Adaptive Importance Sampling
(and applications to Bayesian Statistics)

Gabriel STOLTZ

(in collaboration with Tony Lelièvre and Nicolas Chopin)

CERMICS & MICMAC project team, ENPC (Marne-la-Vallée, France)

http://cermics.enpc.fr/∼stoltz/

Sampling: The metastability issue, and a possible cure

Description of the system

- Configuration $x \in \mathcal{D}$, distributed according to $\pi(dx) = Z^{-1} f(x)\, dx$
- Statistical physics:
  - positions q, momenta $p = M\dot{q}$
  - Microscopic description of a classical system (N particles):

$$(q, p) = (q_1, \dots, q_N,\ p_1, \dots, p_N) \in \mathcal{D}$$

  - For instance, $\mathcal{D} = \mathcal{M} \times \mathbb{R}^{3N}$ with $\mathcal{M} = \mathbb{R}^{3N}$ or $\mathbb{T}^{3N}$
- Hamiltonian (all the physics is contained in V)

$$H(q, p) = \sum_{i=1}^{N} \frac{p_i^2}{2m_i} + V(q_1, \dots, q_N)$$

- Example: pair interactions $V(q_1, \dots, q_N) = \sum_{1 \le i < j \le N} v(|\ q_j - q_i\ |)$

Extracting macroscopic properties: Statistical physics

- Given the structure and the laws of interaction of the particles, what are the macroscopic properties of the matter composed of these particles?
- Equilibrium thermodynamic properties (pressure,...):

$$\langle A \rangle = \int_{\mathcal{D}} A(q,p)\, d\mu(q,p)$$

- Integral in a high dimensional space...
- Choice of thermodynamic ensemble $\equiv$ choice of probability measure $d\mu$:
  - microcanonical (NVE, constant energy) ;
  - canonical (NVT, "constant temperature") : Boltzmann measure

$$d\mu_{\mathrm{NVT}} = \frac{1}{Z_{\mathrm{NVT}}} \exp(-\beta H(q,p))\, dq\, dp, \quad \beta = 1/(k_B T)$$

  - Other choices are possible (grand-canonical, constant pressure,...)
- Certain properties can not be computed this way (free energy, entropy)!

Sampling a Gibbs measure: Overdamped Langevin dynamics

- SDE on the configurational part only (momenta trivial to sample)

$$dq_t = -\nabla V(q_t)\,dt + \sigma\,dW_t,$$

where $(W_t)_{t\geq 0}$ is a standard Wiener process of dimension dN

- Invariance of the canonical measure

$$d\pi(q) = Z^{-1}\mathrm{e}^{-\beta V(q)}\,dq, \qquad Z = \int_{\mathcal{M}} \mathrm{e}^{-\beta V(q)}\,dq$$

if steady state of Fokker-Planck equation $\partial_t \psi_t = \mathrm{div}\left(\nabla V \psi_t + \frac{\sigma^2}{2}\nabla\psi_t\right)$

- Fluctuation/dissipation relation $\sigma = (2/\beta)^{1/2}$
- Invariance + irreducibility (elliptic process):

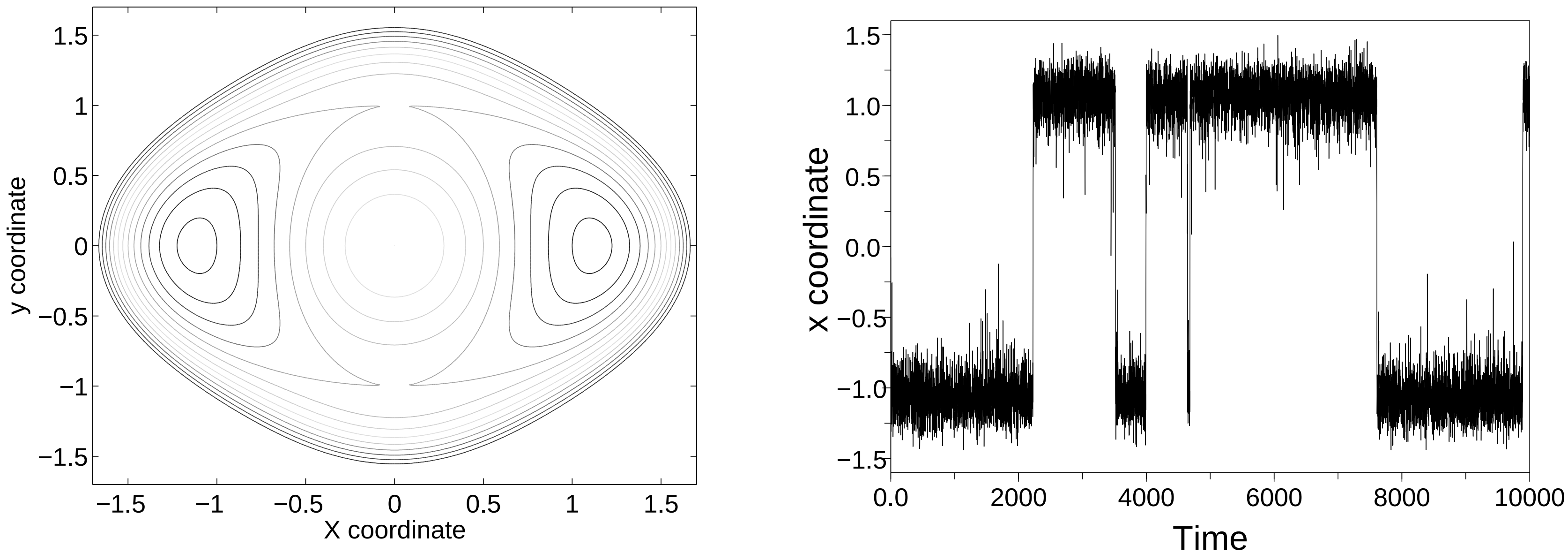
$$\lim_{T\to\infty} \frac{1}{T}\int_0^T A(q_t^x)\,dt = \int_{\mathcal{M}} A(q)d\pi \quad \text{a.s.}$$

Metastability (1)

Numerical discretization of the overdamped Langevin dynamics:

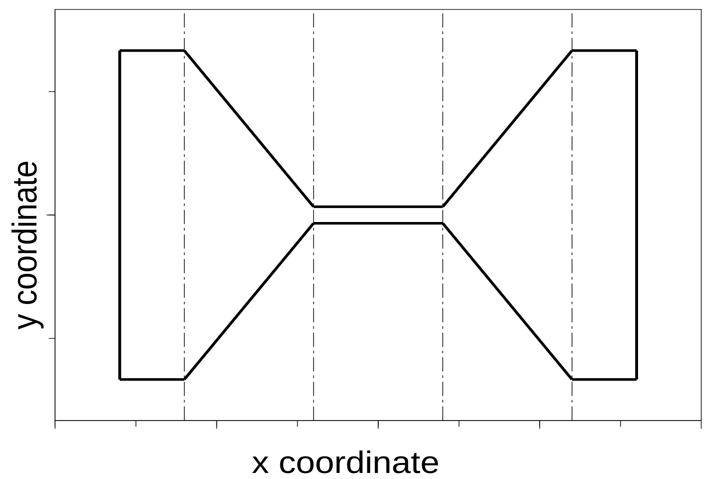
$$q^{n+1} = q^n - \Delta t \nabla V(q^n) + \sigma\sqrt{\Delta t}\, U^n$$

where $U^n \sim \mathcal{N}(0,1)$ i.i.d.

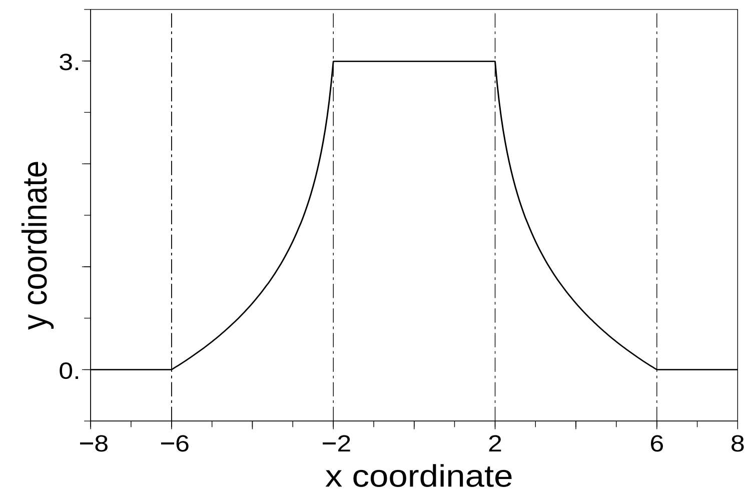


Projected trajectory in the x variable for $\Delta t = 0.01$, $\beta = 6$.

- Although the trajectory average converges to the phase-space average, the convergence may be slow...
- Slowly evolving macroscopic function of the microscopic degrees of freedom
- Two origins : energetic or entropic barriers (in fact, free energy barrier)



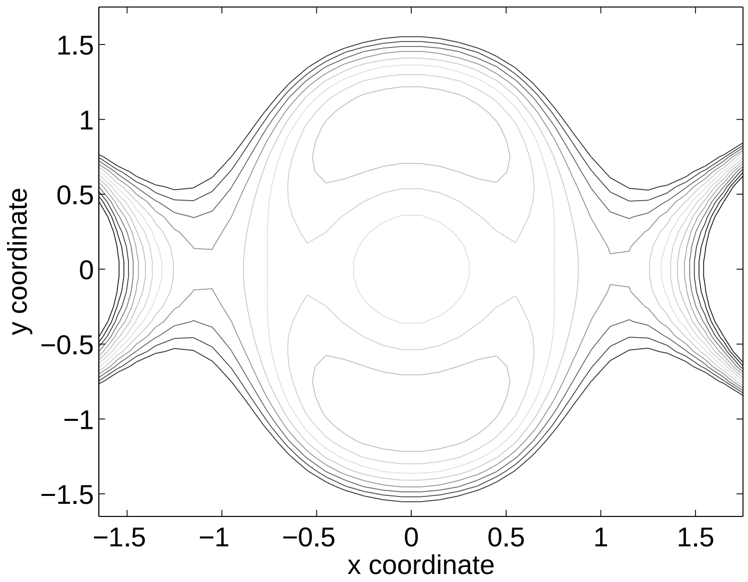
(a) Entropic barrier.

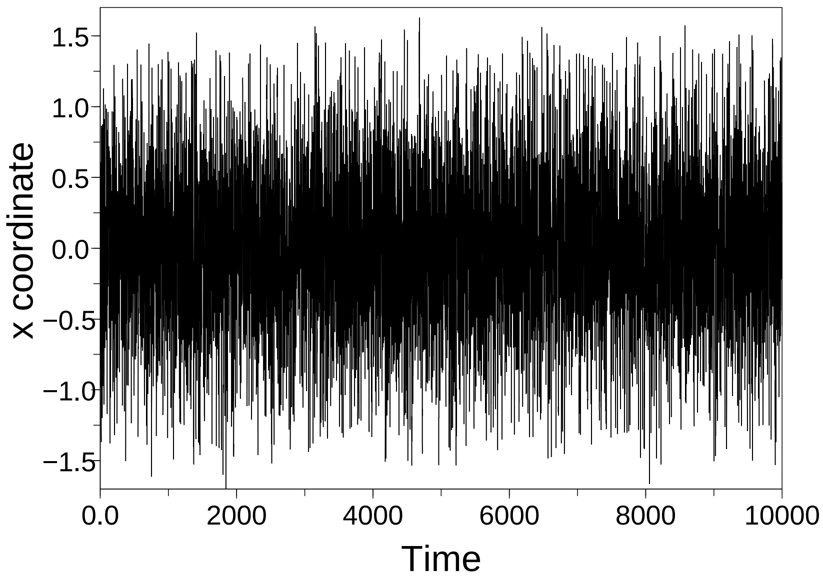


(b) Associated free energy.

Metastability (3)

- Assume the free energy F associated with the slow direction x has been computed, and sample the modified potential $\mathcal{V}(x,y) = V(x,y) - F(x)$.





Projected trajectory in the x variable for $\Delta t = 0.01$, $\beta = 6$.

- Many more transitions! The variable x is uniformly distributed.
- Reweighting with weights $\mathrm{e}^{-\beta F(x)}$ to compute canonical averages

Computation of free energy differences (1)

- Absolute free energy

$$F = -\frac{1}{\beta} \ln Z, \qquad Z = \int_{\mathcal{D}} \mathrm{e}^{-\beta E(x)}\, dx$$

- Motivation (Gibbs, 1902):
  - canonical measure $\mu(dq\,dp) = Z^{-1} \exp(-\beta H(q,p))\, da\, dp$
  - start from the thermodynamic identity $F = U - TS$
  - average energy $U = \displaystyle\int H \mu$
  - entropy $S = -k_{\mathrm{B}} \displaystyle\int \mu \ln \mu$
- Can be computed for ideal gases, and solids at low temperature
- Usually only free energy differences matter!

- Alchemical transition: indexed by an external parameter λ (force field parameter, magnetic field,...)

$$F(1) - F(0) = -\beta^{-1} \ln \left(\frac{\displaystyle\int_{\mathcal{D}} \mathrm{e}^{-\beta E_1(x)}\, dx}{\displaystyle\int_{\mathcal{D}} \mathrm{e}^{-\beta E_0(x)}\, dx} \right) ;$$

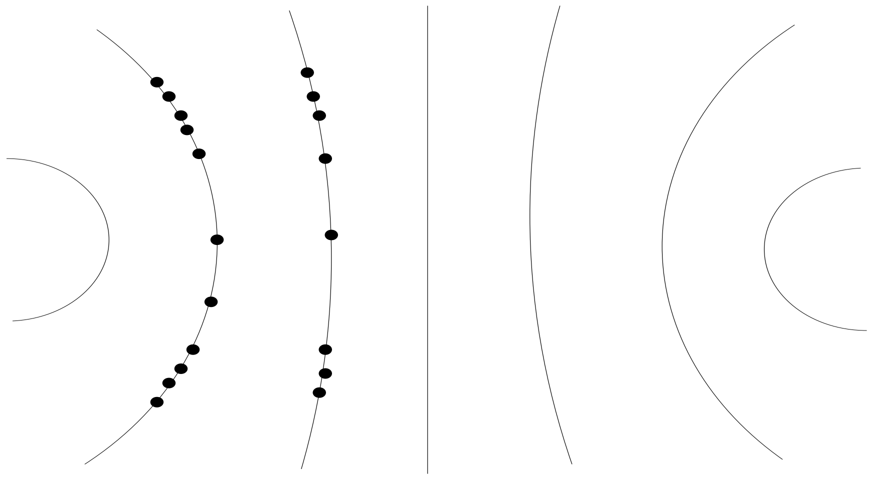
Typically, $E_\lambda = (1-\lambda)E_0 + \lambda E_1$

- (given) reaction coordinate $\xi \,:\, \mathcal{D} \to \mathbb{R}^m$ (angle, length,...):

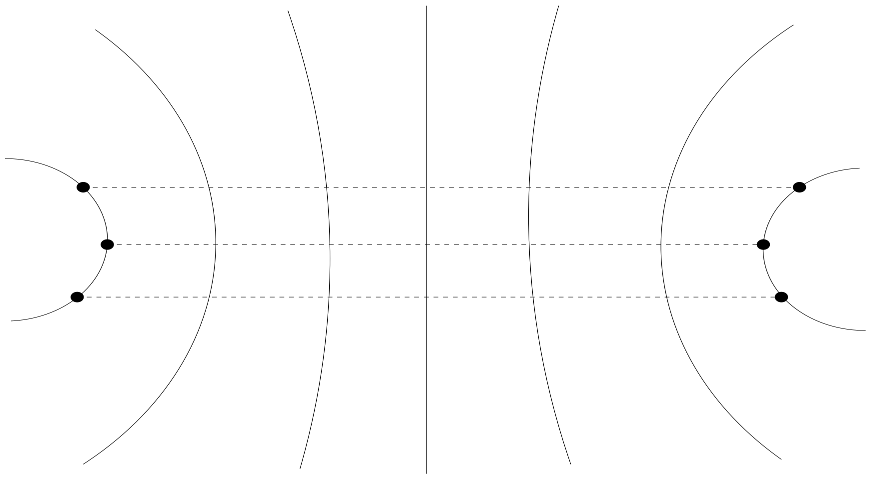
$$F(z_1) - F(z_0) = -\beta^{-1} \ln \left(\frac{\displaystyle\int_{\mathcal{D}} \mathrm{e}^{-\beta E(x)}\, \delta_{\xi(x)-z_1}\, dx}{\displaystyle\int_{\mathcal{D}} \mathrm{e}^{-\beta E(x)}\, \delta_{\xi(x)-z_0}\, dx} \right) .$$

Recall $\delta_{\xi(x)-z}(dx) = |\nabla\xi(x)|^{-1}\sigma_{\Sigma_z}(dx)$, submanifold $\Sigma_z = \xi^{-1}\{z\}$

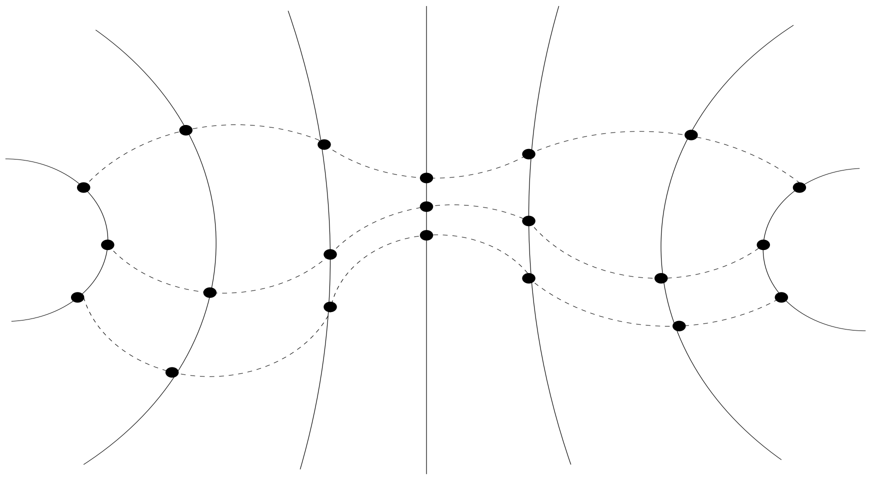
Cartoon comparison of the methods



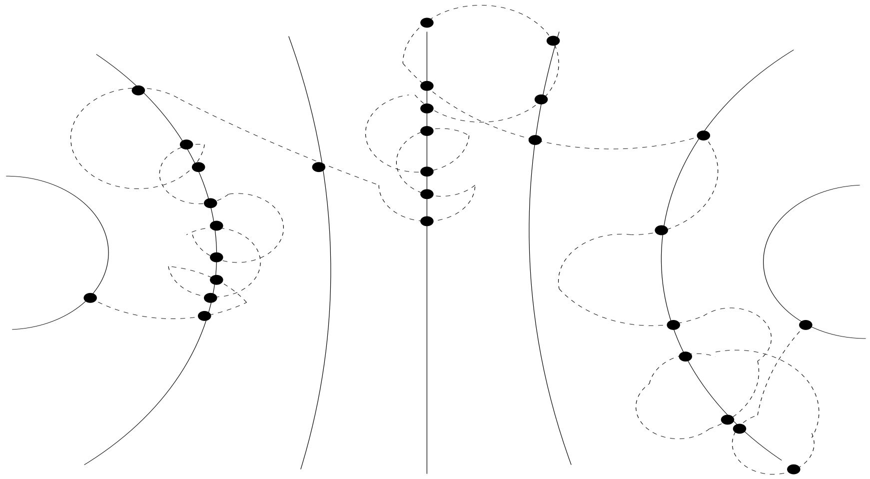
(a) Thermodynamic integration



(b) Free energy perturbation



(c) Nonequilibrium switching dynamics



(d) Adaptive dynamics

Mathematical classification (april 2008)

Free energy perturbation	$\longrightarrow$	Homogeneous MCs and SDEs
Thermodynamic integration	$\longrightarrow$	Projected MCs and SDEs
Nonequilibrium dynamics	$\longrightarrow$	Nonhomogenous MCs and SDEs
Adaptive dynamics	$\longrightarrow$	Nonlinear SDEs and MCs
Selection procedures	$\longrightarrow$	Particle systems and jump processes

Adaptive dynamics: The example of ABF

Adaptive dynamics (1)

- Adaptive methods (*Adaptive biasing force*,[a] *nonequilibrium metadynamics*,[b] etc)
  - General framework[c]
  - Convergence proof in a limiting case[d]
- Simplified setting: $\lambda \in \mathbb{R}/\mathbb{Z}$, $V_\lambda(q) \equiv V(q,\lambda)$

$$\begin{cases} dq_t = -\nabla_q V(q_t,\lambda_t)\,dt + \sqrt{2\beta^{-1}}\,dW_t^q \\ d\lambda_t = -\partial_\lambda V(q_t,\lambda_t)\,dt + \sqrt{2\beta^{-1}}\,dW_t^\lambda \end{cases}$$

so that $F(\lambda_2) - F(\lambda_1) = -\beta^{-1} \ln \dfrac{\overline{\psi}_{\rm eq}(\lambda_2)}{\overline{\psi}_{\rm eq}(\lambda_1)}$, with $\overline{\psi}_{\rm eq}(\lambda) = \displaystyle\int_{\mathcal{D}} \mathrm{e}^{-\beta V(q,\lambda)}\,dq$

---

[a] Darve and Pohorille, *J. Chem. Phys.* (2001)

[b] Bussi, Laio and Parrinello, *Phys. Rev. Lett.* (2006)

[c] T. Lelièvre, M. Rousset and G. Stoltz, *J. Chem. Phys.* (2007)

[d] T. Lelièvre, M. Rousset and G. Stoltz, *Nonlinearity* (2008)

Adaptive dynamics (2)

- Metastable sampling in the λ variable... Introduction of a bias in the dynamics of λ to force the exploration
- The ideal case would be

$$\left\{\begin{array}{rcl} dq_t &=& -\nabla_q V(q_t,\lambda_t)\,dt + \sqrt{2\beta^{-1}}\,dW_t^q \\ d\lambda_t &=& -\partial_\lambda\left[V(q_t,\lambda_t) - F(\lambda_t)\right]\,dt + \sqrt{2\beta^{-1}}\,dW_t^\lambda \\ \partial_\lambda F(z) &=& \mathbb{E}_{\rm eq}\Big(\partial_\lambda V(q,\lambda)\Big) \end{array}\right.$$

- A natural approximation is to use the current estimate of the force

$$\left\{\begin{array}{rcl} dq_t &=& -\nabla_q V(q_t,\lambda_t)\,dt + \sqrt{2\beta^{-1}}\,dW_t^q \\ d\lambda_t &=& -\partial_\lambda\left[V(q_t,\lambda_t) - F_{\rm bias}(t,\lambda_t)\right]\,dt + \sqrt{2\beta^{-1}}\,dW_t^\lambda \\ \partial_\lambda F_{\rm bias}(t,z) &=& \mathbb{E}\Big(\partial_\lambda V(q_t,z)\Big) \end{array}\right.$$

General case: some geometry...

- Additional terms related to the fact that $|\nabla\xi| \neq 1$
- Reaction coordinate case

$$\pi^\xi(dz) = \left(\int_{\Sigma(z)} Z_\pi^{-1} \mathrm{e}^{-\beta E(x)} \, |\nabla\xi(x)|^{-1} \, \sigma_{\Sigma(z)}(dx) \right) dz = \mathrm{e}^{-\beta F(z)} \, dz.$$

- Mean force $\nabla F(z) = \displaystyle\int_{\Sigma(z)} f(x) \, \pi^\xi(dx \,|\, z)$ with

$$f(x) = \frac{\nabla\xi(x) \cdot \nabla V(x)}{|\nabla\xi(x)|^2} - \frac{1}{\beta} \div \left(\frac{\nabla\xi(x)}{|\nabla\xi(x)|^2} \right)$$

- Dynamics $\left\{ \begin{array}{rcl} dq_t & = & -\nabla\Big(V - F_t \circ \xi\Big)\, dt + \sqrt{\dfrac{2}{\beta}}\, dW_t \\ \partial_z F_t(z) & = & \mathbb{E}\Big(f(q_t) \,\Big|\, \xi(q_t) = z \Big) \end{array} \right.$

- In practice, the following conditional expectation is required for the update of the bias:

$$\mathbb{E}\Big(\partial_\lambda V(q_t,\lambda)\Big) = \frac{\displaystyle\int_{\mathcal{D}} \partial_\lambda V(q,\lambda)\,\psi_t(q,\lambda)\,dq}{\displaystyle\int_{\mathcal{D}} \psi_t(q,\lambda)\,dq}$$

- There are two (complementary) strategies to compute it:
  - using a large number of replicas $(q_t^{i,M},\lambda_t^{i,M})_{i=1,\dots,M}$ of the system which all contribute to the same free energy profile

$$\psi_t(q,\lambda) \simeq \frac{1}{M}\sum_{i=1}^{M} \delta^{\varepsilon}{}_{(q_t^{i,M},\lambda_t^{i,M})-(q,\lambda)};$$

  - resorting to some time average

$$\psi_t(q,\lambda) \simeq \frac{1}{T}\int_{t-T}^{t} \delta^{\varepsilon}{}_{(q_s,\lambda_s)-(q,\lambda)}\,ds.$$

Adaptive dynamics: convergence

- Adaptive biasing force = nonlinear PDE on the law $\psi_t(q,\lambda)$:

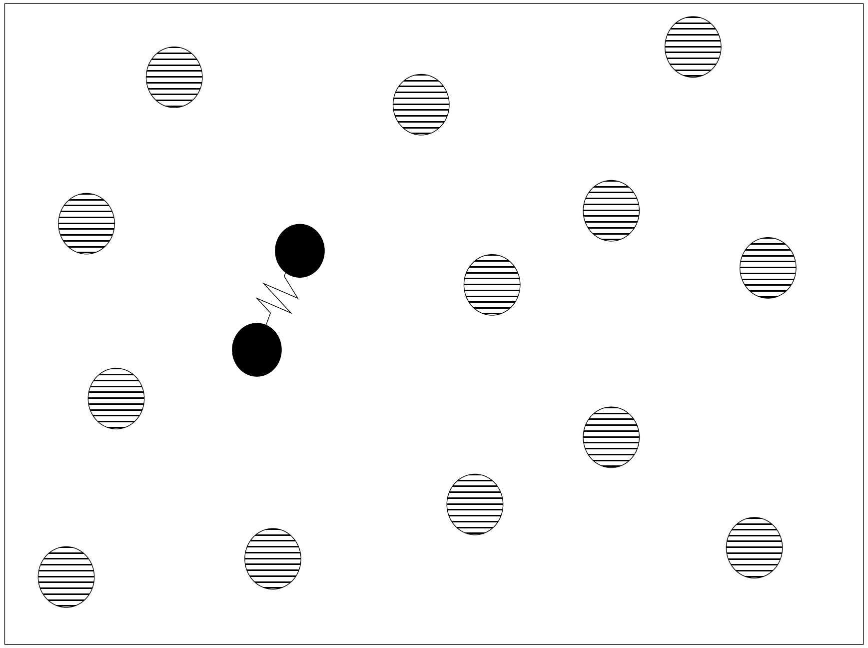
$$\left\{\begin{aligned} &\partial_t \psi_t = \mathrm{div}\Big(\nabla(V - F_{\mathrm{bias}}(t,\lambda))\psi_t + \beta^{-1}\nabla\psi_t\Big), \\ &\partial_\lambda F_{\mathrm{bias}}(t,\lambda) = \frac{\displaystyle\int_{\mathcal{D}} \partial_\lambda V(q,\lambda)\,\psi_t(q,\lambda)\,dq}{\displaystyle\int_{\mathcal{D}} \psi_t(q,\lambda)\,dq}. \end{aligned}\right.$$

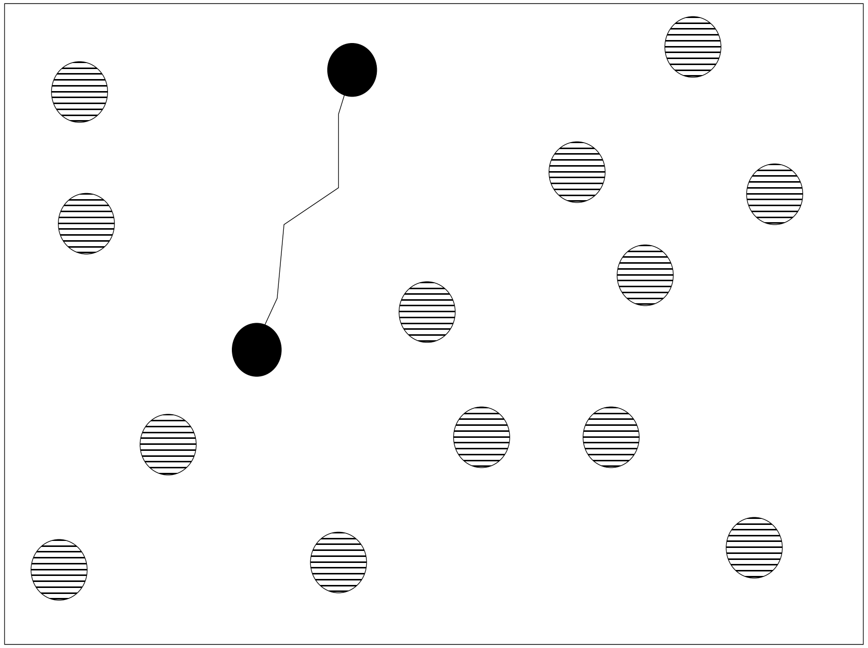
- Simple diffusion for the marginals $\partial_t \overline{\psi}_t = \partial_{\lambda\lambda}\,\overline{\psi}_t$
- Entropic method: decomposition[a] of the total entropy $H(\psi_t\,|\,\psi_\infty) = \displaystyle\int_{\mathcal{M}\times\mathbb{T}} \ln\left(\frac{\psi_t}{\psi_\infty}\right)\psi_t$ into a macroscopic contribution (marginals in λ) and a microscopic one (conditioned measures)
- Convergence of the microscopic entropy provided some uniform logarithmic Sobolev inequality on the conditioned measures holds

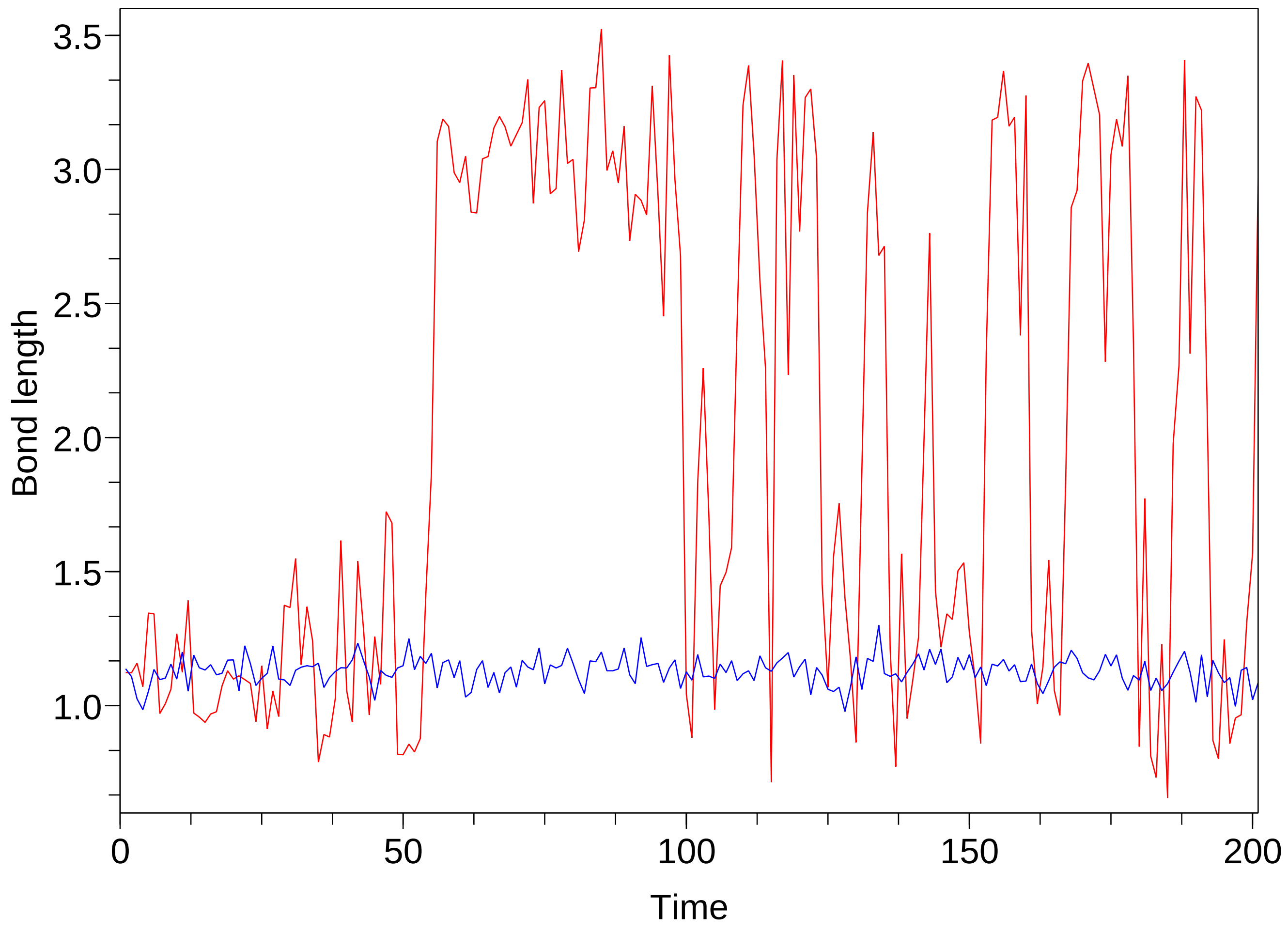
[a] T. Lelièvre, M. Rousset and G. Stoltz, *Nonlinearity* **21** (2008) (merci Felix Otto)

Application: Solvatation effects on conformational changes (1)

- Two particules (q_1,q_2) interacting through $V_{\rm S}(r) = h\left[1 - \frac{(r-r_0-w)^2}{w^2}\right]^2$
- Solvent: particules interacting through the purely repulsive potential $V_{\rm WCA}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] + \varepsilon$ if $r \leq r_0$, 0 if $r > r_0$
- Reaction coordinate $\xi(q) = \frac{|q_1 - q_2| - r_0}{2w}$, compact state $\xi^{-1}(0)$, stretched state $\xi^{-1}(1)$



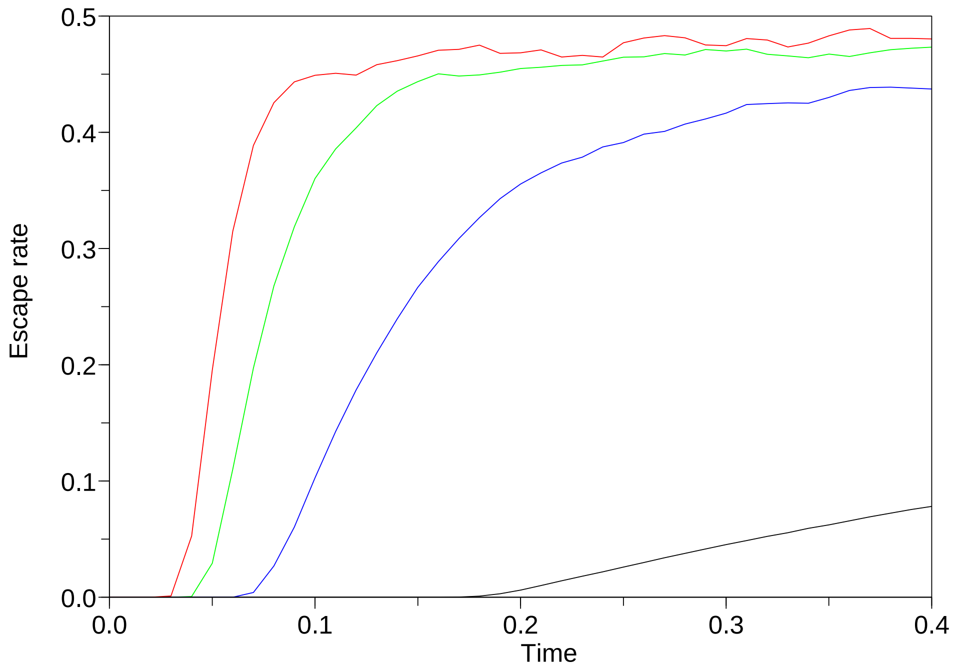




Blue: without biasing term. Red: adaptive biasing force.

Parameters: $h = 10$, density $\rho = 0.25\,\sigma^{-2}$, $w = 1$, $\beta = 3$, $\varepsilon = 1$, $\tau = 0.1$

- Add a selection term in the dynamics $\partial_t \psi = \mathcal{L}^*_\psi \psi + \Big(S_{t,\psi} - \overline{S}_{t,\psi} \Big) \psi$
- For instance, $S_{t,\psi^\xi}(z) = c(t) \dfrac{\Delta_z \psi^\xi(z)}{\psi^\xi}(z)$ leads to an enhanced diffusion $\partial_t \overline{\psi}_t(\lambda) = \big(\beta^{-1} + c(t)\big) \Delta_z \psi^\xi$



Transition rates with increasing selection strenghts.

Application to Bayesian statistics: sampling mixture models

- Distribution of N_{data} values approximated by a mixture of N Gaussians
- Parameters of the mixture

$$x = (q_1, \ldots, q_{N-1}, \mu_1, \ldots, \mu_N, v_1, \ldots, v_N) \in \mathcal{S}_{N-1} \times [\mu_{\min}, \mu_{\max}]^N \times [v_{\min}, +\infty)^N$$

where $\mathcal{S}_{N-1} = \left\{ (q_1, \ldots, q_{N-1}) \;\middle|\; 0 \leq q_i \leq 1, \; \sum_{i=1}^{N-1} q_i \leq 1 \right\}.$

- Weight $q_N = 1 - \sum_{i=1}^{N-1} q_i$
- Corresponding mixture $f(y \,|\, x) = \sum_{i=1}^{N} q_i \sqrt{\frac{v_i}{2\pi}} \exp\left(-\frac{v_i}{2}(y - \mu_i)^2\right)$,
- Likelihood $\Pi(y \,|\, x) = \prod_{i=1}^{N_{\text{data}}} f(y_i \,|\, x).$
- Initial conditions: equal weights, means and variances for the gaussians

Description of the prior

- Random beta model[a] for mixtures: $\beta \sim \Gamma(g,h)$ is an additional variable

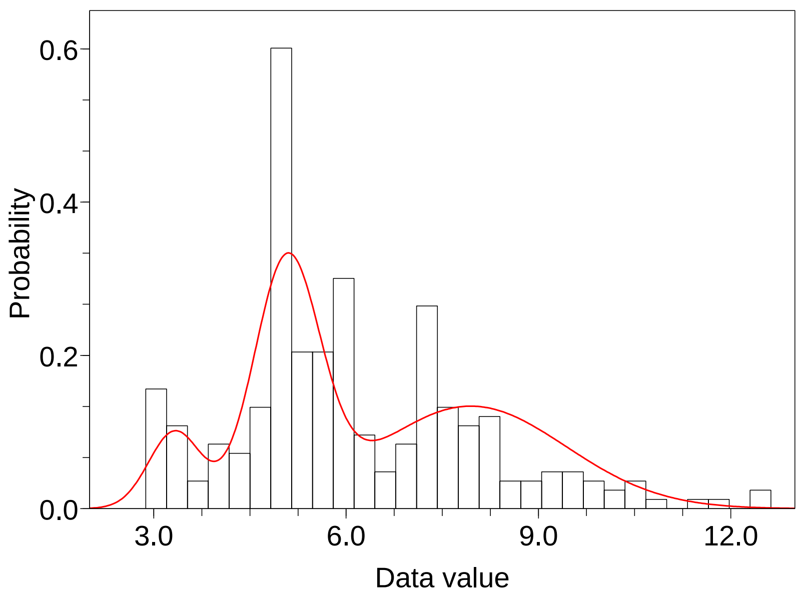
$$(q_1,\dots,q_K) \sim \mathrm{Dirichlet}_K(1,\dots,1), \quad \mu_k \sim \mathcal{N}\left(M, \frac{R^2}{4}\right), \quad v_k \sim \Gamma(\alpha,\beta)$$

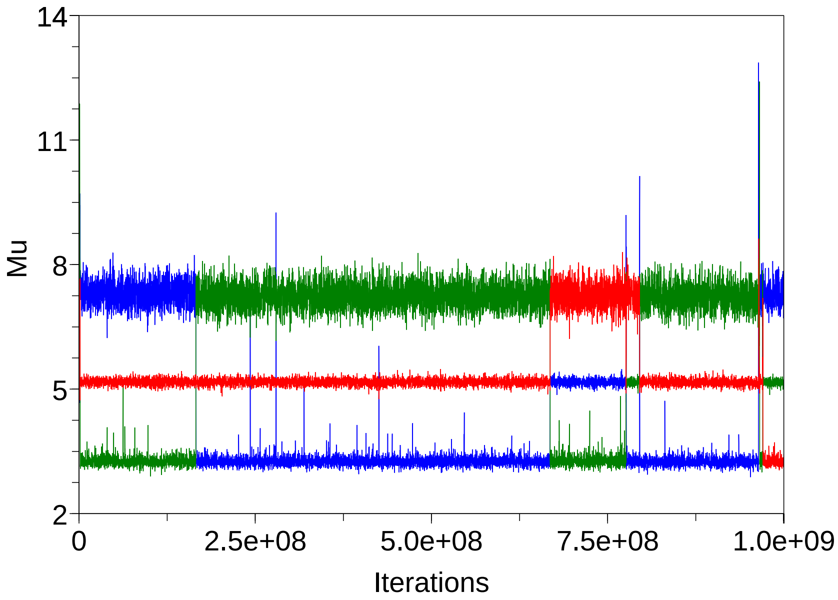
- Parameters: M is the mean of the data, the range $R = \max_{1 \le i \le N_{\mathrm{data}}} y_i - \min_{1 \le i \le N_{\mathrm{data}}} y_i$, and $\alpha = 2$, $g = 0.2$ and $h = 100g/(\alpha R^2)$
- Monte-Carlo dynamics: Metropolis random walk with gaussian proposals characterized by $(\sigma_q, \sigma_\mu, \sigma_v, \sigma_\beta)$
- Binning procedure: mean force and bias in bin (z_i, z_{i+1})

$$F_n^{\Delta z}(z) = \frac{\displaystyle\sum_{j=0}^{n} f(x_j)\,\mathbf{1}_{z_i \le \xi(x^j) \le z_{i+1}}}{\displaystyle\sum_{j=0}^{n} \mathbf{1}_{z_i \le \xi(x^j) \le z_{i+1}}}, \quad A_n(z) = \sum_{k=0}^{i-1} \Delta z\, F_n^{\Delta z}\left(k + \frac{1}{2}\Delta z\right)$$

[a] S. Richardson and P. J. Green. *J. Roy. Stat. Soc. B*, 59(4):731–792, 1997.

Fish data

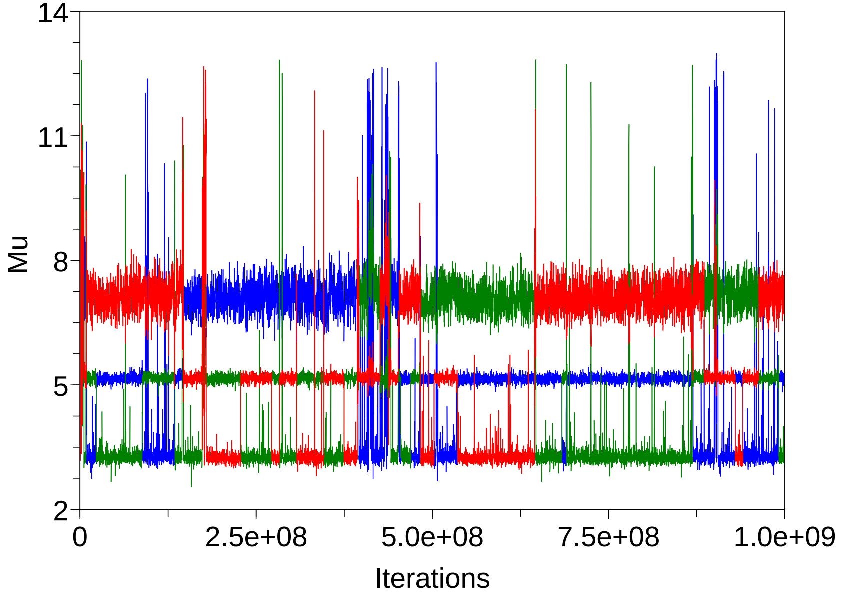


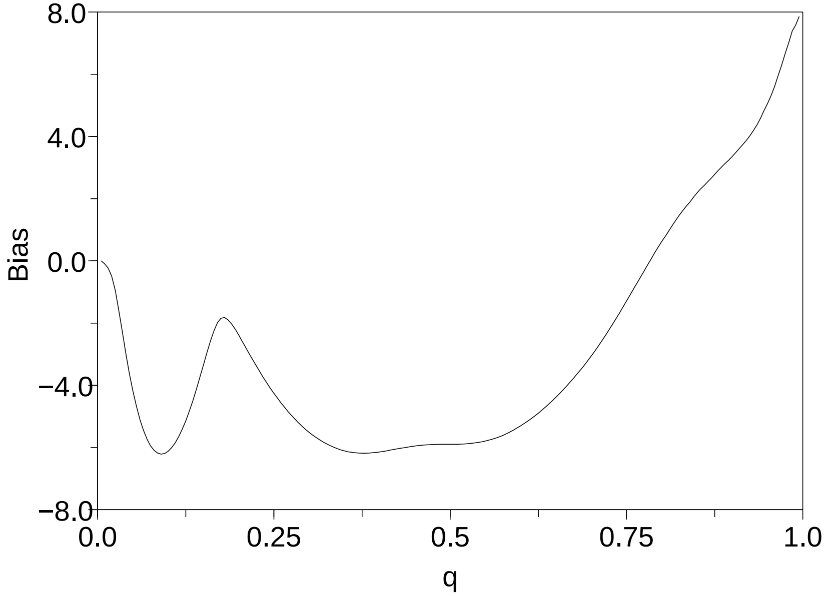


Left: Fish data, and a possible fit using the last configuration from the trajectory plotted in the right picture.

Right: Typical sampling trajectory, gaussian random walk with $(\sigma_q, \sigma_\mu, \sigma_v, \sigma_\beta) = (0.005, 0.025, 0.05, 0.005)$.

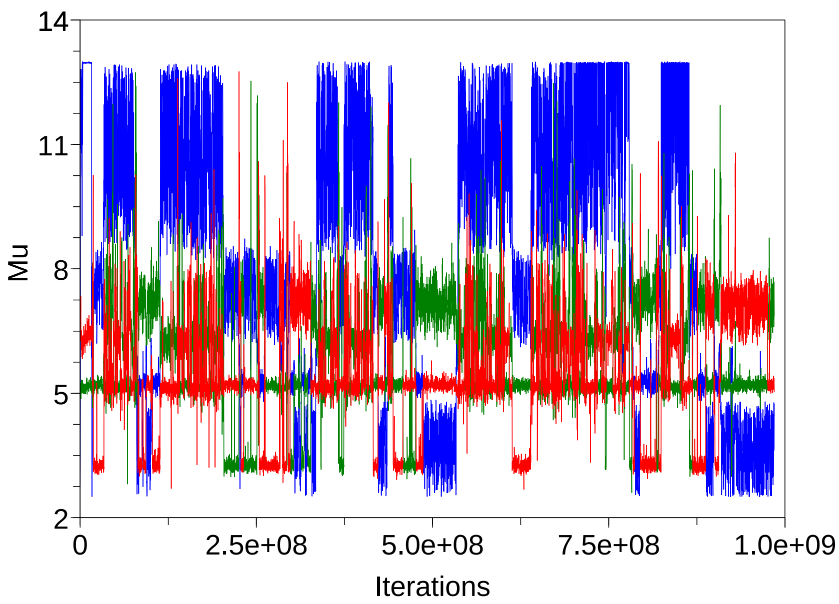
Fish data: biasing q_1

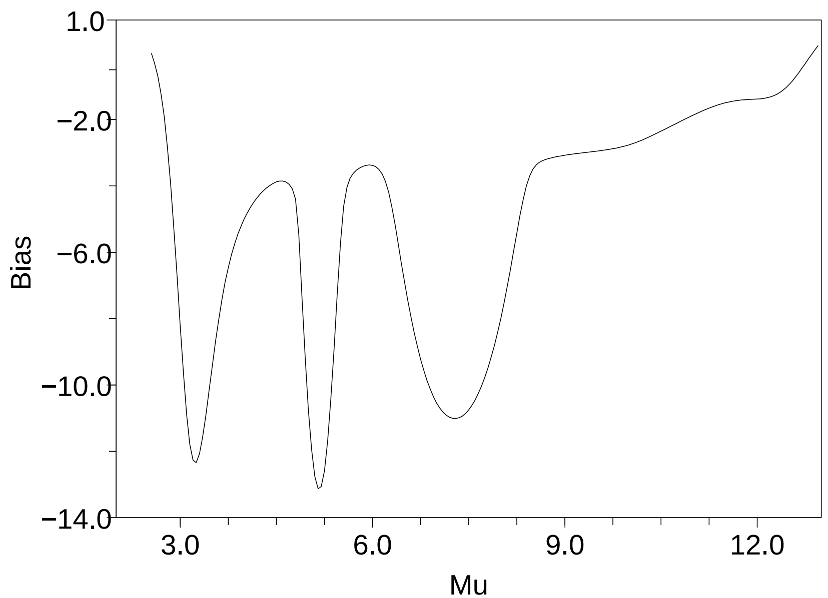




Left: Typical sampling trajectory when the reaction coordinate is q_1.
Right: Associated biasing potential at the end of the simulation.

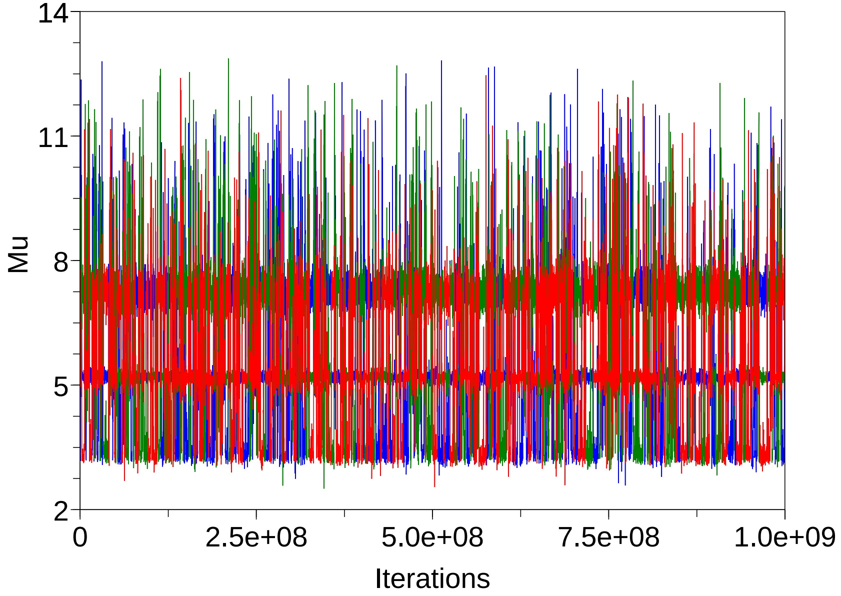
Fish data: biasing μ_1

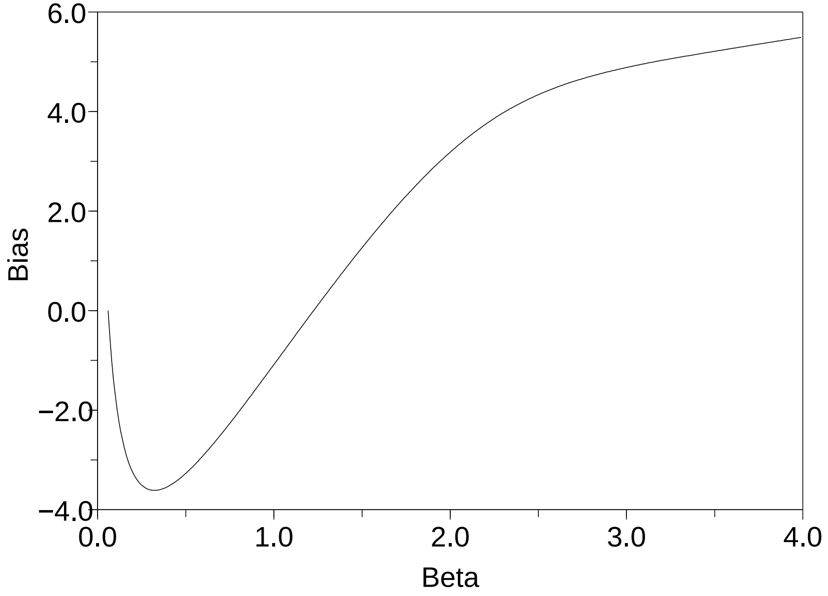




Left: Typical sampling trajectory when the reaction coordinate is μ_1.
Right: Associated biasing potential at the end of the simulation.

Fish data: biasing β

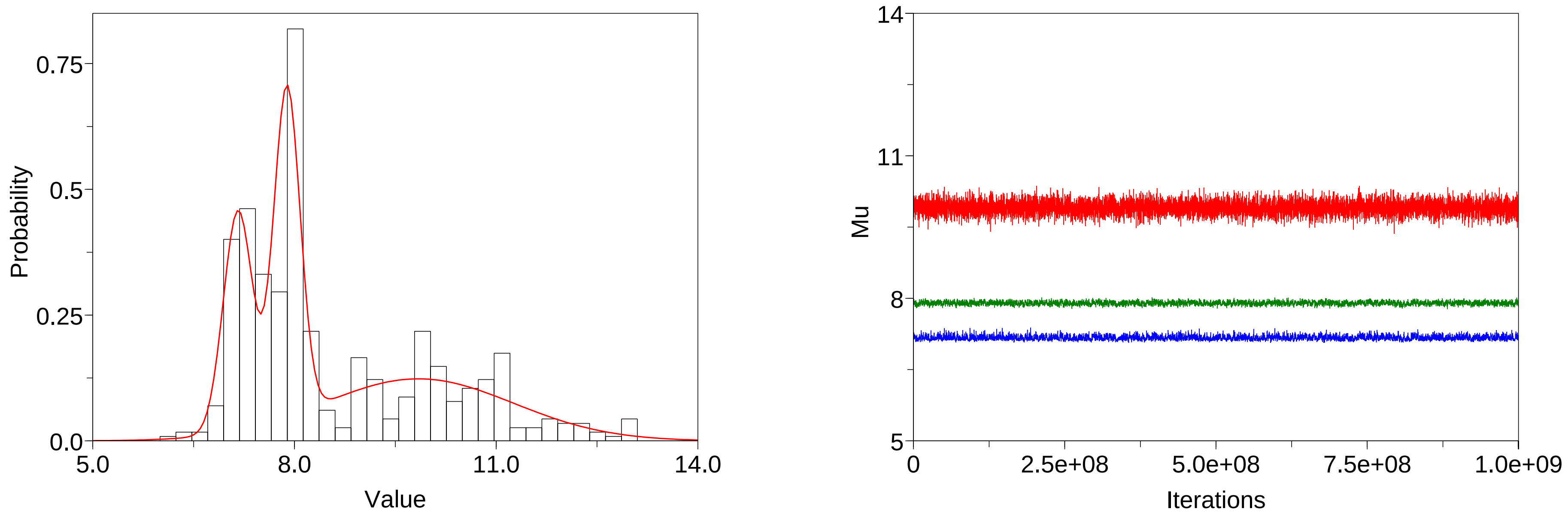




Left: Typical sampling trajectory when the reaction coordinate is β.

Right: Associated biasing potential at the end of the simulation.

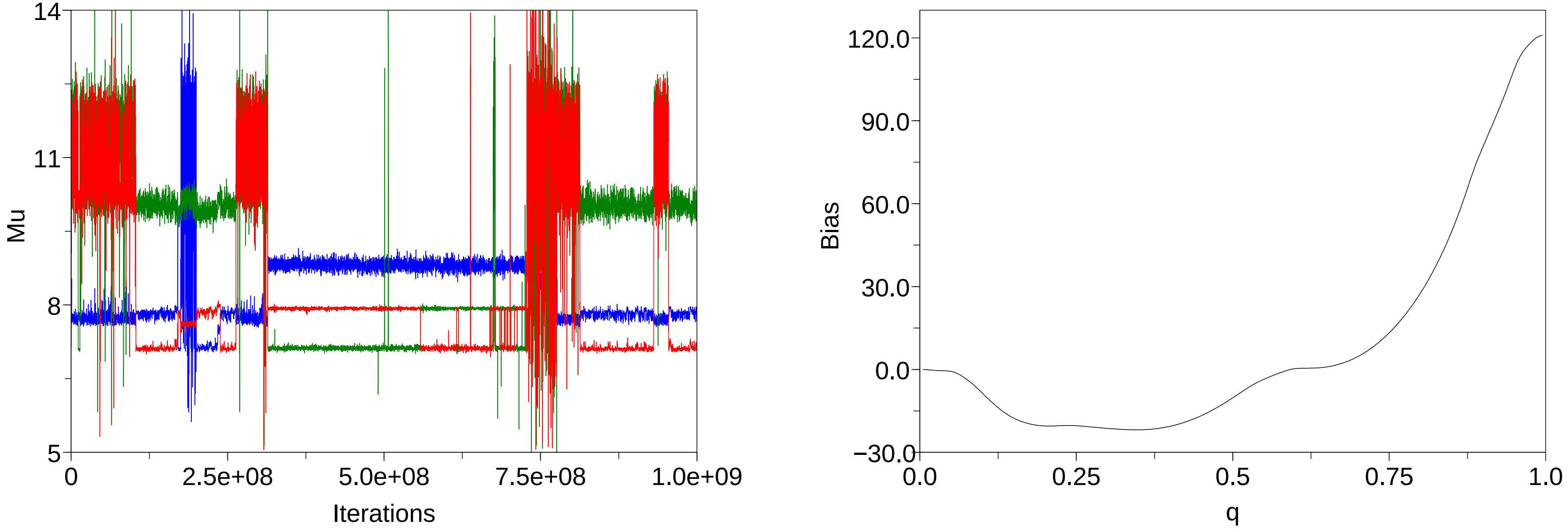
Hidalgo data



Left: Hidalgo data, and a possible fit using the last configuration from the trajectory plotted in the right picture.

Right: Typical sampling trajectory, gaussian random walk with $(\sigma_q, \sigma_\mu, \sigma_v, \sigma_\beta) = (0.001, 0.05, 0.1, 0.005)$.

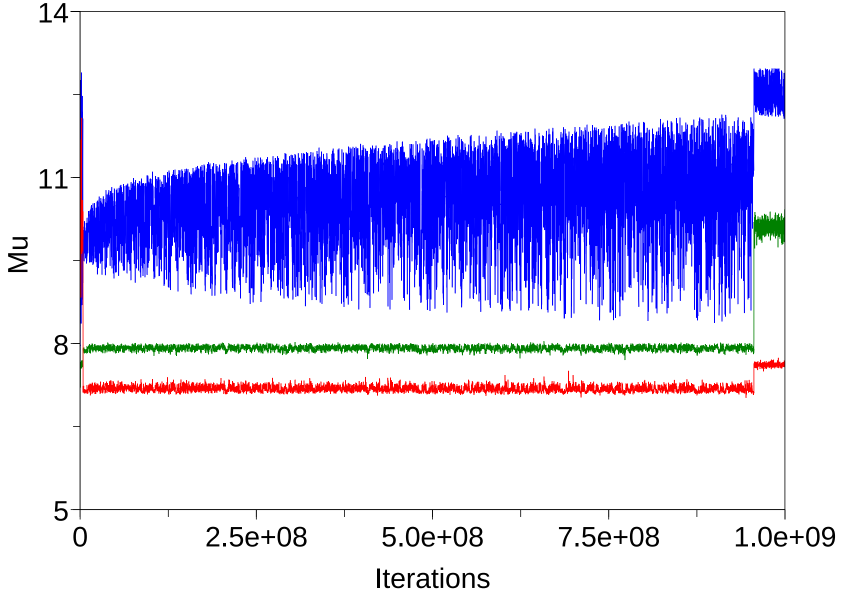
Hidalgo data: biasing q_1

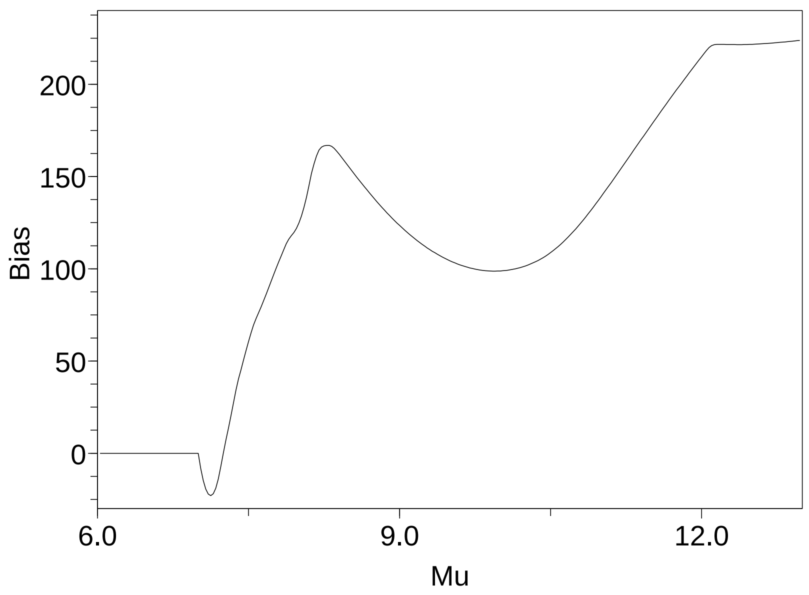


Left: Typical sampling trajectory when the reaction coordinate is q_1.

Right: Associated biasing potential at the end of the simulation.

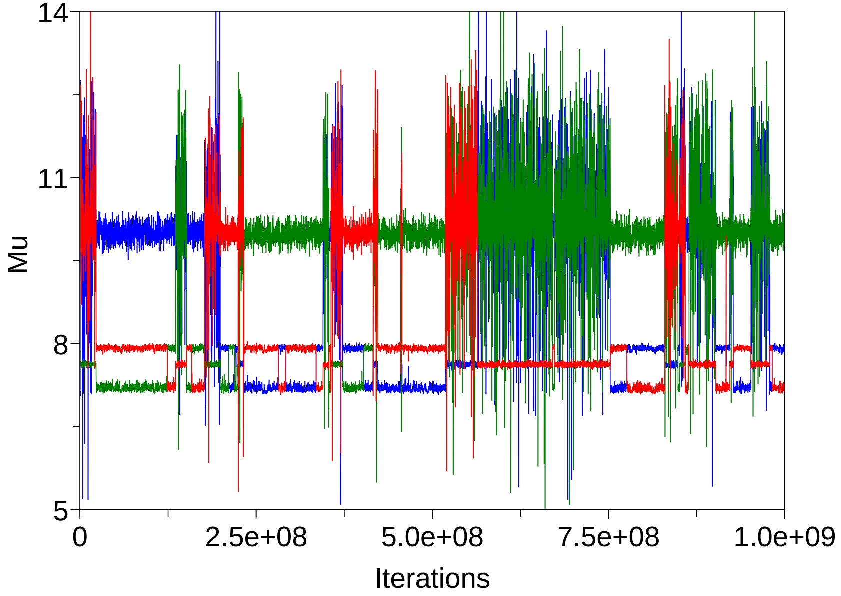
Hidalgo data: biasing μ_1

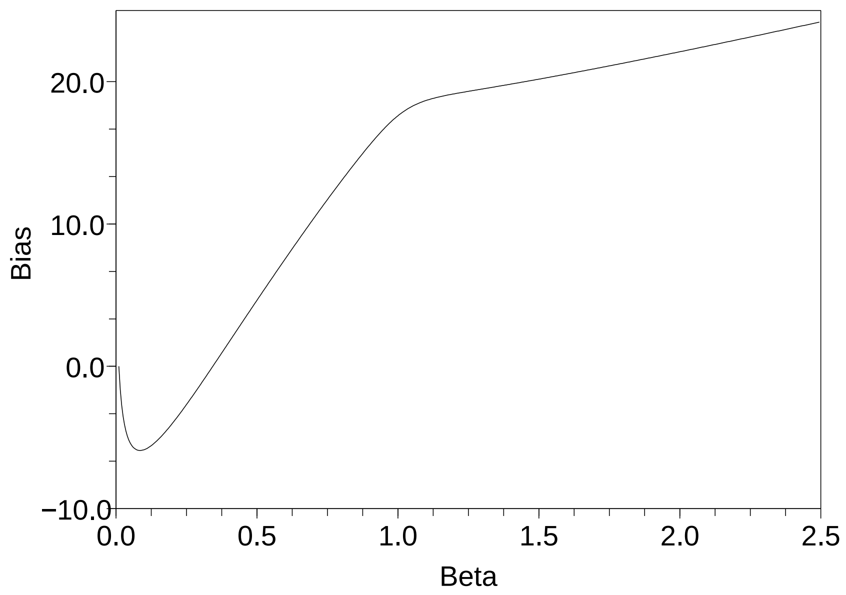




Left: Typical sampling trajectory when the reaction coordinate is μ_1.
Right: Associated biasing potential at the end of the simulation.

Hidalgo data: biasing β





Left: Typical sampling trajectory when the reaction coordinate is β.
Right: Associated biasing potential at the end of the simulation.