journal*, 99(2):595–599, 2010.

[19] A. Swetnam and M. Allen. Improving the wang–landau algorithm for polymers and proteins. *Journal of computational chemistry*, 32(5):816–821, 2011.

[20] W. Janke and W. Paul. Thermodynamics and structure of macromolecules from flat-histogram Monte Carlo simulations. *Soft matter*, 12(3):642–657, 2016.

[21] W. Atisattaponga and P. Marupanthornb. A 1/t algorithm with the density of two states for estimating multidimensional integrals. *Computer Physics Communications*, 220(122–128), 2017.

[22] S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability.* Springer Science & Business Media, 2012.

[23] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Annals of applied probability*, pages 1–9, 1998.

[24] M.K. Cowles and B.P. Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[25] G. Roberts and J. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

[26] I. Andricioaei, J.E. Straub, and A.F. Voter. Smart darting Monte Carlo. *The Journal of Chemical Physics*, 114(16):6994–7000, 2001.

[27] C. Sminchisescu and M. Welling. Generalized darting Monte Carlo. *Pattern Recognition*, 44(10):2738–2748, 2011.

[28] Z. Li and H.A. Scheraga. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *PNAS*, 84(19):6611–6615, 1987.

[29] A. Roth, T. Dreyfus, C.H. Robert, and F. Cazals. Hybridizing rapidly growing random trees and basin hopping yields an improved exploration of energy landscapes. *J. Comp. Chem.*, 37(8):739–752, 2016.

[30] J.S. Rosenthal and G.O. Roberts. Coupling and ergodicity of adaptive MCMC. *Journal of Applied Probablity*, 44:458–475, 2007.

[31] M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[32] M. Girolami and B. Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[33] C. Schön, M. Wevers, and M. Jansen. 'entropically'stabilized region on the energy landscape of an ionic solid. *Journal of Physics: Condensed Matter*, 15(32):5479, 2003.

[34] C. Schön and M. Jansen. Prediction, determination and validation of phase diagrams via the global study of energy landscapes. *Int. J. of Materials Research*, 100(2):135, 2009.

[35] A. Chevallier and F. Cazals. A generic framework for Wang-Landau type algorithms. *Submitted*, 2020.

[36] F. Cazals and T. Dreyfus. The Structural Bioinformatics Library: modeling in biomolecular science and beyond. *Bioinformatics*, 7(33):1–8, 2017.

[37] P. Smith. The alanine dipeptide free energy surface in solution. *The Journal of chemical physics*, 111(12):5568–5579, 1999.

[38] S. Somani and D. J. Wales. Energy landscapes and global thermodynamics for alanine peptides. *The Journal of Chemical Physics*, 139(12), 2013.

[39] H. Stamati, C. Clementi, and L. Kavraki. Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Proteins: Structure, Function, and Bioinformatics*, 78(2):223–235, 2010.

[40] S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

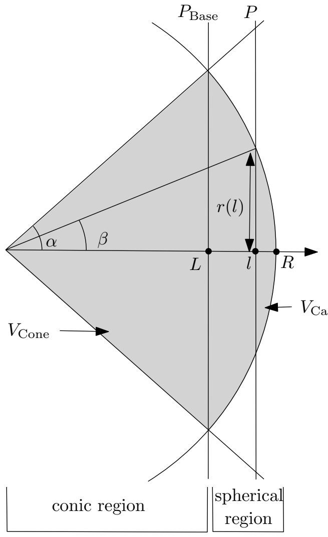
6 Appendix: uniform sampling in a hypercone

We wish to sample uniformly at random in the intersection of a cone of aperture α intersected with a n-dimensional ball $B^n(R)$ as described in Fig.15. The procedure therefore yields a random point in the cone, and the direction defined by this point and the apex of the cone. The algorithm and the calculations use the notations of Fig. 15.

6.1 Uniform direction in a cone: algorithm overview

The classical strategy to uniformly sample on the unit sphere consists of picking a random vector of independent and identically distributed (iid) normally distributed random variables, and to normalize the obtained vector. In a similar spirit, to sample uniformly the intersection between a cone and the unit sphere S^{n-1}, we sample uniformly the intersection between the cone and the unit ball bounded by S^{n-1}, and renormalize the result. In the following, we sketch the algorithm, and refer the reader to the supplemental section for full details (Fig. 15 and supplemental section 6).

Figure 15 Uniform sampling the intersection between the unit ball and a cone. The volume defined by the grey region is the union of a conic region and of a spherical region.



The algorithm proceeds in three steps:

- (i) Decide whether one samples from the conic or the spherical regions,
- (ii) Pick a slice in the cone or spherical cap,
- (iii) Sample the slice.

More formally, consider a $n-1$ dimensional ball of radius 1 and the cone of angle α. We define:

- $\text{Vol}_n(1)$: the volume of the unit ball in dimension n,
- I_x: the incomplete Beta factor – Def. in SI Section 6.2,

- the volumes of the conic and spherical regions respectively – Fig 15 and supplemental section 6:

$$V_{n,\alpha}^{\text{Cone}} = \frac{\text{Vol}_n(1)}{2} I_{\sin^2\alpha}(\frac{n+1}{2}, \frac{1}{2}) \tag{25}$$

$$V_{n,\alpha}^{\text{Cap}} = \frac{L^n}{n} \tan^{n-1}(\alpha). \tag{26}$$

Using these notations, the three aforementioned steps go as follows – details in section 6:

- (i) Draw $u \in [0, V_{n,\alpha}^{\text{Cone}}(1) + V_{n,\alpha}^{\text{Cap}}(1)]$. If $u < V_{n,\alpha}^{\text{Cone}}(1)$, we pick in the cone (i.e. $l < L$), else we pick in the cap (i.e. $l > L$).
- (ii) Pick a slice of the cone or the cap at distance l from the center, using the density

$$f_{cone}(l) = C_{cone} r(l)^{n-1} 1_{l \leq L} = C_{cone} \tan(\alpha)^{n-1} l^{n-1} 1_{l \leq L} \tag{27}$$

or

$$f_{cap}(l) = C_{cap} r(l)^{n-1} 1_{l > L} = C_{cap}(1 - l^2)^{\frac{n-1}{2}} 1_{l > L} \tag{28}$$

with C_{cone} and C_{cap} normalization constants used to define probability densities.

- (iii) Draw uniformly at random in the corresponding $n-1$ ball, using the density from Eq. (34).

6.2 Pre-requisites

Special functions. We shall need the Beta and incomplete Beta functions, defined by

$$\begin{cases} B(a,b) = \int_0^1 t^{x-1}(1-t)^{y-1} dt, \\ B(x;a,b) = \int_0^x t^{x-1}(1-t)^{y-1} dt (\text{ with } 0 < x < 1). \end{cases} \tag{29}$$

Using both, one defines the regularized incomplete Beta factor

$$I_x(a,b) = \frac{B(x;a,b)}{B(a,b)}. \tag{30}$$

Spheres and balls: surface and volume. The surface area of a sphere of a $n-1$ sphere $S^{n-1}(R)$ of radius R in $\mathbb{R}^d$

$$\text{Area}_{n-1}(R) = R^{n-1} \frac{2\ \pi^{n/2}}{\Gamma(d/2)} \equiv R^{n-1} A_n. \tag{31}$$

The volume of the corresponding ball $B^n(R)$ satisfies

$$\text{Vol}_n(R) = R \frac{\text{Area}_n(R)}{n} = R^n \frac{2}{n} \frac{\pi^{n/2}}{\Gamma(n/2)} = R^n \frac{\pi^{n/2}}{\Gamma(n/2+1)} \equiv R^n V_n. \tag{32}$$

To generate a point X uniformly at random on the unit the unit sphere S^n, we generate a point $X = (x_1, \dots, x_n)^t$ whose coordinates are iid Gaussian with $\mu = 0$ and $\sigma = 1$. The corresponding density is given by

$$f_G(X) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{2}}. \tag{33}$$

To obtain a unit vector, we normalize the latter as $\frac{X}{\|X\|}$. (NB: due to normalization the coordinates of this vector are not independent.)

Random generation within a ball. To generate a point uniformly at random inside $B^n(R = 1)$, observe that the volume of $B^n(r) = r^n V_n$. Differentiating yields

$$\frac{d}{dr}(r^n V_n) = dr^{n-1} V_n. \tag{34}$$

Therefore one generate a random value using the density dr^{n-1} for $r \in [0, 1]$.

Spherical caps of the n-dimensional ball. We consider a conic region inside the n-dimensional ball, consisting of the union of a pyramid and that of a spherical cap defined by the cone of aperture α (Fig. 15). Surface and volume of such a cap is easily computed [40].

To compute the volume of the cap, we integrate the volume of a $n-1$ dimensional sphere or radius $r \sin\beta$ whose height element is $d(r\cos\beta) = r\sin\beta$:

$$V_{n,\alpha}^{\text{Cap}}(r) = \int_0^\alpha \text{Vol}_{n-1}(r\sin\beta) d(r\cos\beta) = \frac{\text{Vol}_n(r)}{2} I_{\sin^2\alpha}(\frac{n+1}{2}, \frac{1}{2}). \tag{35}$$

Note that the incomplete Beta factor as the probability for a point of the ball to also be inside the spherical cap.

To compute the surface of the cap, we integrate the area of a $n-1$ dimensional sphere or radius $r\sin\beta$ with arc element $rd\beta$:

$$A_{n,\alpha}^{\text{Cap}}(r) = \int_0^\alpha \text{Area}_{n-1}(r\sin\beta) r d\beta = \frac{\text{Area}_{n-1}(r)}{2} I_{\sin^2\alpha}(\frac{n-1}{2}, \frac{1}{2}). \tag{36}$$

6.3 Algorithm to uniformly sample a hypercone

6.3.1 Sampling from f_{cap}

The previous algorithm requires sampling from f_{cap} defined in eq.(28). The most straightforward way to sample from a probability density is to compute the inverse of the cumulative distribution function ($F(x) = \int_{-\infty}^x f(y)dy$). This requires to compute a primitive of the density. However, there is no simple analytic expression for the primitive of f_{cap}. Hence, we fall back to rejection sampling with a well chosen base distribution such that the rejection rate do note depend on the dimension n.

Observe that while $(1-l^2)^{\frac{n-1}{2}}$ do not have a simple primitive, the function $l(1-l^2)^{\frac{n-1}{2}}$ do. Therefore we define

$$g_{cap}(l) = MC_{cap} l (1-l^2)^{\frac{n-1}{2}} \tag{37}$$

with M such that for all l, $g_{cap}(l) \geq f_{cap}(l)$ which is required for rejection sampling. The optimal choice for M is:

$$M = \frac{1}{L} = \frac{1}{cos\alpha}.$$

L $\tilde{g}_{cap}$ the renormalized version of g_{cap}. Assuming we can sample point from $\tilde{g}_{cap}$, the acceptance rate for each l in the rejection algorithm used with f_{cap} and g_{cap} is

$$\frac{f_{cap}(l)}{g_{cap}(l)} = \frac{1}{lM} \leq \frac{1}{M}$$

as $l \leq 1$. Hence the acceptance rate do not depend on n and only on α the opening of the cone.

Sampling from $\tilde{g}_{cap}$: To sample from $\tilde{g}_cap$ we compute the inverse of its cumulative distribution.

Let

$$B(x) = \int_L^x l(1-l^2)^{\frac{n-1}{2}} dl$$

using the change of variable $y = 1 - l^2$, we deduce:

$$\begin{aligned} B(x) &= \left[-\frac{(1-y^2)^{(1+n)/2}}{1+n} \right]_L^x \\ &= \frac{(1-L^2)^{(1+n)/2}}{1+n} - \frac{(1-x^2)^{(1+n)/2}}{1+n} \end{aligned}$$

The cumulative distribution for $\tilde{g}_{cap}$ is

$$\begin{aligned} F(x) &= 1_{x>L} \frac{B(x)}{B(1)} \\ &= 1_{x>L} \left(1 - \frac{(1-x^2)^{(1+n)/2}}{(1-L^2)^{(1+n)/2}} \right) \end{aligned}$$

And its inverse:

$$F^{-1}(x) = \sqrt{1 - (1-L^2)(1-x)^{2/(n+1)}}$$

Hence we can sample from $\tilde{g}_{cap}$.

6.3.2 Sampling from f_{cone}

The inverse CDF for f_{cone} is straightforward to compute:

$$F_{cone}^{-1}(x) = Lx^{1/n}$$

Therefore we can sample from f_{cone}.

6.4 Changing the cone axis

The previous section algorithm generates a point in a cone whose axis is fixed: $e_1 = (1, 0, ...0)$. In practice, the axis of a cone is aligned with the gradient of the potential energy – Section 3.4.

To handle arbitrary cones, we apply a linear transformation. We describe here how to apply this transformation with a contained complexity. Let $d \in \mathbb{R}^n \setminus \{e_1\}$ be the desired axis of the cone.

Let H be the hyperplane orthogonal to e_1. In the algorithm, we generates points in H. Suppose we generate $(x_2, ..., x_n)$ in H. For any orthonrmal basis $\epsilon_2, ..., \epsilon_n$ of H, the points $x_2\epsilon_2 + ... + x_n\epsilon_n$ will have the same distribution in H. Hence we try to find a basis $\epsilon_2, ..., \epsilon_n$ adapted to our problem.

We choose $\epsilon_2 = \frac{d - <d,e_1>e_1}{\|d - <d,e_1>e_1\|}$.

We complete this base with $\epsilon_3, ..., \epsilon_n$, and we will see that the choice of these $\epsilon_3, ..., \epsilon_n$ do not matter.

Let R the rotation such that $R(e_1) = d$ and $R(\epsilon_i) = \epsilon_i$ for $i > 2$.

Let $H_0 = Vect(e_1)$, $H_1 = Vect(e_1, \epsilon_2)$ and $H_2 = Vect(\epsilon_3, ..., \epsilon_n)$.
Let $x \in \mathbb{R}^n$. Then there exists u_1, u_2 and v such that

$$x = u_1 e_1 + u_2 \epsilon_2 + u_3 v$$

with $v = x- < x, e_1 > e_1 - < x, \epsilon_2 > \epsilon_2 \in H_2$. u_1, u_2 and v are straightforward to compute. We easily get:

$$R(x) = R(u_1 e_1 + u_2 \epsilon_2) + u_3 v$$

Thus the transformation R can be reduced to a simple rotate in $\mathbb{R}^2$. Let $\theta =< e_1, d >$. Then

$$R(u_1 e_1 + u_2 \epsilon_2) = u_1 d + u_2 \left(cos(\theta + \pi/2) e_1 + sin(\theta + \pi/2) \epsilon_2 \right)$$

Thus we full transform is as follow:

- compute
$$\epsilon_2 = \frac{d- < d, e_1 > e_1}{\|d- < d, e_1 > e_1\|}$$
- compute $u_1 =< x, e_1 >$, $u_2 =< x, \epsilon_2 >$ and $v = x - u_1 e_1 - u_2 \epsilon_2$
- compute $\theta = (e_1, x)$ and $\tilde{d} = (cos(\theta + \pi/2) e_1 + sin(\theta + \pi/2) \epsilon_2)$
- $R(x) = u_1 d + u_2 \tilde{d} + v$

7 Appendix: transition probability for darting

7.1 Notations

We give here a detailed computation of the transition probability for darting given by eq. 20. We use the same notations than in section 3.5. Let us write P_{dart} the Markov kernel associated to the darting move. Let x be a point of $\mathcal{E}$. The transition kernel has a density, hence we write $P(x, y)$ instead of $P(x, dy)$. For a minimum k, let $H(k)$ the Hessian of U at m_k. Let $\lambda_1, ..., \lambda_n$ its eigenvalues and $e_1, .., e_n$ its eigenvectors as an orthonormal basis. Finally let $A_k \subset \mathcal{E}$ be the basin of attraction of minimum k and let k_x the minimum such that $x \in A_{k_x}$. We consider the following rescaling of state space:

$$h_k(y) = m_k + \sum_i \sqrt{\lambda_i}(y - m_k|e_i)e_i. \tag{38}$$

Let $\tilde{U}_k(z) = U(h_k^{-1}(z))$ the potential energy in the rescaled space. Let $\tilde{f}_k(u, T_U)$ the application which associates the first intersection between $m_k + \alpha u$ and $\tilde{U} = T_U + U(m_k)$ with $\alpha > 0$. Formally, $\tilde{f}_k$ is an application defined on $S^{n-1} \times \mathbb{R}^+$.

Let $f_k(u, T_U) = h^{-1}(\tilde{f}(u, T_U))$. Also let $f_{k*}(\mu_{k,x})$ be the pushforward measure of $\mu_{k,x}$ by f_k.

Then, the Markov kernel seen as an operator on measures is given by:

$$P_{dart}(x, .) = \frac{1}{K} \sum_k \frac{f_{k*}(\mu_{k,x})}{\int \mu_{k,x}} \tag{39}$$

where $\mu_{k,x}$ is the product measure of the Lebesgue measure on S^{n-1} and the Lebesgue measure of

$$I_k(x) = [U(x) - U(k_x) + U(k) - \beta, U(x) - U(k_x) + U(k) + \beta] \tag{40}$$

7.2 Assumptions

We first define assumptions ensuring that function f_k defines a bijection between the set of directions and the restriction of the target energy level set surface to the basin of a local minimum. Intuitively, the level set surface seen from a local minimum should be star shaped, and the energy range considered should not contain any critical value associated with a saddle point. More formally:

Assumption 1. *For every local minimum $k \leq K$, $u \in S^{n-1}$, $T_U \in [U(m_k), U(m_k) + M]$, the intersection $\{y|y = m_k + \alpha u, \alpha > 0\} \cap \{y|U(y) = T_U\} \cap A_k$ is a single point. See Fig. 16 and Fig. 17.*

Doing a line search for every minimum is expensive. Using constant M defined in Section 3.5(see paragraph *When to jump*), we introduce the following assumption to simplify Eq. (39), which essentially states that the points considered belong to the Voronoi region of the local minimum m_k:

Assumption 2. *For every y such that $U(y) - U(m_{k_y}) \leq M$, then for every $k \leq K$,*

$$\|y - m_{k_y}\| \leq \|y - m_k\|$$

The simplified expression for Eq. (39) reads as

$$P(x,y) = \frac{1}{K}\frac{f_{k_y*}(\mu_{k_y,x})}{\int \mu_{k_y,x}}$$

where k_y is the closest minimum to y. As a final observation, assumptions 1 and 2 are true if M is small enough (using a second order Taylor expansion for the proof at the bottom of the local minima)

Figure 16 Not allowed by assumption 1 as there are multiple intersection point between a direction and the restriction of an energy level set to a basin.

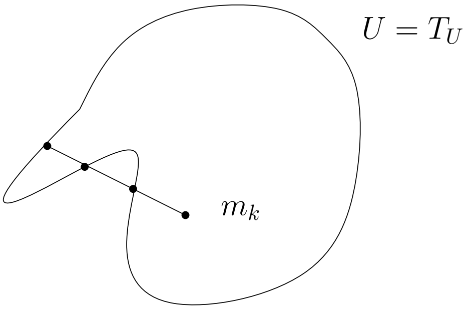
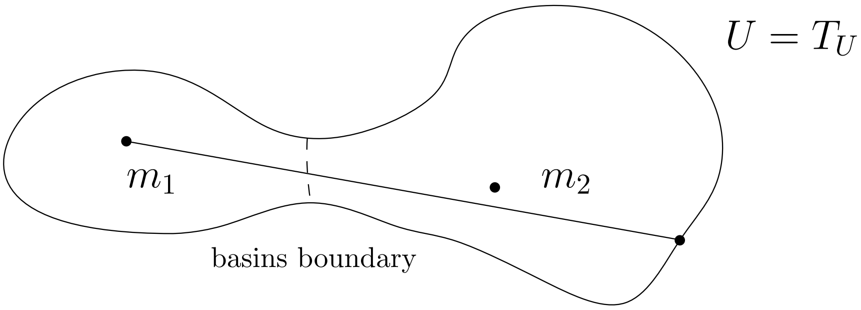


Figure 17 Not allowed by assumption 1 as selected directions yield intersection points outside the basin of m_1.



7.3 Derivation of the transition probability

Under assumption 1, f_k is a bijection from $S^{n-1} \times [U(m_k), U(m_k) + M]$ to the connected component containing m_k of the set of point $\{y|U(y) \leq U(m_k) + M\}$. Hence its inverse is well defined. The density of the pushfoward measure can be computed using the usual change of variable formula:

$$f_{k*}(\mu_{k,x})(y) = |J(f_k^{-1})(y)| 1_{I_k}(U(y))$$

For notation simplicity, we consider a fixed k and write $f = f_k$ and $h = h_k$ for the following computation. The inverse of f has the following expression:

$$\tilde{f}^{-1}(z) = \left(\frac{z - m_k}{\|z - m_k\|}, \tilde{U}(z) \right)$$

Let $z = h(y)$ and $u = \frac{z-m_k}{\|z-m_k\|}$, and choose $w_1, ..., w_{n-1}$ in $\mathbb{R}^n$ such that $w_1, ..., w_{n-1}, u$ is an orthonormal basis of $\mathbb{R}^n$. Let $l = \|z - m_k\|$. Then:

$$\frac{\partial \tilde{f}^{-1}}{\partial w_i}(z) = \left(\frac{1}{l} w_i, \frac{\partial \tilde{U}}{\partial w_i}(y)\right)$$

Observe that $w_1, ..., w_{n-1}$ is an orthonormal basis of the tangent space of S^{n-1} at u. Then considering that $\tilde{f}^{-1}$ is an application from an open set of $\mathbb{R}^n$ to $S^{n-1} \times \mathbb{R}^+$, the Jacobian of f^{-1} becomes:

$$J(\tilde{f}^{-1})(z) = \begin{pmatrix} \frac{1}{l} & 0 & ... & 0 & 0 \\ 0 & \frac{1}{l} & ... & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & ... & \frac{1}{l} & 0 \\ (\nabla \tilde{U}(z)|w_1) & (\nabla \tilde{U}(z)|w_2) & ... & (\nabla \tilde{U}(z)|w_{n-1}) & (\nabla \tilde{U}(z)|u) \end{pmatrix}$$

Hence

$$|J(\tilde{f}^{-1})(z)| = \frac{1}{l^{n-1}}(\nabla \tilde{U}(z)|u)$$

And using $\tilde{U}(z) = U(h^{-1}(z))$,

$$\frac{\partial \tilde{U}}{\partial u}(z) = \nabla U(y)^T J(h^{-1})(z)u \tag{41}$$

$$= \nabla U(y)^T J(h^{-1})(z)\frac{z - m_k}{l} \tag{42}$$

$$= \nabla U(y)^T J(h^{-1})(z)(h(y) - m_k)\frac{1}{l} \tag{43}$$

$$= \nabla U(y)^T (y - m_k)\frac{1}{l} \tag{44}$$

Where the simplification in equation 44 is justified by the fact that $h(y) - m_k = J(h)(y - m_k) = J(h^{-1})^{-1}(y - m_k)$. Combining the two previous equations:

$$|J(\tilde{f}^{-1})(z)| = \frac{1}{l^n}\nabla U(y)^T (y - m_k)$$

We deduce:

$$|J(f^{-1})(y)| = |J(h)|\frac{1}{l^n}\nabla U(y)^T (y - m_k)$$

The Jacobian matrix of h is easy to compute:

$$|J(h)| = \prod_{i \leq n} \sqrt{\lambda_i}$$

Hence we deduce:

$$f_{k*}(\mu_{k,x})(y) = 1_{I_k}(U(y))\frac{1}{l^n}\nabla U(y)^T (y - m_k)\prod_{i \leq n} \sqrt{\lambda_i} \tag{45}$$

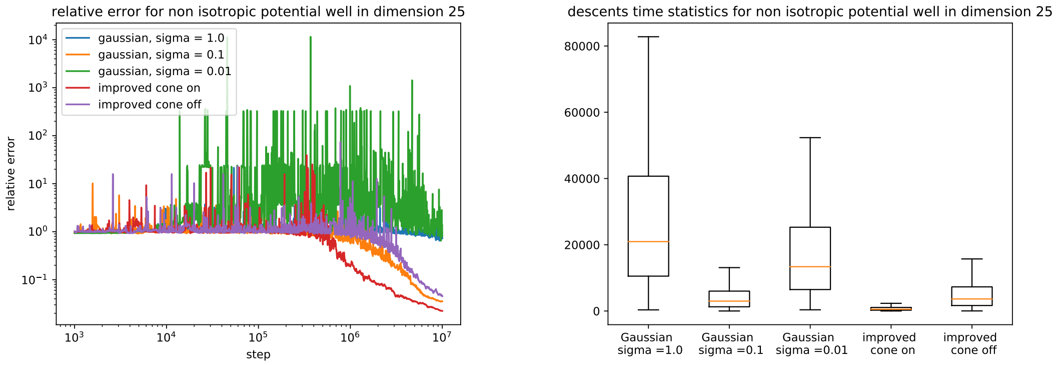
The rescaling factor for measure $\mu_{k,x}$ is:

$$\int \mu_{k,x} = \frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right) 2\beta} \tag{46}$$

Injecting equations 45 and 46 into equation 39 allows us to compute $P_{dart}(x, y)$.

,

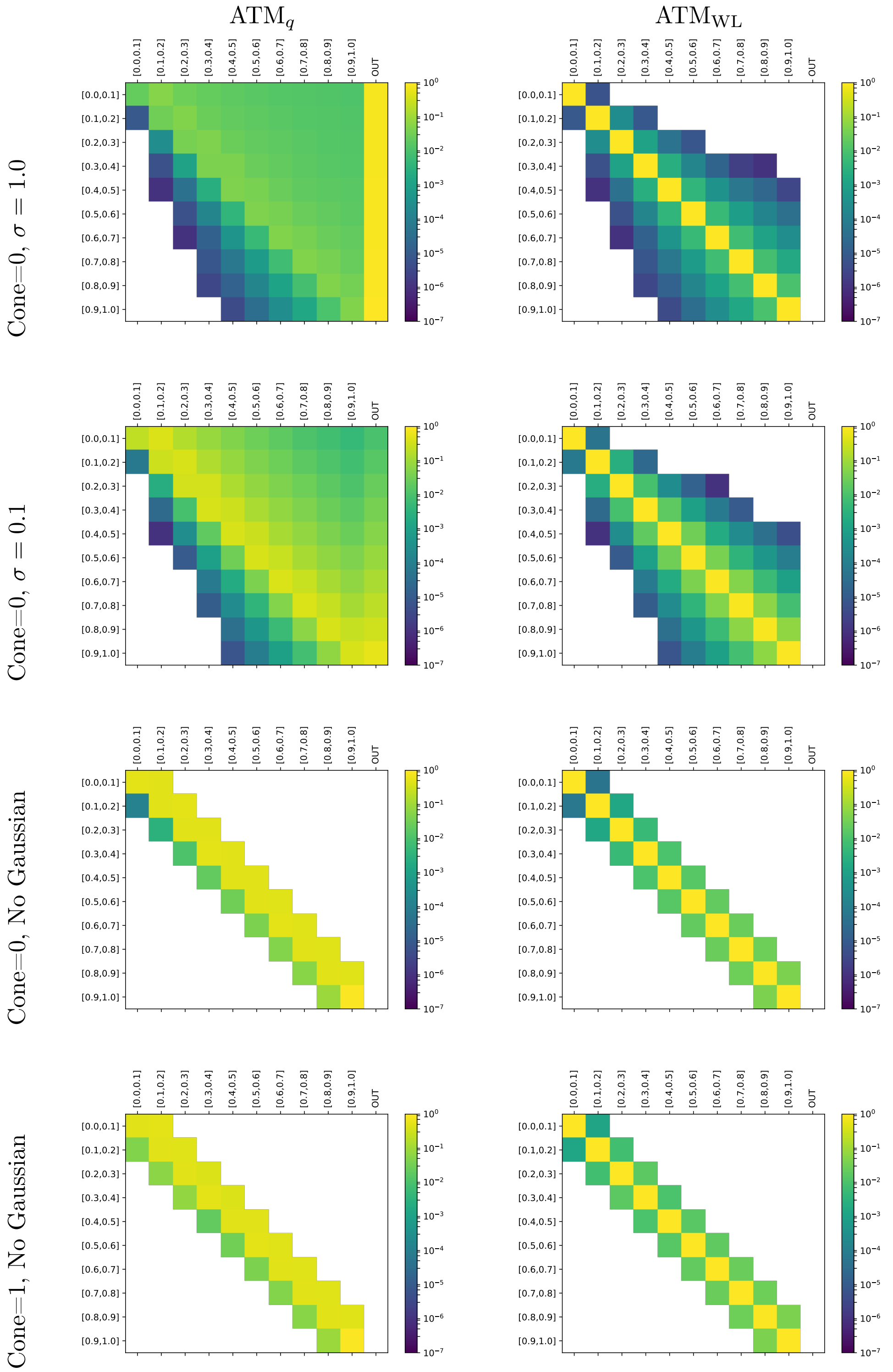
Figure 18 Non isotropic single well in dimension 25: comparison of the five proposals. Values have been averaged over 30 runs. The five proposals used are the three Gaussian based proposals, plus our combined proposal with and without the cone improvement. **(Left) Comparison of the evolution of relative error – Eq. (22) (Right) Box plot of the descending times.**



8 Appendix: results

8.0.1 Single well potential: non isotropic

Figure 19 Anisotropic single well, $n = 25$, : aggregated transition matrices.



Contents

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399