On Sequential Monte Carlo (SMC) strategies for Target Distributions

M. Rousset [1,2]
F. Cérou, A. Guyader, B. Delyon, T. Lelièvre, G. Stoltz, C.E. Bréhier, L. Goudenège, P. Héas. (PhDs: F. Ernoult, K. Tit).

[1]Inria Rennes Bretagne Atlantique

[2]IRMAR, Université de Rennes 1

RESIM 2021

Aim of the talk

- **'Target probability distribution'**: defined as a density w.r.t to a easily simulable distribution, density given up to a normalizing constant. E.g.: posterior distribution, Gibbs probability.
- **SMC = particle methods= Importance splitting**. As opposed to MCMC methods. Start with a sample of N 'particles'. Algorithms output: sample of N particles (approx. indep.) with distribution the 'target'.
- **Aim of the talk:** How to think about adaptivity to speed up sims. Nota Bene: Casual chat, not in papers !

E.g.: Rare event problem

- $\pi(dx)$ a reference probability on $S\,(=\mathbb{R}^d)$ that can be exactly simulated (e.g. Gaussian, uniform).
- $\mathrm{score} : \mathbb{R}^d \to \mathbb{R}$ a given computable function.
- Assume $\pi(\{\mathrm{score} > 0\}) = 1$. Problem: for $\underline{s = 1}$:

$$\begin{cases} \text{Estimate} \quad p_s := \pi(\{\mathrm{score} > s\}) \ll 1 \\ \text{Simulate according to 'target'} \quad \eta_s(dx) := \pi(dx | \,\mathrm{score}(x) > s). \end{cases}$$

Idea

Estimate/Simulate ' **"smoothly" and sequentially'** *the path*

$$s \mapsto (p_s, \eta_s), \quad s \in [0, 1].$$

Generalization

- $\frac{1}{z_0}\mathrm{e}^{-V_0(0)}\pi(dx)$ a reference probability on $S = \mathbb{R}^d$ that can be exactly simulated (e.g. Gaussian, uniform). Choose $z_0 = 1$.
- $(s, x) \mapsto V_s(x) : \mathbb{R} \times \mathbb{R}^d \times \to \mathbb{R}$ a given computable function (called potential). (Optional: $\nabla_x V_s(x)$ is available).
- Problem, for $s := 1$:

$$\begin{cases} \text{Estimate the normalization:} & z_s := \pi(\mathrm{e}^{-V_s()}) \\ \text{Simulate according to 'target':} & \eta_s(dx) := \frac{1}{z_s}\mathrm{e}^{-V_s(x)}\pi(dx). \end{cases}$$

- Previous rare event model is particular case for:

$$V_s(x) = \begin{cases} +\infty & \text{if } \mathrm{score}(x) \leqslant s \\ 0 & \text{if } \mathrm{score}(x) > s \end{cases}$$

Manifold Generalization[1]

- $\frac{1}{z_0}\mathrm{e}^{-V(x,0)}\pi_0(dx)$ a target probability on $S = \mathbb{R}^d$ that can be exactly simulated (e.g. Gaussian, uniform). $z_0 = 1$.
- Target : $\mathrm{e}^{-V_s} d\pi_s / z_s$.
- $s \mapsto \pi_s$ a path of mutually singular non-negative reference measures and a family of computable maps $i_{s,s'} : \mathbb{R}^d \to \mathbb{R}^d$ with $s, s' \in \mathbb{R}$ such that:

$$\pi_{s'} = i_{s,s'}[\pi_s] \quad \text{(push-forward)}$$

Example

$\pi_s := 2d' < 2d$-dimensional phase-space volume of a parametric family of co-tangent spaces $s \mapsto T^*\Sigma_s \subset \mathbb{R}^{2d}$. $i_{s,s'}$ is a simulable symplectic projection.

[1] Lelièvre-Stoltz-Rousset, *Langevin dynamics with constraints and computation of free energy differences*, 2012

High Dimensional Applications

- Sampling w.r.t. Gibbs distribution. Tempering: $\pi_s \propto \mathrm{e}^{-sU(x)}\pi(dx)$.
- Bayesian statistics: $\pi =$ prior distribution on model(s). $-V(s,x) =$ (smoothed) log-likelihood from $s \times n_{\mathrm{obs}}$ datas.
- $\pi =$ physical Markovian trajectory (Thermostatted Molecular Dynamics). Score = 'minimum distance' of path from a molecular configuration.

Sequential Monte-Carlo a.k.a. Importance Splitting

Define: $0 = s_{(0)} < \ldots < s_{(i_{\max})} = 1$ a given, finite ladder of scores.

$X^n_{s_{(i)}}$ state of particle n at iteration i.

General Form of the Algorithm with Weighted Particles:

(0) Simulate N independent particles according to $\eta_0 = \frac{1}{z_0}\mathrm{e}^{-V_0}\pi$.

Iterate on $i = 1 \ldots i_{\max}$:

(i) **Weights:** update the 'importance weight' of each particle $n \in (1, N)$ by $\mathrm{e}^{-V_{s_{(i)}}(X^n_{s_{(i-1)}})+V_{s_{(i-1)}}(X^n_{s_{(i-1)}})}$ (target: $\mathrm{e}^{-V_{s_{(i)}}}\pi$).

(i) **Selection (optional)** kill and/or split particles and update weights. E.g.: triggered if weights are too degenerate.

(i) **Mutation**: modify ('mutate') (all or some or none) particles with Markov Chain Monte Carlo transition $M_{s_{(i)}}(x, dx')$ that leaves invariant the target $\eta_{s_{(i)}}(dx) := \frac{1}{z_{s_{(i)}}}\mathrm{e}^{-V(x,s_{(i)})}\pi(dx)$.

Sequential Monte-Carlo a.k.a. Importance Splitting

Estimators:

- Target measures $\eta_s = \frac{1}{z_s} \mathrm{e}^{-V(x,s)} \pi(dx)$ are estimated by weighted empirical measures with normalization

$$\eta^N_{s_{(i)}} := \sum_{n=1}^{N} \mathrm{Weight}^n_{s_{(i)}} \delta_{X^n_{s_{(i)}}} / \sum_{n=1}^{N} \mathrm{Weight}^n_{s_{(i)}}.$$

- Normalizations are estimated by the average weights over particles

$$z^N_{s_{(i)}} := \frac{1}{N} \sum_{n=1}^{N} \mathrm{Weight}^n_{s_{(i)}}$$

Fun Remark: Includes MCMC !

- Pick a ladder where all scores (except first) $\to 1$.
- NO selection, ONLY Mutations.
- GET: N MCMC with η_0 prior initial condition.

Basic Refs

Papers:

- Del Moral Doucet Jasra *Sequential Monte Carlo samplers* 2006.
- A Beskos, A Jasra, N Kantas, A Thiery *On the convergence of adaptive sequential Monte Carlo methods* 2016
- F Cérou, P Del Moral, T Furon, A Guyader *Sequential Monte Carlo for rare event estimation* 2012
- F Cérou, A Guyader, *Adaptive Multilevel Splitting for rare event analysis*, 2007.
- In Phys.: 'Jarzynski equality'
- Freddy Bouchet and al..

Books

- Liu *Monte Carlo Strategies*
- Chopin *Introduction To Sequential Monte Carlo*
- Doucet, Freitas, Gordon *Sequential Monte Carlo in Practice*
- Del Moral *Feynman-Kac formula*

Classification of re-sampling or selection scheme

Definition

A selection or re-sampling scheme draw branching numbers $B_n \in \mathbb{N}$, $n = 1 \dots N$ *such that:*

$$\widetilde{\text{weight}}\, \mathbb{E}[\sum_{n=1}^{\tilde{N}} \delta_{\tilde{X}^n}] = \widetilde{\text{weight}}\, \mathbb{E}[\sum_{n=1}^{N} B^n \delta_{X^n}] = \sum_{n=1}^{N} \text{weight}^n \delta_{X^n}.$$

The branching numbers define a new particle system $\tilde{X}_1, \dots, \tilde{X}_{\tilde{N}}$ *with* $\tilde{N} = \sum_n B_n$ *particles and common weight* $\widetilde{\text{weight}}$.

- $B^n \geqslant 1$ *: selection of* **splitting type**.
- $B^n \leqslant 1$ *: selection of* **killing type**.
- $B^n \geqslant 1$ *and* $\mathbb{E}(B_n)$ *is independent on* n*:* **neutral bearing**.

What is 'adaptivity' ?

- A 'non-adaptive' SMC/Importance Splitting algorithm consist of: i) **preset** ladder of scores $0 = s_{(0)} < \ldots < s_{(i_{\max})} = 1$, ii) **preset** choice of mutations M_s leaving targte η_s invariant.
- Many 'adaptive' variants (e.g. Adaptive Multilevel Splitting, see after) are presented as follows: the choice of the scores is **random, adaptive**.
- In this talk I propose the 'mindset':

Idea

Interpret **'Adaptive scores'** *as* $\rightarrow$ **'Triggered and/or adaptive mutations'**.

'Adaptive scores' = nothing happens for many scores because of adaptivity of the triggering of mutations.

Adaptive and Triggered Mutations

Consider the mutation M_s after the selection step in the algo.
Vocabulary:

- **Preset Mutations**: M_s is preset, applied to all particles at each score $\rightarrow$ non-adaptive, 'Feynman-Kac-Del Moral structure'.
- **Adaptive Mutations**: The mutation kernel M_s is random and depends on the past particle empirical distribution. E.g.: if M_s is based on accept/reject, proposal is adaptively tuned to target an average acceptance rate $r_0 \in (0, 1)$.
- **(Triggered) Mutations-If-Selection**: A mutation kernel M_s is applied only when selection step is triggered.
- **(Triggered) Mutations-On-Child**: A mutation kernel M_s applied only to children when a neutral bearing selection is triggered.

Adaptive and Triggered Mutations

Example (Mutations-If-Selection)

- *Compute the relative variance (Effective Sample Size) of weights at each score/iteration.*
- *If relative variance greater than a treshhold: trigger selection.*
- *If selection has been triggered, mutations on all particles are triggered.*

Example (Mutations-On-Child)

- *Special case of Mutations-If-Selection.*
- *Resampling/selection is split in two parts: i) re-sample/select according to the weights BUT so that final sample size $N - K < N$. ii) K new particles are added by independent picking of particles (neutral bearing).*
- *Triggered mutations are applied on the K children in ii) ONLY.*

Adaptive/Triggered Mutation variant

Remarks

- *Triggered Mutations is a kind of adaptivity.*
- *The goal of Triggered Mutations (If-Selection, On-Child) is to save computational power by avoiding mutations (hence evaluation of V or ∇V) if simple weighting is sufficient.*
- *Consistency of Adaptive mutations: large sample $N \to +\infty$.*
- *Well-known rare event case: Adaptive Multilevel Splitting (AMS) algorithm (see after).*
- *AMS in the dynamical setting has a hidden non-adaptive Feynman-Kac-Del Moral structure (see below).*

The Feynman-Kac-Del Moral structure

- For non-adaptive = preset mutations, the algorithm can be derived from a Feynman-Kac formula:

$$\int \varphi(x) \mathrm{e}^{-V_{s_{(i)}}(x)} \pi(dx) =$$

$$\mathbb{E}\left[\varphi(X_{s_{(i)}}) \mathrm{e}^{-\sum_{i'=1}^{i} V_{s_{(i')}}(X_{s_{(i'-1)}}) - V_{s_{(i'-1)}}(X_{s_{(i'-1)}})}\right]$$

  where $X_{s_{(i)}}$, $i \geqslant 0$ is a Markov chain with $X_0 \sim \eta_0$ and probability transition $M_{s_{(i)}}$.

- The algoritm is then: simulating independently N chains with weights. Additional re-sampling/selection to prevent weight degeneracy.
- Nota Bene: in Del Moral, re-sampling/selection is put in a (very slightly restrictive) 'mean-field' form.

Jarzynski equality

Remark

The Feynman-Kac formula before is known in physics as 'Jarzynski equality'. In that case:

- *s is reaction coordinate or a thermodynamic parameter.*
- *Target is a canonical Gibbs distribution (mechanical system thermostatted).*
- *Mutation is Newton dynamics with parameter s + random perturbation at given temperature (Langevin).*
- *Weight* $= e^{-Work/(k_b T)}$ *!!*
- *Exists experimentally !!*

The Feynman-Kac-Del Moral structure

Proposition (Unbiasedness)

Un-normalized estimators are unbiased for algorithms following the Feynman-Kac-Del Moral structure.

Proof.

First remark that $\int \varphi \, \mathrm{e}^{-V_{s_{(i)}}} d\pi = \mathbb{E}[\varphi(X_{s_{(i)}}) \mathrm{e}^{-V_{s_{(i)}}(X_{s_{(i-1)}}) + V_{s_{(i-1)}}(X_{s_{(i-1)}})} \times \ldots \times \mathrm{e}^{-V_{s_{(1)}}(X_{s_{(0)}}) + V_{s_{(0)}}(X_{s_{(0)}})}] =: \mathbb{E}[Q^{0 \to i}(\varphi)(X_0)]$ where $i \mapsto X_{(i)}$ is the MCMC chain used in the mutation step. Then check that for $i \leqslant i_0$

$$i \mapsto z^N_{s(i)} \int Q^{i \to i^0}(\varphi) \, d\eta^N_{s(i)} \text{ is a martingale.}$$

□

Consistency[2] when $N \to +\infty$

Proposition (Asymptotic Unbiasedness)

Consider any algorithm with adptive features continuous w.r.t involved estimators. In the large sample size limit $N \to +\infty$, for each i,

$$(z^N_{s_{(i)}}, \eta^N_{s_{(i)}}) \xrightarrow[N \to +\infty]{\mathbb{P}} (z_{s_{(i)}}, \eta_{s_{(i)}})$$

Proof–(has to be made generically).

By induction $i \to i+1$. □

[2] A Beskos, A Jasra, N Kantas, A Thiery *On the convergence of adaptive sequential Monte Carlo methods* 2016

High dimension requires scarse mutations

- High Dimension $d \gg 1$: weights that are $\times$ by $\mathrm{e}^{-V_{s_{(i+1)}}(X_{s_{(i)}})+V_{s_{(i)}}(X_{s_{(i)}})}$ at each iteration have exponential variance with d (typically).

Example

In $\mathbb{R}^d$, if coordinates of X are i.i.d. and V has a sum form over coordinates and is smooth w.r.t. s, by CLT, non-degeneracy of weights requires:

$$s^{(i+1)} - s^{(i)} \sim \frac{1}{\sqrt{d}} \xrightarrow{d\to+\infty} 0.$$

- Tempting to not mutate at each $s^{(i)}$.
- Idea: switch to a continuum of scores:

$$s \in \left\{s^{(0)}, \ldots, s^{(I)}\right\} \quad \text{becomes} \quad s \in [0,1].$$

Indexing the algorithm by selection events

'Same' algorithm, new representation:

- **Non-Triggered Mutations**: Each particles evolve independently according to a Markov process with generator L_s invariant with respect to target $\eta_s \propto \mathrm{e}^{-V_s}\pi$.

Example

Piecewise constant Markov jump process

$$L_s(\varphi)(x) = \lambda_s(M_s(\varphi)(x) - \varphi(x)), \quad \eta_s M_s = \eta_s$$

can be simulated: i) mutations occur at random score (higher than s_0 with proba $\mathrm{e}^{-\int_0^{s_0}\lambda_s ds}$), ii) mutations with M_s.

- Other examples: discretization of a Stochastic Differential Equation, or Piecewise Deterministic Markov Process.

Re-Indexing the algorithm by splitting events

Initialize particles and set $S_{(0)} = 0$. Mutate all particles with L_s on $s \in [0, 1]$. **Iterate on** j:

(j) **Weights:** compute the 'importance' weight for $s \in [0, 1]$ of particles so that it targets η_s for each s, e.g.: $\mathrm{e}^{-\int_0^s \partial_{s'} V_{s'}(X_{s'}) ds'}$.

(j) **Selection** Compute the next random score

$$S_{(j)} := \inf \left\{ s \geqslant S_{(j-1)} | \text{Criteria}_s^N == 1 \right\}$$

e.g.: Criteria_s = weight degeneracy (Effective Sample Size) at s.

Then perform selection/re-sampling according to weights.

(j) **Triggered Mutations**: additional Mutations-If-Selection with $\tilde{M}_{S_{(j)}}$ (option: On-Child, Adaptive).

(j) **Preset Mutations**: mutate with L_s on $s \in [S^{(j)}, 1]$ new ($\Leftrightarrow$ all !) particles.

(Exit) Stop if $S^{(j)} = 1$ else $j \to j + 1$.

Re-Indexing the algorithm by splitting events

Remarks

- *Preset mutations are simulated by ANTICIPATION (can be adjusted to decrease cost).*
- *Mutations with L_s can be adaptive BUT adaptivity must NOT depend on ANTICIPATION.*
- *Unbiasedness/Feynman-Kac/Del Moral structure[a] holds if i) L_s non-adaptive, ii) no Triggered-Mutation .*
- *AMS in 'static setting' is an example with ONLY Triggered Mutations-On-Child (see after).*
- *AMS in 'dynamic setting' is an example with PSEUDO-triggered Mutation-On-Child: they are in fact anticipated preset mutations, (see after).*

[a] See also Brehier Gazeau Goudenege Lelievre Rousset GAMS 2016

Static[3] AMS algorithm

Let $k < N$ given. Assume rare event setting with:

- $\pi :=$ anything simulable.
- $\mathrm{e}^{-V_s} = \mathbf{1}_{\text{score}>s}$.
- $L_s = 0$, only triggered mutations.
- Selection $=$ killing $+$ neutral bearing. Triggered by k particles with lowest score which are killed and then neutrally borne.
- Mutation-If-Selection with Mutation-On-Child. $\tilde{M}_s$ is a MCMC kernel reversible w.r.t. π with rejection if proposal has score $\leqslant s$.

[3] F Cérou, P Del Moral, T Furon, A Guyader *Sequential Monte Carlo for rare event estimation* 2012

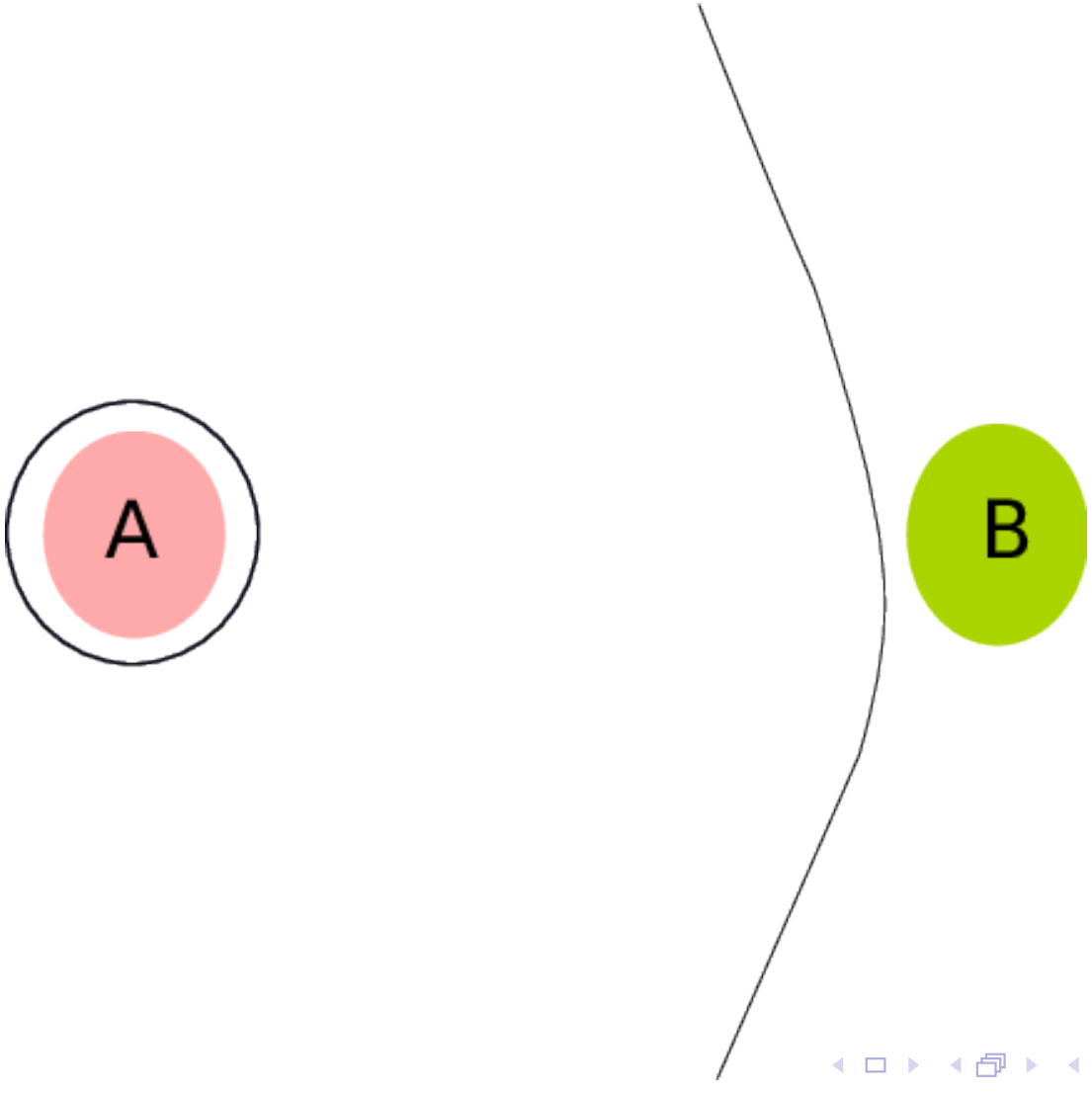
Dynamical[4] AMS algorithm

- π = Law of a Markov chain / process.
- $\mathrm{e}^{-V_s} = \mathbf{1}_{\text{score}>s}$, score = $\max(\xi(\text{path}))$.
- L_s = generator of π **starting from first hitting time of** $\{\xi > s\}$. N.B.: do nothing if score not attained.
- Selection = killing + neutral bearing. Triggered by k particle killed.
- Preset mutation of all particles with L_s. Mutations of old particles already simulated by ANTICIPATION.

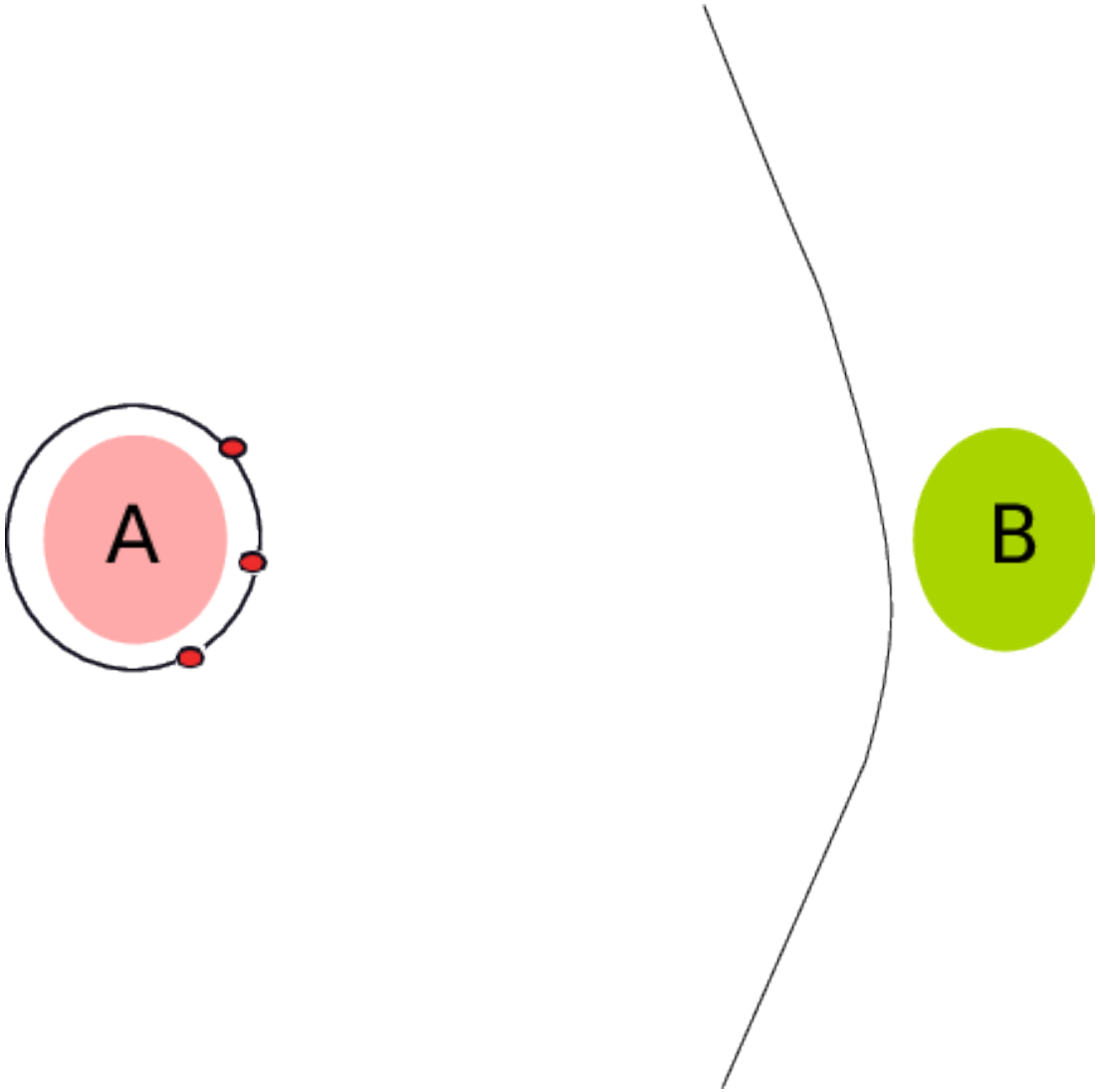
[4] F Cérou, A Guyader, *Adaptive multilevel splitting for rare event analysis*

Adaptive Multilevel Splitting

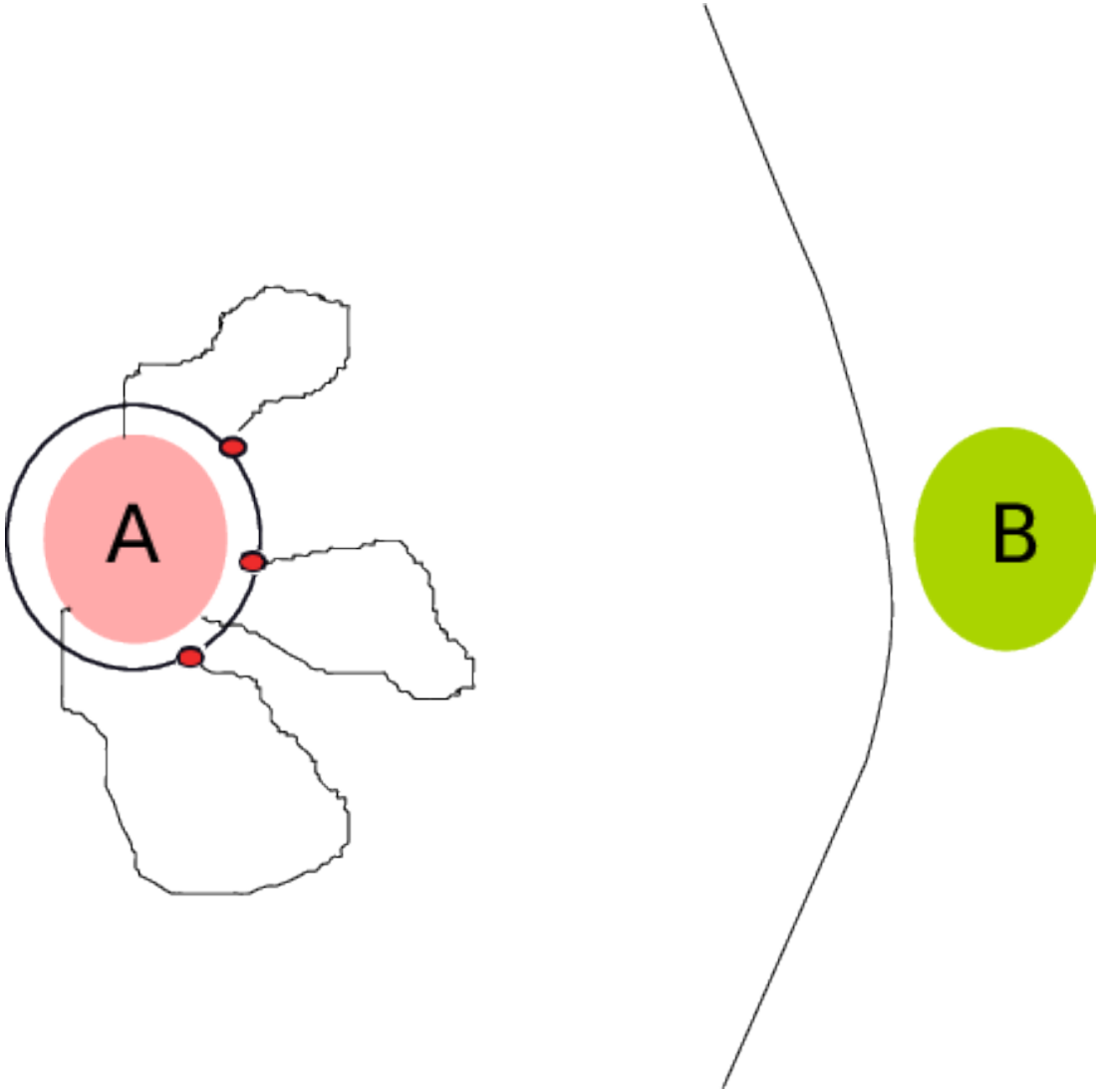
– Black line: $\{\xi = constant\}$.



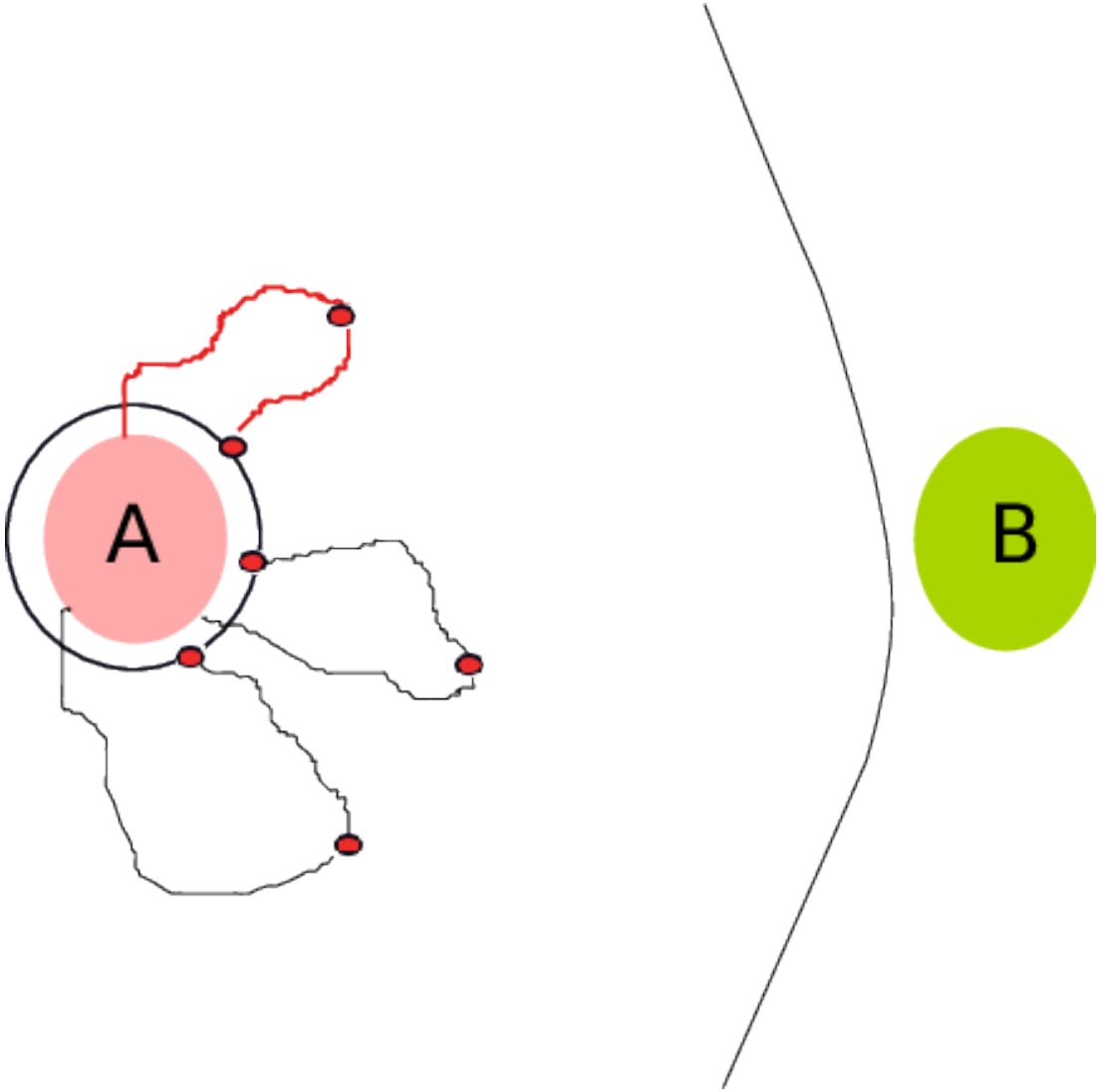
Adaptive Multilevel Splitting



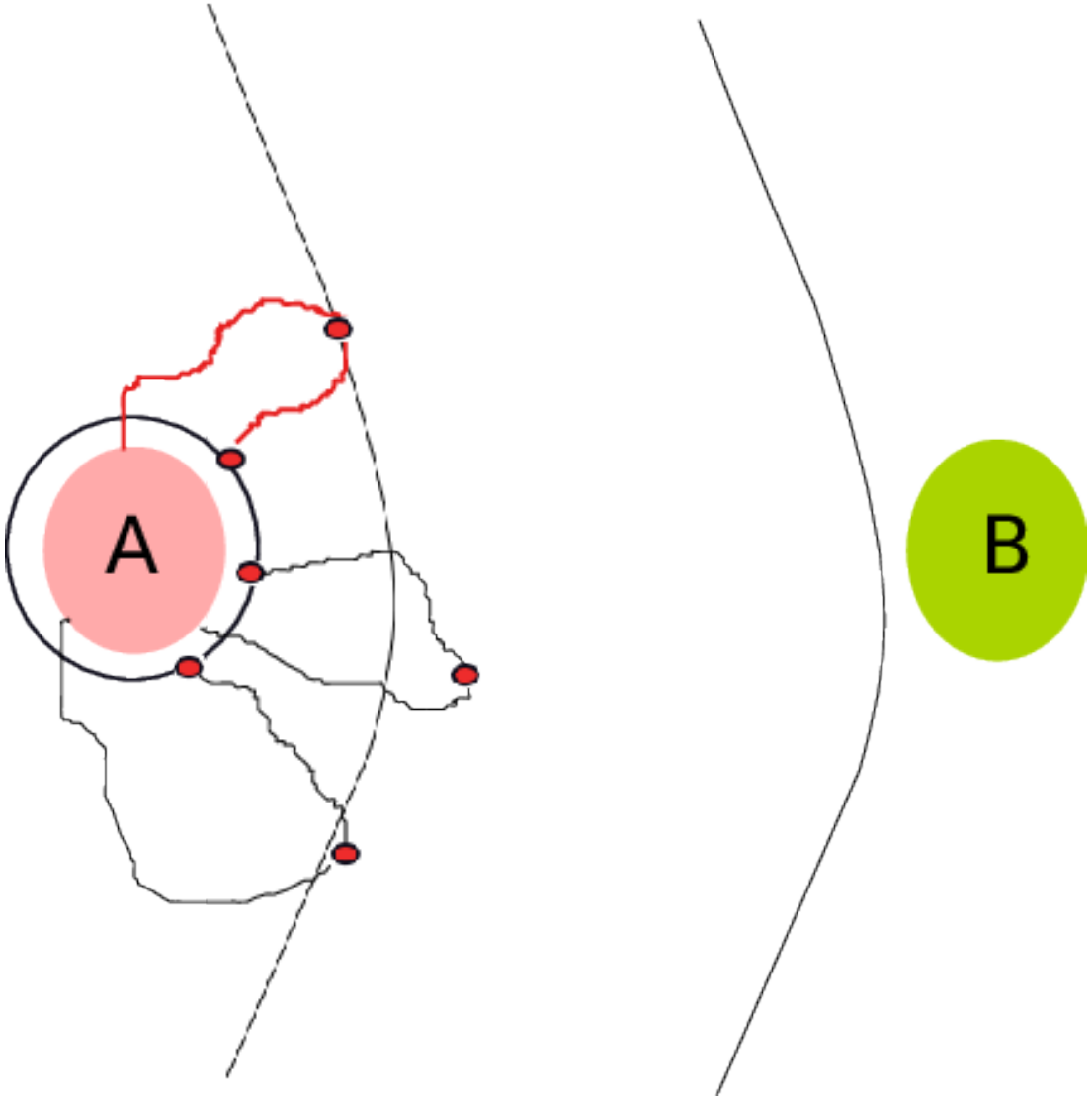
Adaptive Multilevel Splitting



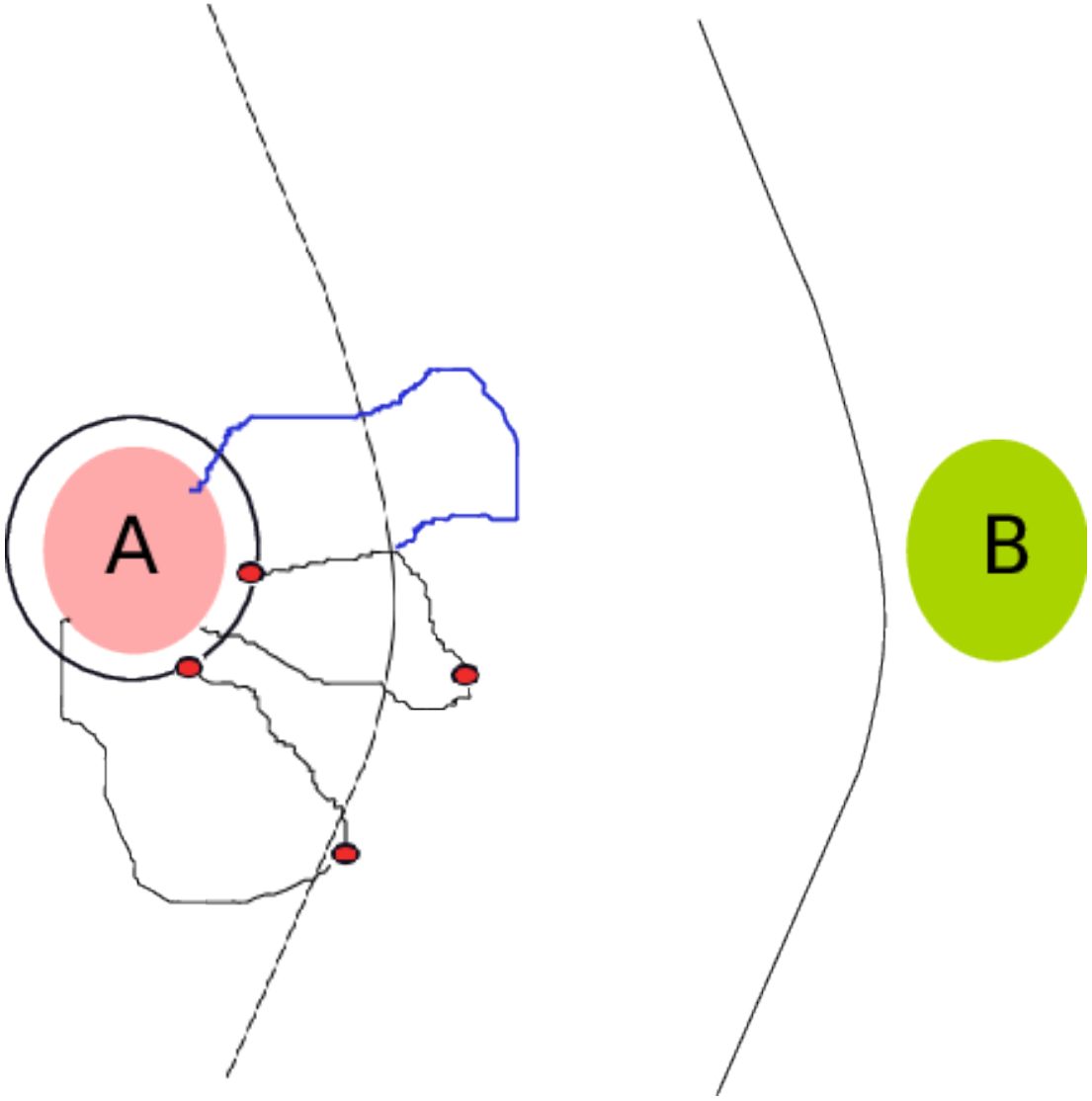
Adaptive Multilevel Splitting



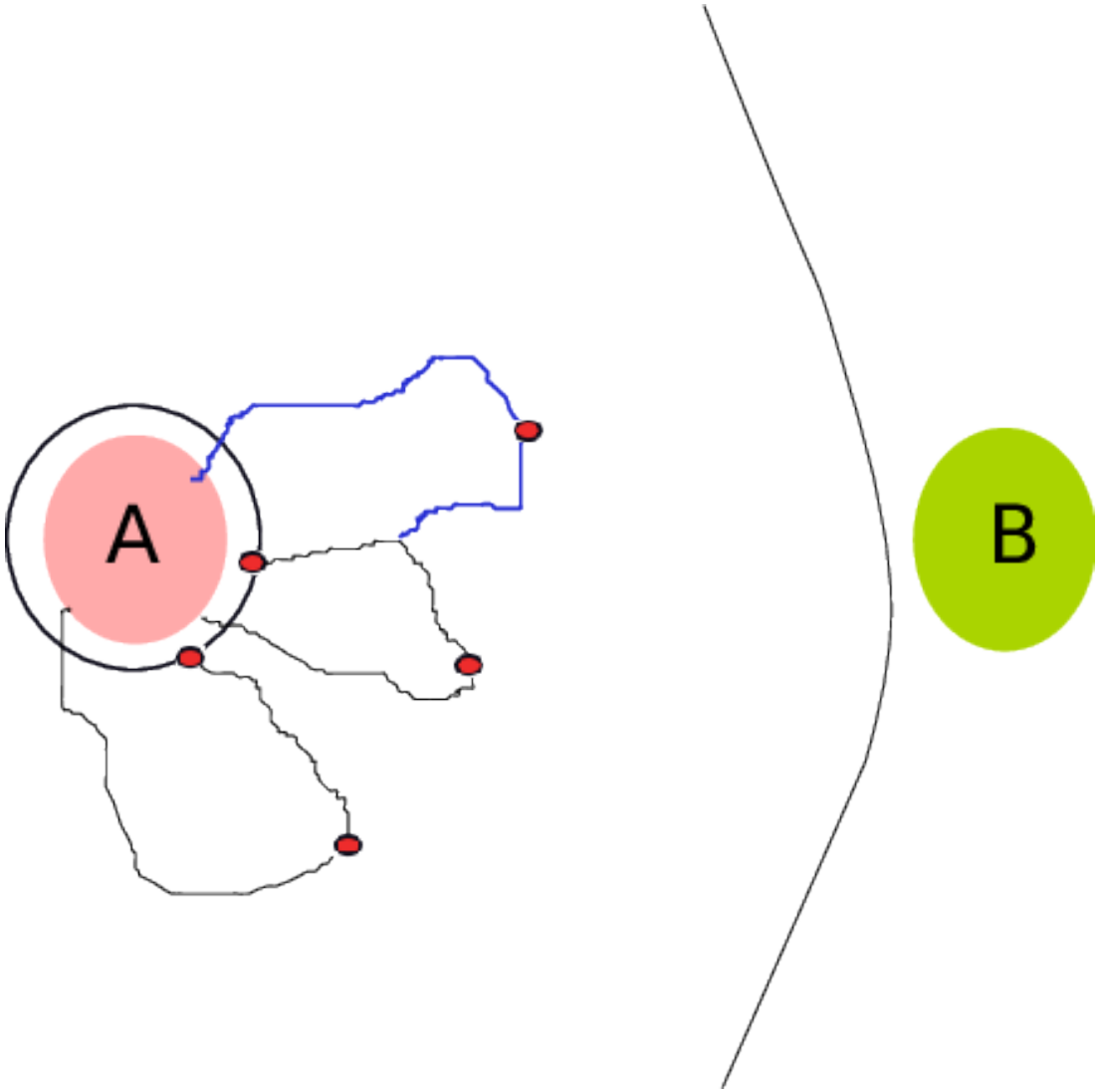
Adaptive Multilevel Splitting



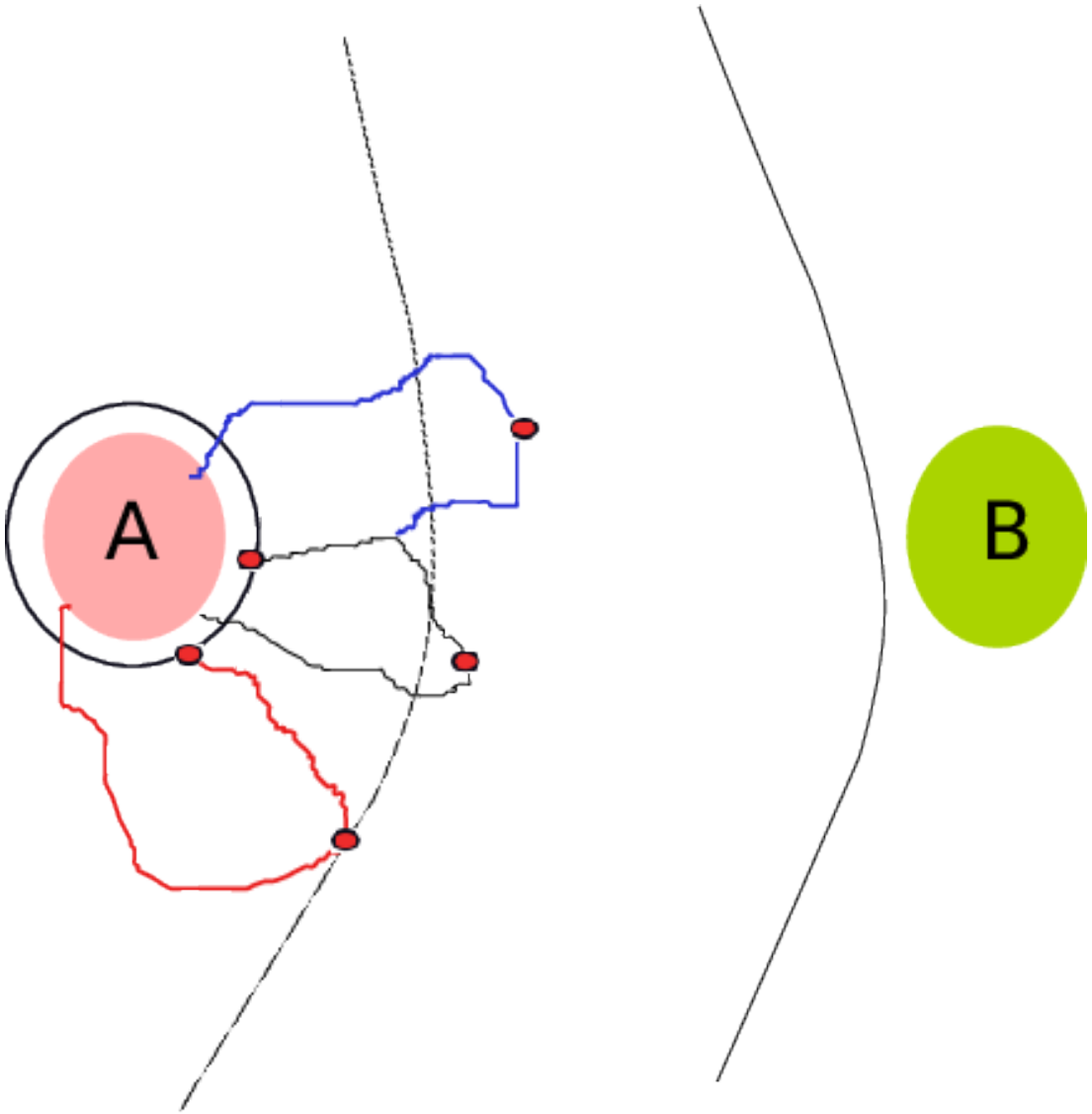
Adaptive Multilevel Splitting



Adaptive Multilevel Splitting



Adaptive Multilevel Splitting



Consistency of static AMS for large mixing

Proposition (Asymptotic Unbiasedness)

Let N be the number of particles be finite and fixed. Assume the mutation kernels associated with Triggered Mutations becomes infinitely mixing that is $M_s \to \eta_s$, then un-normalized estimators becomes unbiased.

Proof–(To be detailed).

Triggered mutations becomes preset mutations given by 'exact target after killing' !! This limit is called the 'idealized case' in the literature[a]. The limit has to be done (e.g. by a coupling argument between M and η) ! □

[a] CE Bréhier, T Lelièvre, M Rousset *Analysis of adaptive multilevel splitting algorithms in an idealized case* 2015

Classification of SMC for 'target' distributions

- Usual obstruction to unbiasedness / Feynman-Kac-Del Moral structure:
  - (Mean-Field) Adaptive Mutation. E.g.: adaptive tuning of rejection rate in Metropolis.
  - Triggered Mutation: Mutation-If-Selection and its special case Mutation-On-Child.
- Algorithms can be indexed either by i) discrete increasing scores $s_{(i)}$, ii) scores associated with effective selection events $..S_{(j)}...$
- Algorithms indexed by effective selection events may exhibit pseudo-adpativity, like dynamic AMS.

Unbiasing any algorithm

In practice using BOTH (biased) adaptive/triggered mutations AND an unbiased Feynman-Kac-Del Moral version is useful for control:

- Run the adaptive version, store the adaptive parameters.
- Dilute the Triggered Mutations into a schedule of Preset Mutations.
- Run the unbiased variant.