Searching for Transition Paths (in Protein Folding)

Henri Orland

IPhT, CEA-Saclay
France

Outline

- The Folding Path problem
- Langevin dynamics and Path integral representation
- Dominant paths
  - Hamilton-Jacobi representation
- Langevin Bridges
  - short time approximation
  - exact numerical solution

1. What is a Protein

Biological Polymers (biopolymers): Proteins, Nucleic Acids (DNA and RNA), Polysaccharides

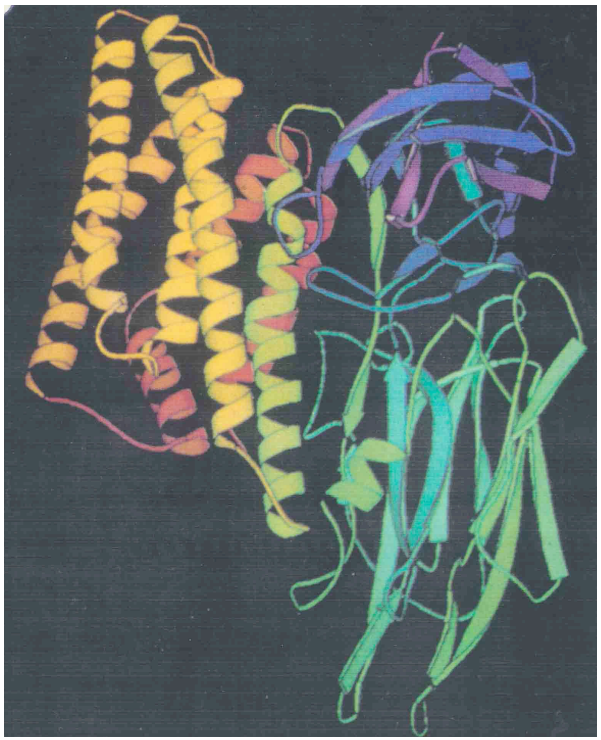
- catalytic activity: enzymes
- transport of ions: hemoglobin (O_2), ion channels
- motor protein
- shell of viruses (influenza, HIV, etc...)
- prions
- food, etc…

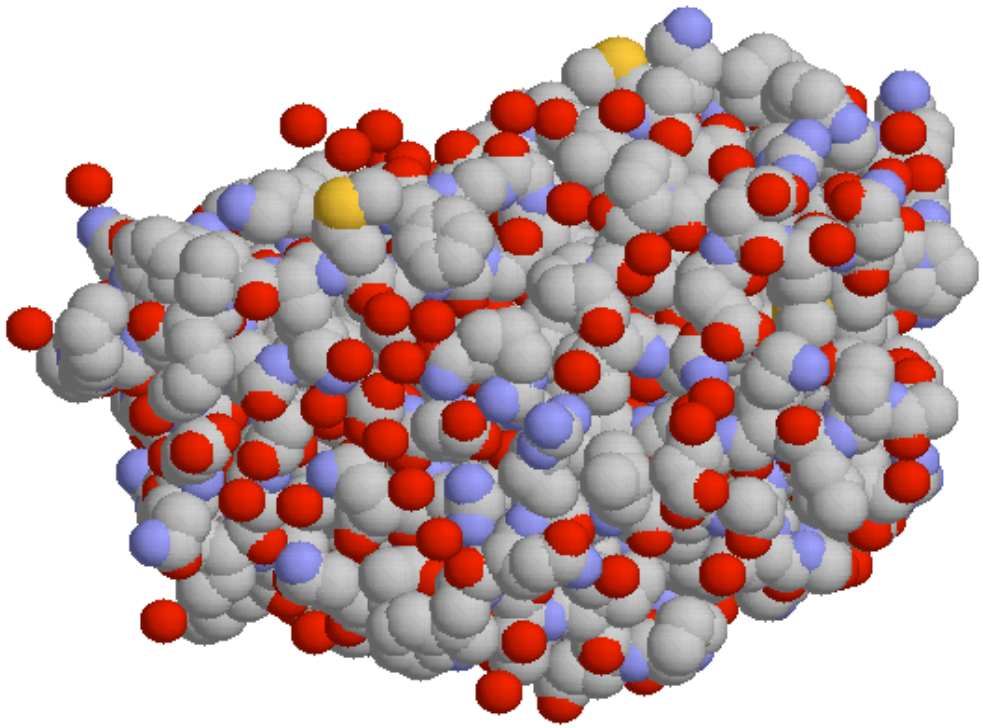
Proteins exist under 2 forms

- Folded or Native: globular unique conformation, biologically active
- Unfolded: random coil, biologically inactive

- Note that a globular polymer has an extensive entropy

$$\mathcal{N} = \mu^N$$





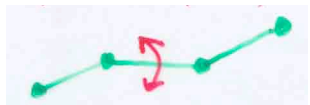
HIV protease (199 residues)

The Protein Folding problem

- A sequence of amino-acids is given by the biologists.
- What is the 3d shape of the corresponding protein?
- To study this problem, try Molecular Dynamics: Karplus, Levitt and Warschel, Nobel prize in Chemistry 2013

Parametrization (CHARMM, AMBER, OPLS, …)






$$E = \sum_{bonds} k_b (b - b_0)^2 + \sum_{valence\ angles} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} k_\varphi (1 + \cos(n\varphi - \delta)) + \sum_{impropers} k_\nu (\nu - \nu_0)^2$$

$$+ \sum_{i<j} 4\varepsilon_{ij} \left((\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6} \right) + \sum_{i<j} \frac{332}{\varepsilon} \frac{q_i q_j}{r_{ij}}$$

in kCal/mol

Use Newton or Langevin dynamics

$$m_i \ddot{r}_i + \gamma_i \dot{r}_i + \frac{\partial E}{\partial r_i} = \eta_i(t)$$

where $\eta_i(t)$ is a Gaussian noise satisfying the fluctuation-dissipation theorem:

$$< \eta_i(t) \eta_j(t) >= 2\gamma_i k_B T \delta_{ij} \delta(t - t')$$

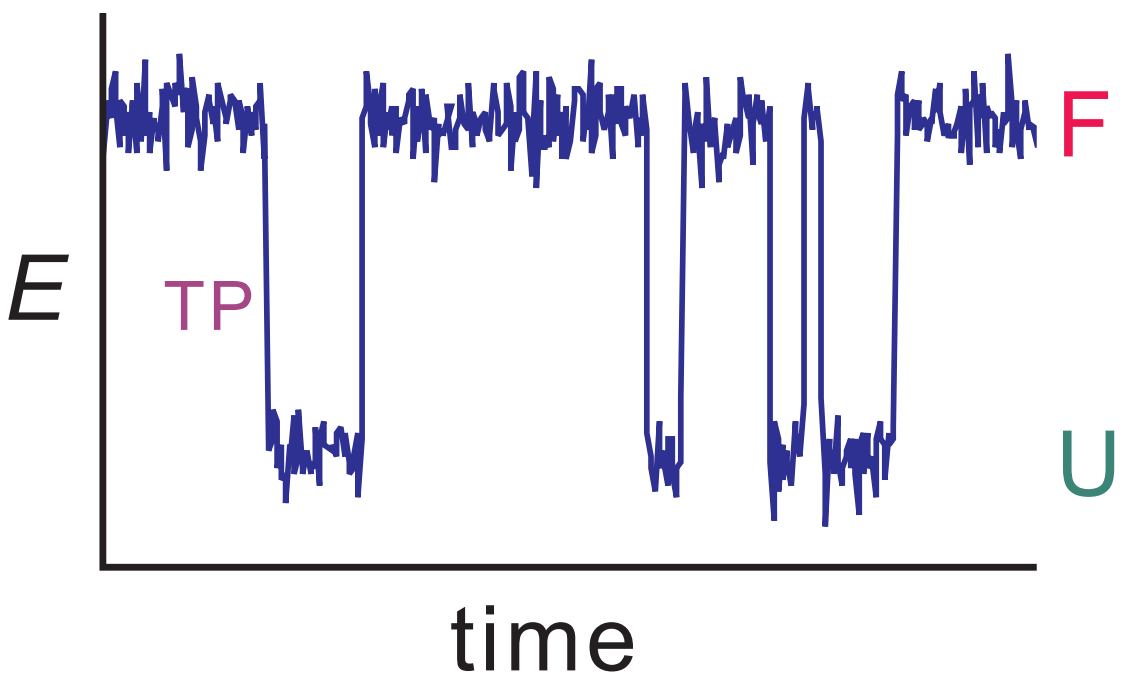
Why it does not work (yet?)?

- To discretize the equations, one must use time steps of the order of $10^{-15}s$
- Large number of degrees of freedom (a few thousand) plus few thousand water molecules
- Force fields not necessarily adapted to folding
- Longest runs: around $1\mu s$ << folding time 1 ms- 1s
- Recently, runs of 1ms on short proteins
- Many metastable states and high barriers

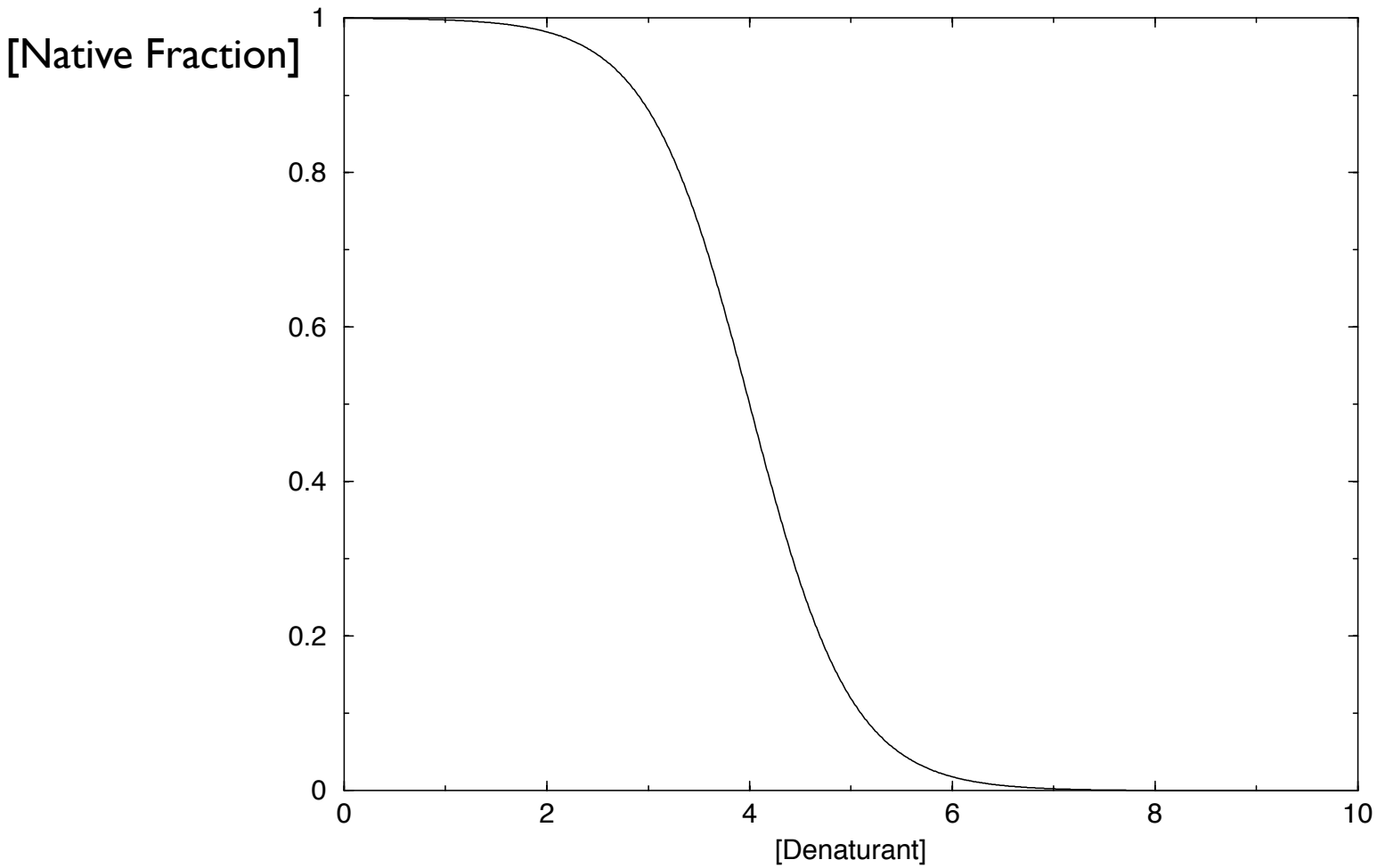
The problem of protein structure prediction is too complicated

Simpler problem: How do proteins fold? How do they go from Unfolded to Native State?

- In given denaturant conditions, a protein spends a fraction of its time in the native state and a fraction of its time in the denatured state.



Denaturation curves



In given denaturant conditions, a fraction of the proteins are native, and the rest are denatured

The Folding Pathway Problem

- The problem: Assume a protein can go from state A to state B. Which pathway (or family of pathways) does the protein take? How are the trajectories from A to B?

Motivation from single molecule experiments

- Examples:
  - from denatured to native in native conditions
  - Allosteric transition between A and B

Difficulty: looking for rare events

Can one describe these reactions in terms of a small set of dominant trajectories with fluctuations around?

Langevin dynamics

- The case of one particle in a potential $U(x)$ at temperature T
- Use Langevin dynamics

$$m\frac{d^2x}{dt^2} + \gamma\frac{dx}{dt} + \frac{\partial U}{\partial x} = \zeta(t)$$

- where γ is the friction and $\zeta(t)$ is a random noise

$$< \zeta(t)\zeta(t') >= 2k_BT\gamma\delta(t-t')$$

Overdamped Langevin dynamics

- At large enough time scale, mass term negligible

$$m\omega^2 \approx \gamma\omega$$

$$\tau \approx 2\pi \frac{m}{\gamma}$$

$$\gamma = \frac{k_B T}{D}$$

$$\tau \approx 10^{-13}\mathrm{s}$$

$D = 10^{-5}\mathrm{cm}^2/\mathrm{s}$ $\qquad$ $m \approx 5.10^{-26}\mathrm{kg}$

- Take overdamped Langevin (Brownian) dynamics

$$\frac{dx}{dt} = -\frac{1}{\gamma}\frac{\partial U}{\partial x} + \eta(t)$$

- with Gaussian noise:

$$< \zeta(t)\zeta(t') >= \frac{2k_BT}{\gamma}\delta(t-t')$$

- γ is the friction coefficient: $D = \frac{k_BT}{\gamma}$

Diffusion coefficient

- Equation of motion is a stochastic equation
- The Probability to find the particle at point *x* at time *t* is given by a Fokker-Planck equation

$$\frac{\partial}{\partial t}P(x,t) = D\frac{\partial}{\partial x}\left(\frac{1}{k_B T}\frac{\partial U}{\partial x}P(x,t) + \frac{\partial P(x,t)}{\partial x}\right)$$

with

$$P(x,0) = \delta(x - x_i)$$

- Fokker-Planck equation looks very much like a Schrödinger equation, except for 1st order derivative. Define

$$P(x,t) = e^{-\frac{\beta U(x)}{2}} Q(x,t)$$

- The function $Q(x,t)$ satisfies an imaginary time Schrödinger equation with a Hamiltonian H

$$-\frac{\partial Q}{\partial t} = HQ$$

- where H is a "quantum" Hamiltonian given by

$$H = \frac{1}{\gamma}\left(-\nabla^2 + \frac{1}{4}(\nabla U)^2 - \frac{k_B T}{2}\nabla^2 U\right)$$

$$P(x_f, t_f | x_i, t_i) = e^{-\frac{U(x_f)-U(x_i)}{2k_B T}} < x_f | e^{-(t_f - t_i)H} | x_i >$$

- Spectral decomposition

$$< x_f | e^{-(t_f - t_i)H} | x_i >= \sum_\alpha e^{-(t_f - t_i)E_\alpha} \Psi_\alpha(x_f)\Psi_\alpha(x_i)$$

$$H\Psi_\alpha(x) = E_\alpha \Psi_\alpha(x)$$

- At large time, the matrix element is dominated by the ground state

$$\Psi_0(x) = \frac{e^{-\beta U(x)/2}}{\sqrt{Z}}$$

$$Z = \int dx e^{-\beta U(x)}$$

with

$$H\Psi_0 = 0$$

so that

$$P(x_f, t_f | x_i, t_i) \approx \frac{e^{-\beta U(x)}}{Z} + e^{-\beta \frac{U(x_f)-U(x_i)}{2}} e^{-(t_f - t_i)E_1} \Psi_1(x_f)\Psi_1(x_i)$$

$\tau = E_1^{-1}$ is the reaction time

- Stationary distribution: the Boltzmann distribution

$$\lim_{t\to+\infty} P(x,t) = P(x) \sim \exp(-U(x)/k_B T)$$

- General form: Path Integral

$$P(x_f,t_f|x_i,t_i) = e^{-\frac{U(x_f)-U(x_i)}{2k_B T}} \int_{x_i}^{x_f} \mathcal{D}x(\tau)\, e^{-S_{eff}[x]/k_B T}$$

- Boundary conditions:

$$x(t_i) = x_i \qquad x(t_f) = x_f$$

Path Integral representation

$$P(x_f, t_f | x_i, t_i) = e^{-\frac{U(x_f)-U(x_i)}{2k_B T}} \int_{x_i}^{x_f} \mathcal{D}x(\tau)\, e^{-S_{eff}[x]/k_B T}$$

- The effective action is given by

$$S_{eff}[x] = \int_{t_i}^{t_f} dt \left(\frac{\gamma}{4}\dot{x}^2 + V_{eff}[x(t)]\right)$$

- and the effective potential is given by

$$V_{eff}[x] = \frac{1}{4\gamma}\left((\nabla U)^2 - 2k_B T \nabla^2 U\right)$$

Saddle-Point method: WKB approximation

work in collaboration with

P. Faccioli, F. Pederiva, M. Sega
University of Trento

To compute the path integral, look for paths which have the largest weight: semi-classical approximation.

- Dominant trajectories: classical trajectories

$$\frac{\gamma}{2}\frac{d^2x}{dt^2} = -\frac{\partial(-V_{eff}[x])}{\partial x}$$

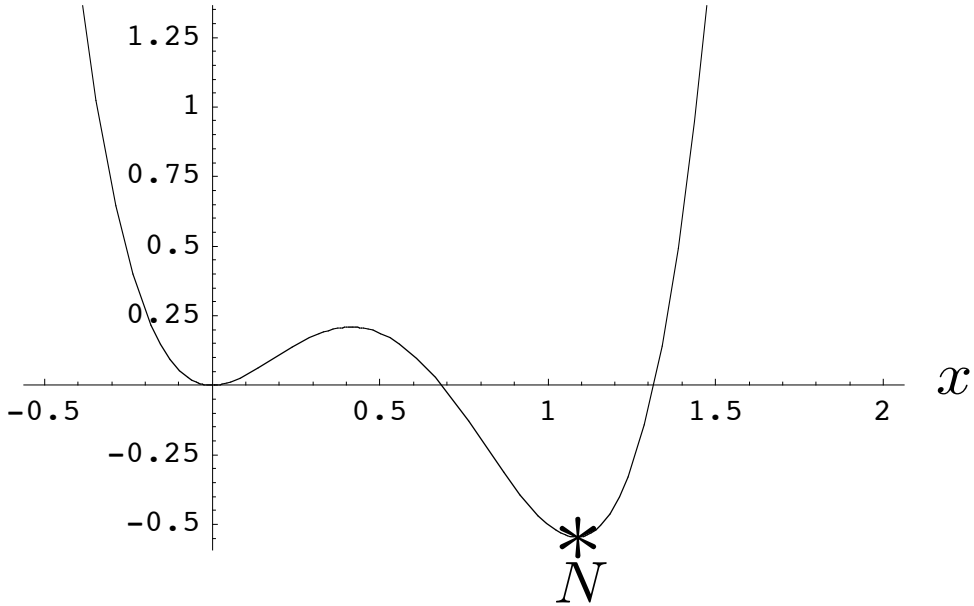
inverted potential

- with correct boundary conditions.

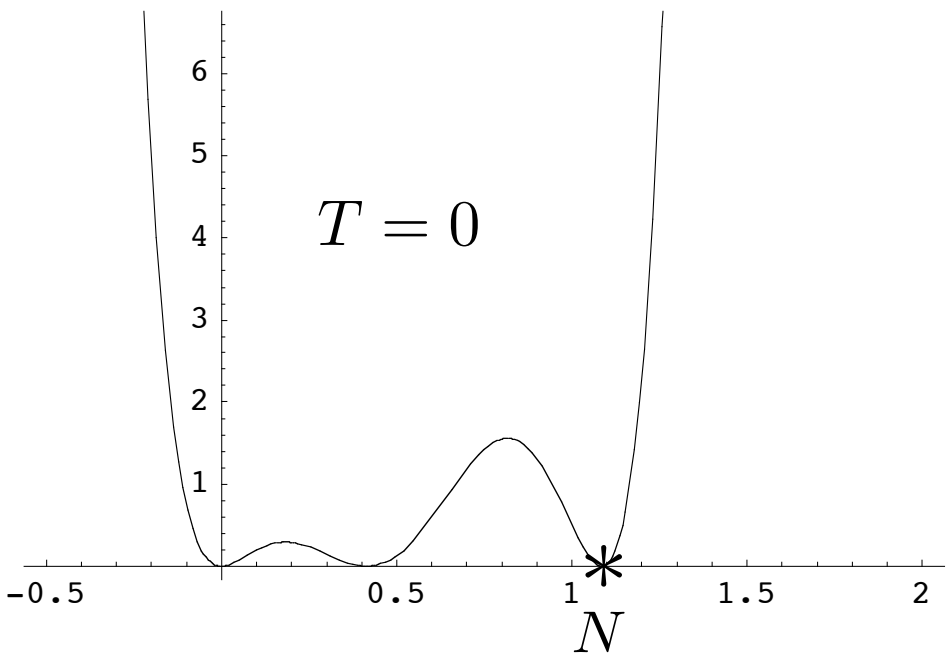
$$x(t_i) = x_i \qquad\qquad x(t_f) = x_f$$

- Problem: one does not know the transition time. Inverse folding rate is equal to mean first passage time (first passage time is distributed).

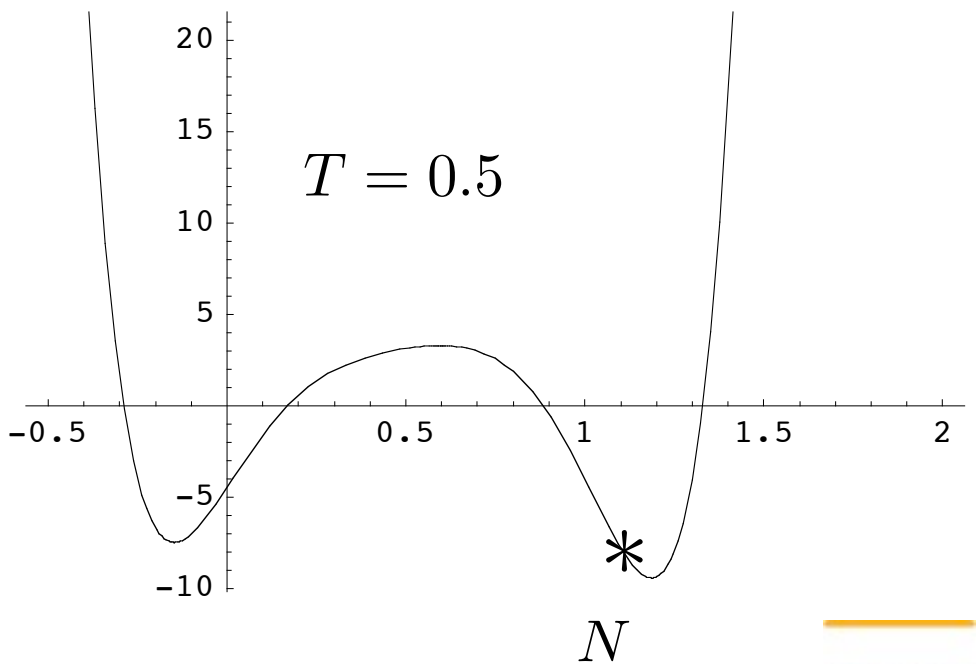
$$U(x) = x^2(5(x-1)^2 - 0.5)$$

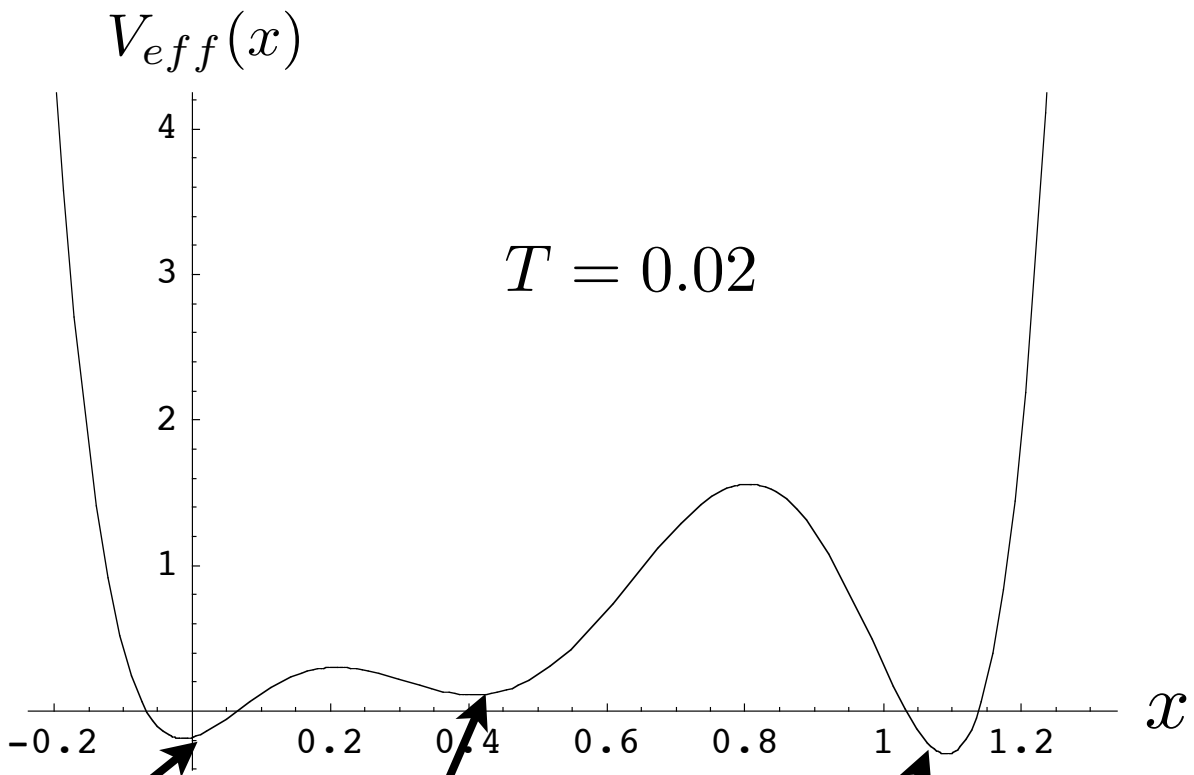


$V_{eff}(x) = U'(x)^2/2 - TU''(x)$



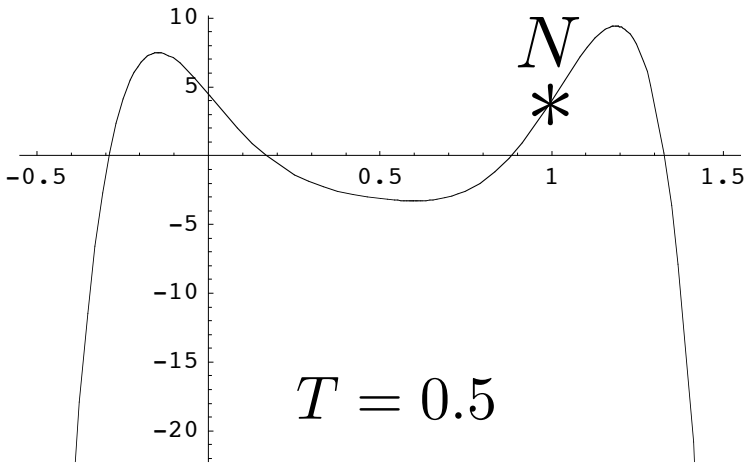
$V_{eff}(x) = U'(x)^2/2 - TU''(x)$

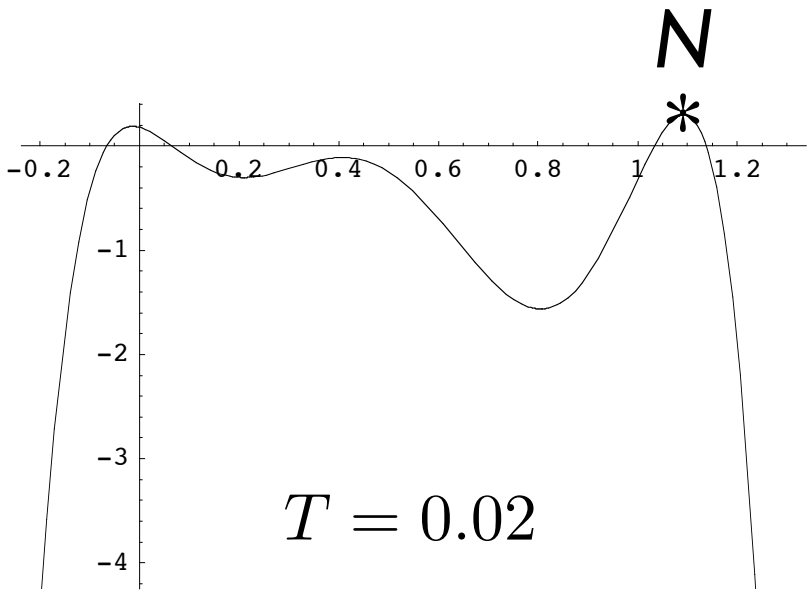




Denatured state

Native





Conserved energy

$$E = \frac{\gamma}{4}\dot{x}^2 - V_{eff}(x)$$

- Solution: go from time-dependent Newtonian dynamics to energy-dependent Hamilton-Jacobi description

$$S_{eff}[x] = \int_{t_i}^{t_f} dt(\frac{\gamma}{4}\dot{x}^2 + V_{eff}[x(t)])$$

- For classical trajectories

$$E_{eff} = \frac{\gamma}{4}\dot{x}^2 - V_{eff}[x]$$

$$S_{eff}[x] = -E_{eff}(t_f - t_i) + \int_{x_i}^{x_f} dx\sqrt{\frac{4}{\gamma}(E_{eff} + V_{eff}[x])}$$

- The method: minimize the Hamilton-Jacobi action

$$S_{HJ} = \int_{x_i}^{x_f} dl \sqrt{2\left(E_{eff} + V_{eff}[x(l)]\right)}$$

- over all paths joining x_i to x_f

dl is an infinitesimal displacement along the path

E_{eff} is a free parameter

- The total time is determined by

$$t_f - t_i = \int_{x_i}^{x_f} dl \sqrt{\frac{1}{2\left(E_{eff} + V_{eff}[x(l)]\right)}}.$$

determine folding time

- The total time is determined by the trajectory and by the energy E_{eff}

  - E_{eff} is not the true energy of the system

$$E = \frac{\gamma}{4}\dot{x}^2 - V_{eff}(x)$$

  - If the final state is an (almost) equilibrium state, then the system should spend a maximum time

$$\dot{x}_f = 0$$

$$E_{eff} = -V_{eff}(x_f)$$

$$= \frac{D}{2k_BT}U''(x_f)$$

- The HJ method is much more efficient than Newtonian mechanics because proteins spend most of their time trying to overcome energy barriers.

- No waiting-times in HJ: work with fixed interval length dl

- For a Protein, minimize

$$S_{HJ} = \sum_{n}^{N-1} \sqrt{2\left(E_{eff} + V_{eff}(n)\right)} \Delta l_{n,n+1} + \lambda P$$

- where $P = \sum_{i}^{N-1} (\Delta l_{i,i+1} - \langle \Delta l \rangle)^2$ and λ is a Lagrange multiplier to fix the interval length

$$\begin{aligned} V_{eff}(n) &= \sum_{i} \left[\frac{D^2}{2(k_B T)^2} \left(\sum_{j} \nabla_j u(\mathbf{x}_i(n), \mathbf{x}_j(n)) \right)^2 \right. \\ &- \left. \frac{D^2}{k_B T} \sum_{j} \nabla_j^2 u(\mathbf{x}_i(n), \mathbf{x}_j(n)) \right] \\ (\Delta l)^2_{n,n+1} &= \sum_{i} (\mathbf{x}_i(n+1) - \mathbf{x}_i(n))^2, \end{aligned}$$

$$E_{eff} = \frac{D}{2k_BT}\sum_i \vec{\nabla}_i^2 H(\vec{r}_1^{(n)},\ldots,\vec{r}_N^{(n)})$$

$$= \frac{D}{2k_BT}\mathrm{Tr}\ \mathcal{H}^{(n)}$$

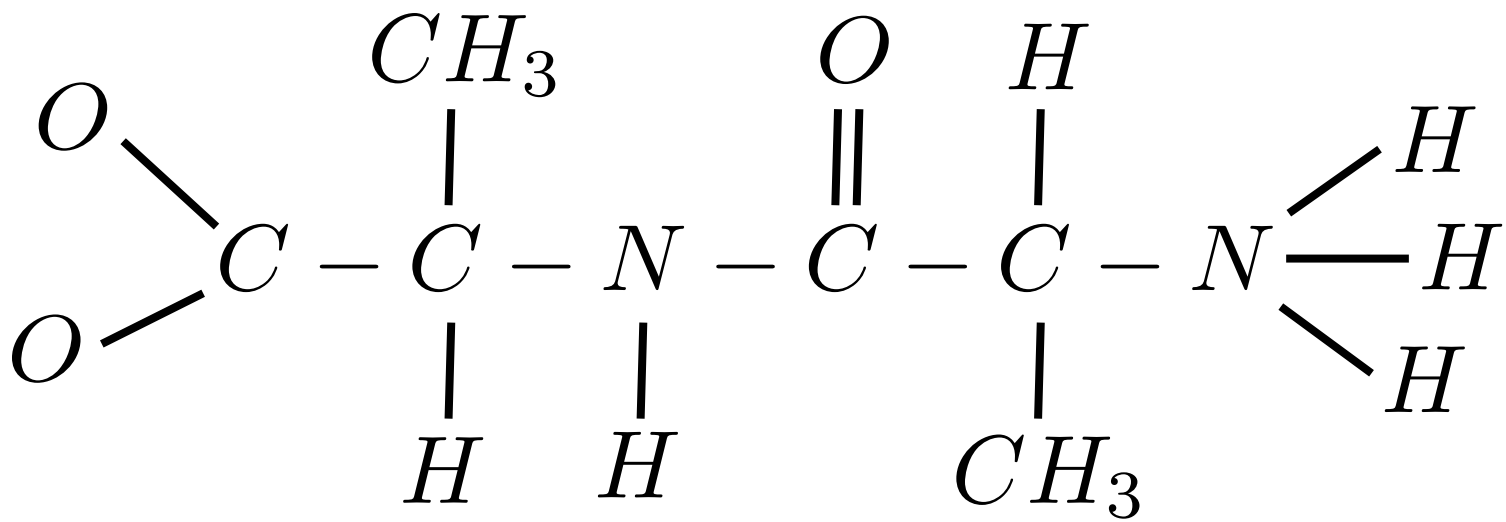
Hessian

- The energy can be evaluated by normal mode analysis or short time MD runs
- The Transition State defined by Commitment Analysis

$$P(x_{ts} \to x_i) = P(x_{ts} \to x_f)$$

$$\frac{U(x_f) - U(x_i)}{2k_BT} = S_{HJ}([x];x_{ts},x_i) - S_{HJ}([x];x_{ts},x_f).$$

Folding of Alanine-dipeptide



Alanine Dipeptide

- Use GROMOS96 force field
- There are four local minima $\mathrm{C7}_{ax} \rightarrow \mathrm{C7}_{eq}$ and

  $\alpha_L \rightarrow \alpha_R$
- The effective energy is computed by few ps MD runs

- Transition states can be obtained by commitment analysis

$$P(x_i, x_{ts}) = P(x_f, x_{ts})$$

which in the saddle-point approximation become

$$\frac{U(x_f) - U(x_i)}{2k_B T} = S_{HJ}([x]; x_{ts}, x_i) - S_{HJ}([x]; x_{ts}, x_f)$$

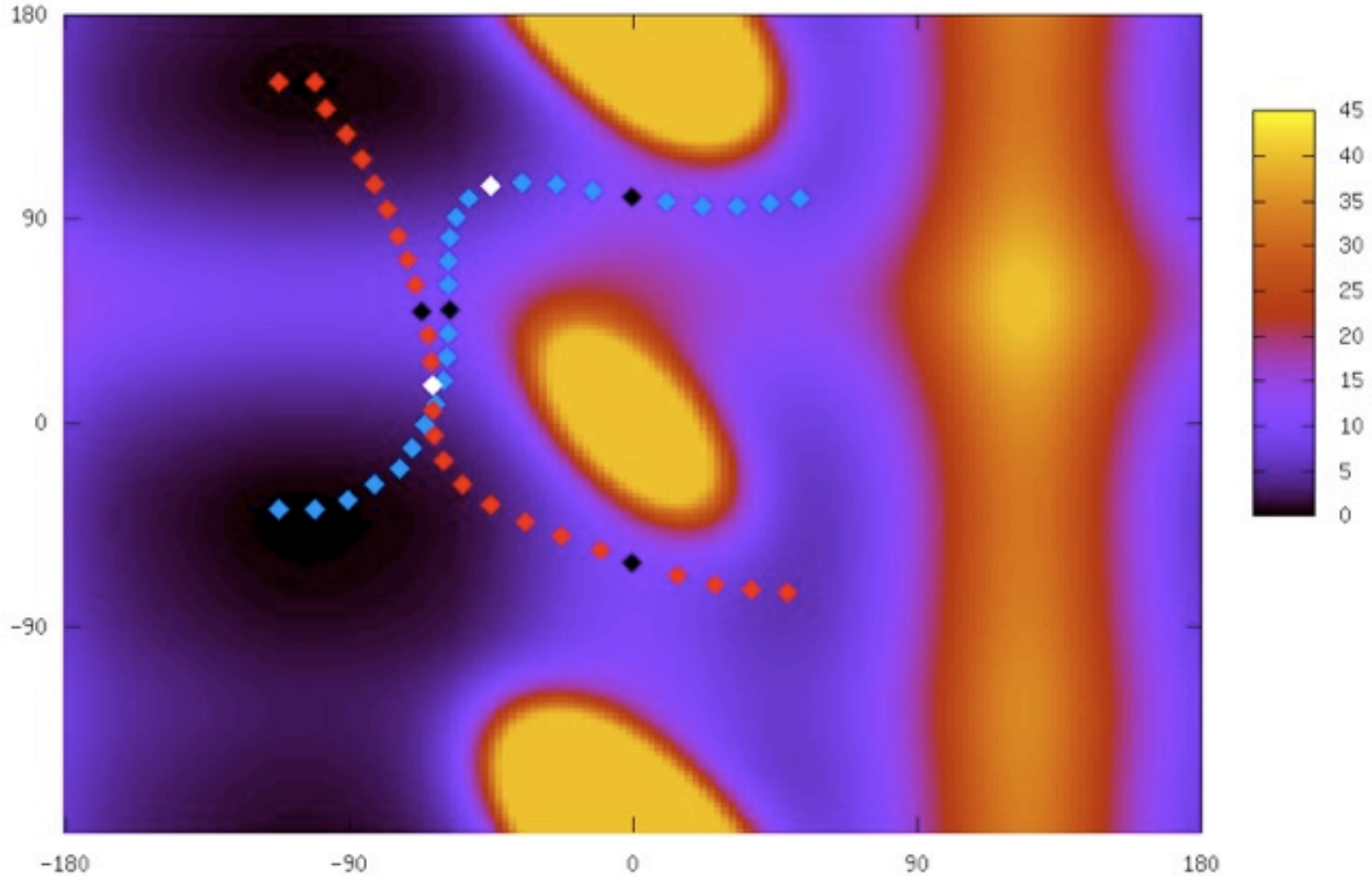


FIG. 1: Dominant Folding Paths for the C7$_{ax}$ →C7$_{eq}$ (red squares) and $\alpha_L \rightarrow \alpha_R$ (blue squares) transitions. In the background, the free energy profile for the ψ and ϕ dihedrals is shown (in units of kJ/mol). Black squares identify the minimum residence time conformations, and the white squares the transition states defined by comittement analysis.

Difficulties with the Method

- Many local minima, particularly with all atom simulations: many routes to folding?
- Optimisation of HJ stuck in the vicinity of initial trajectory
- How to overcome these difficulties?

Langevin Bridges

- Consider paths starting at $(x_0, 0)$ and conditioned to end at (x_f, t_f)
- The conditional probability for such a path to be at (x, t) is given by

$$\mathcal{P}(x,t) = \frac{1}{P(x_f, t_f | x_0, 0)} Q(x,t) P(x,t)$$

FP equation

adjoint FP equation

$$P(x,t) = P(x, t | x_0, 0)$$

$$Q(x,t) = P(x_f, t_f | x, t)$$

Fokker-Planck and adjoint

FP

$$\frac{\partial P}{\partial t} = D\frac{\partial}{\partial x}\left(\frac{\partial P}{\partial x} + \beta\frac{\partial U}{\partial x}P\right)$$

adjoint FP

$$\frac{\partial Q}{\partial t} = -D\frac{\partial^2 Q}{\partial x^2} + D\beta\frac{\partial U}{\partial x}\frac{\partial Q}{\partial x}$$

conditional probability

$$\frac{\partial \mathcal{P}}{\partial t} = D\frac{\partial}{\partial x}\left(\frac{\partial \mathcal{P}}{\partial x} + \frac{\partial}{\partial x}\left(\beta U - 2\ln Q\right)\mathcal{P}\right)$$

- Modified Langevin equation for conditioned paths

$$\frac{dx}{dt} = -\frac{D}{k_B T}\frac{\partial U}{\partial x} + 2D\frac{\partial \ln Q}{\partial x} + \eta(t)$$

$$Q(x,t) = P(x_f, t_f | x, t)$$

$$Q(x,t) = e^{-\frac{\beta}{2}(U(x_f)-U(x))} < x_f | e^{-(t_f - t)H} | x >$$

$$\frac{dx}{dt} = 2\frac{k_B T}{\gamma}\frac{\partial}{\partial x} \ln < x_f | e^{-(t_f - t)H} | x > + \eta(t)$$

$$\frac{dx}{dt} = < \dot{x}(t) > + \eta(t)$$

- Example: Brownian bridges

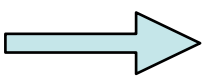
$$U(x) = 0$$

$$P(x_f, t_f | x, t) = \sqrt{\frac{1}{4\pi D(t_f - t)}} e^{-\frac{(x_f - x)^2}{4D(t_f - t)}}$$

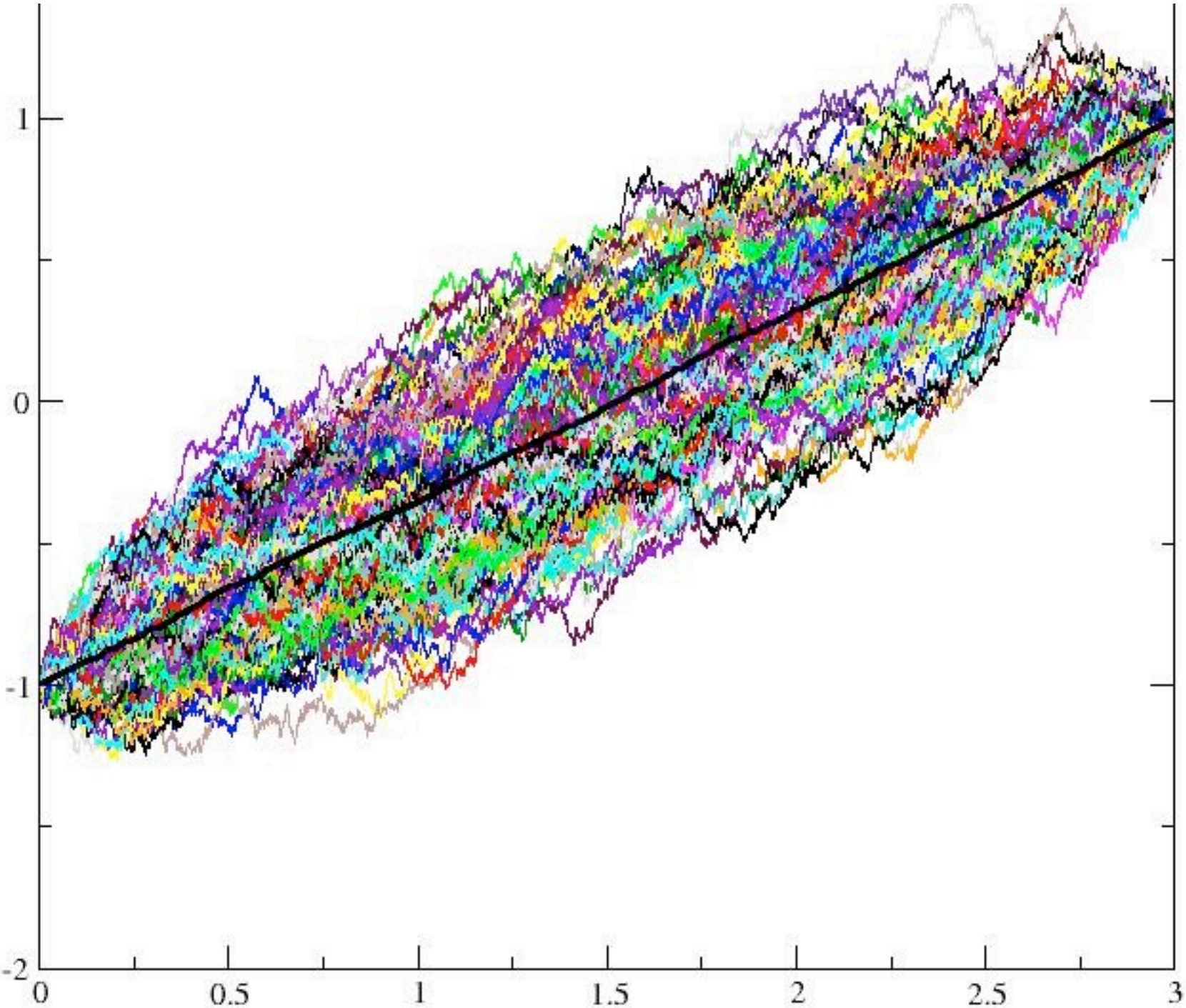
- Conditioned Langevin equation becomes

$$\frac{dx}{dt} = \frac{x_f - x}{t_f - t} + \eta(t)$$

$$\frac{dX}{dt} = \frac{x_f - X}{t_f - t}$$



average is linear in time


1
0
-1
-2
0
0.5
1
1.5
2
2.5
3

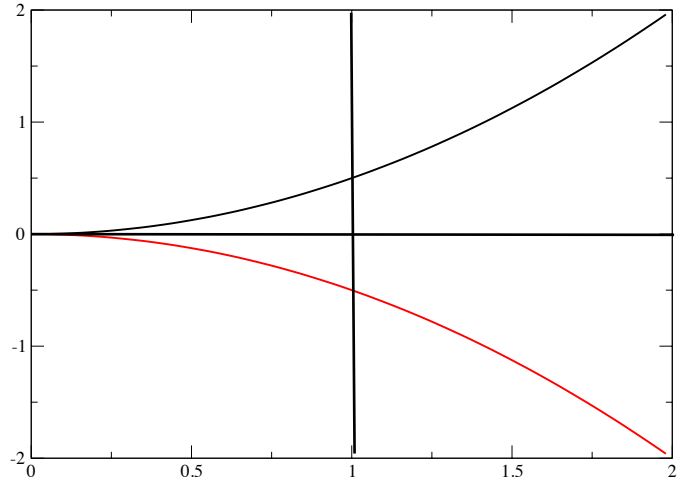
- Example: the Harmonic oscillator

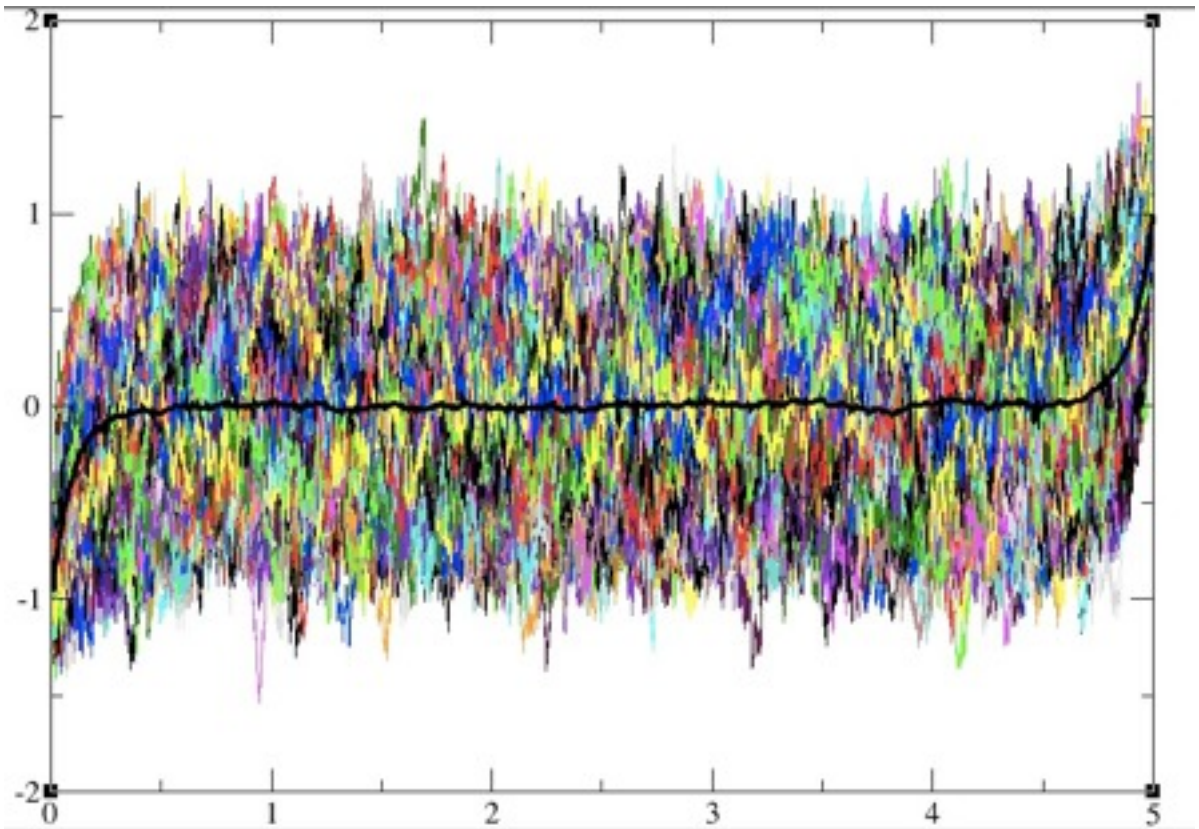
$$U(x) = \frac{1}{2} K x^2$$

- Bridge equation

$$\frac{dx}{dt} = \frac{K}{\gamma} \frac{x_f - x \cosh \frac{K}{\gamma}(t_f - t)}{\sinh \frac{K}{\gamma}(t_f - t)} + \eta(t)$$

- Note that this equation does not depend on the sign of *K*: same trajectories for well or barrier





x=-2.

500 trajectories between -1 and +1

- In general, we don't know how to calculate the function *Q(x,t)*. We need to make approximations.
- Some important requirements:
  - *Q(x,t)>0*
  - Detailed balance: $Q(x,t) = P(x_f, t_f | x, t)$ and we should have

$$\frac{P(x_f, t_f | x, t)}{P(x, t_f | x_f, t)} = e^{-\frac{U(x_f) - U(x)}{k_B T}}$$

  - Local in time and space (for simplicity and tractability)

- Langevin equation for conditioned paths

$$\frac{dx}{dt} = -\frac{D}{k_B T}\frac{\partial U}{\partial x} + 2D\frac{\partial \ln Q}{\partial x} + \eta(t)$$

$$Q(x,t) = P(x_f, t_f | x, t)$$

$$Q(x,t) = e^{-\frac{\beta}{2}(U(x_f)-U(x))} < x_f | e^{-(t_f - t)H} | x >$$

$$\frac{dx}{dt} = 2\frac{k_B T}{\gamma}\frac{\partial}{\partial x} \ln < x_f | e^{-(t_f - t)H} | x > + \eta(t)$$

Short transition path time approximation

- In the Kramers picture, there are 2 time scales:
  - Kramers time, or waiting time, or folding time

$$\tau_K \approx e^{\frac{\Delta E}{k_B T}}$$

  - Transition path time (Hummer, Szabo)

$$\tau_{TP} \approx \log \frac{\Delta E}{k_B T}$$

  - We will assume

$$\tau_{TP} << \tau_K$$

$$Q(x,t) = e^{-\frac{U(x_f)-U(x)}{2k_BT}} < x_f|e^{-H(t_f-t)}|x >$$

For short times, use the Trotter formula (Baker, Campbell, Haussdorf). To satisfy detailed balance, use symmetric form

$$e^{-\varepsilon(H_0+V_1)} = e^{-\varepsilon V_1/2} e^{-\varepsilon H_0} e^{-\varepsilon V_1/2} + O(\varepsilon^3)$$

$$I(x,t) =< x_f|e^{-H(t_f-t)}|x >$$

$$= e^{-(t_f-t)(V_1(x_f)+V_1(x))/2} < x_f|e^{-(t_f-t)H_0}|x >$$

$$< x_f|e^{-H_0(t_f-t)}|x > = \sqrt{\frac{1}{4\pi D(t_f-t)}}\, e^{-\frac{(x_f-x)^2}{4D(t_f-t)}}$$

- For short enough time, putting all the terms together we obtain the (approximate) Langevin bridge equation

$$\frac{d\vec{x}}{dt} = \frac{\vec{x}_f - \vec{x}}{t_f - t} - \frac{1}{4\gamma^2}(t_f - t)\nabla V(\vec{x}) + \vec{\eta}(t)$$

$$V(\vec{x}) = (\nabla U)^2 - 2k_B T \nabla^2 U$$

This equation is to be integrated with initial condition x_i
Equation is local.

- Works very well for “short times”
- For longer times, need to reweight the paths

$$w(\{x(t)\}) = \exp\left(-\frac{\gamma}{4k_BT}\int_0^{t_f} dt\left(\left(\frac{d\vec{x}}{dt}+\frac{1}{\gamma}\nabla U\right)^2-\left(\frac{d\vec{x}}{dt}-\frac{\vec{x}_f-\vec{x}}{t_f-t}+\frac{t_f-t}{4\gamma^2}\nabla V(\vec{x})\right)^2\right)\right)$$

true weight

actual weight

- then for any observable

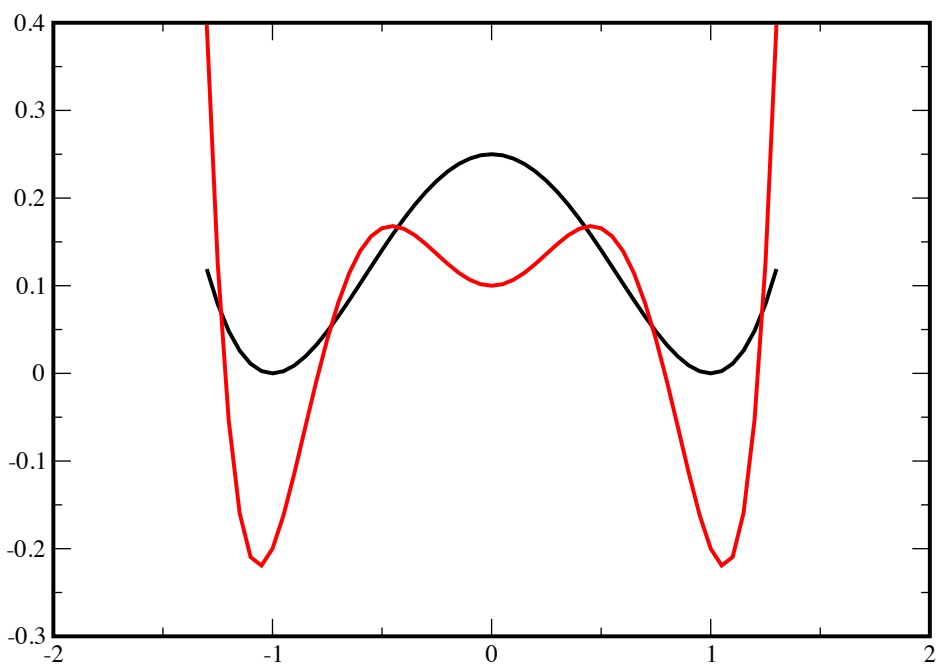
$$< A >= \sum_{\{x(t)\}} w(\{x(t)\})A(\{x(t)\})$$

Example: Quartic double well

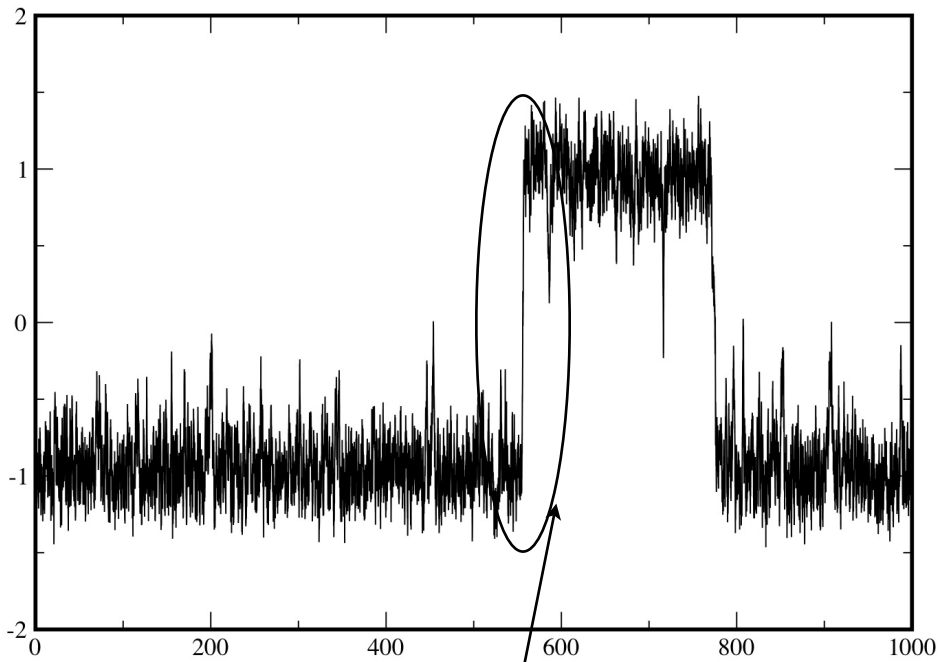
- We take

$$U(x) = \frac{1}{4}(x^2 - 1)^2$$

$$V(x) = \frac{1}{4k_BT}(U'^2(x) - 2k_BTU''(x))$$

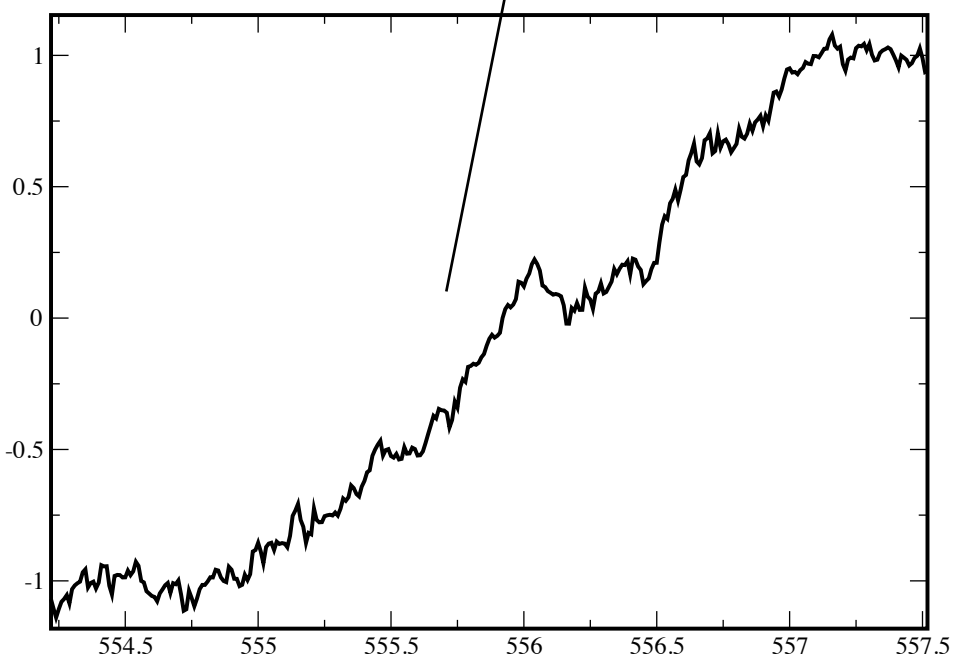


$T = 0.05$

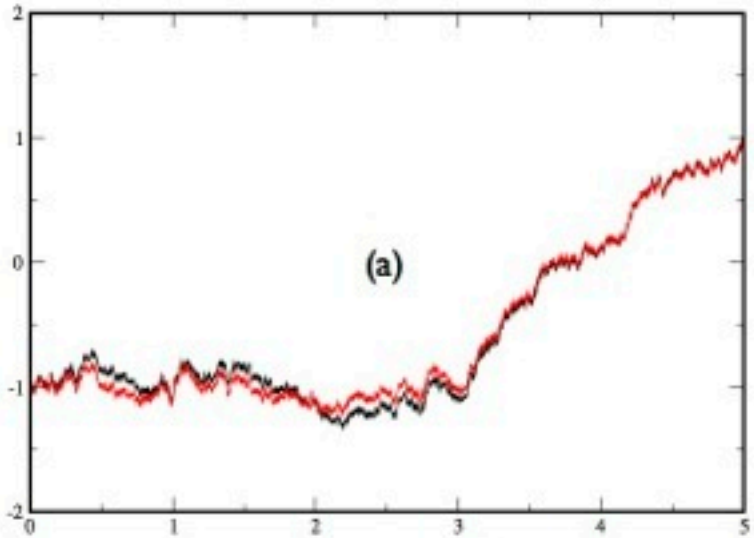


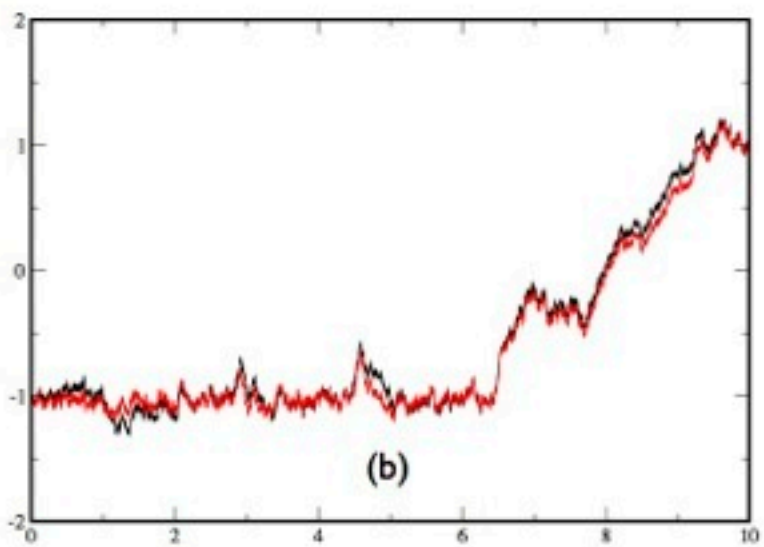
Langevin

transition region



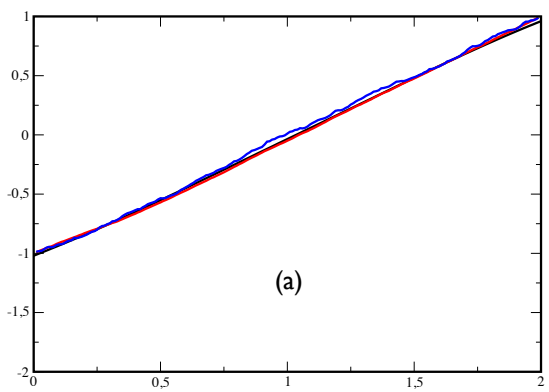
Trajectories



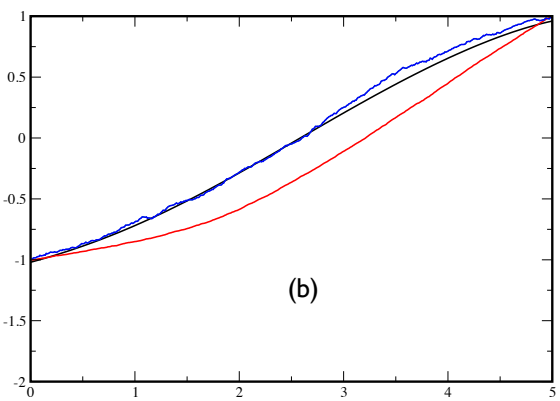


Averages and Observables

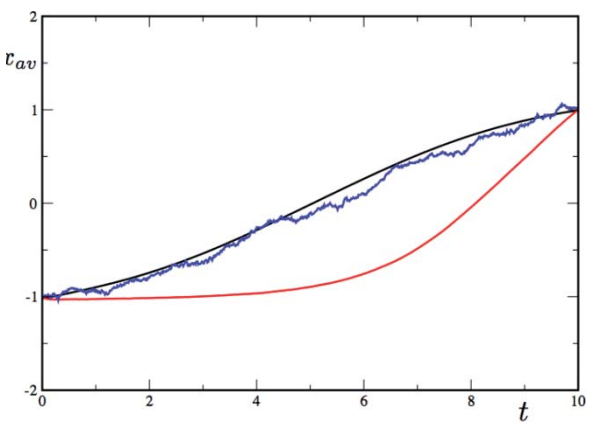
- Average trajectory: exact (black), approximate (red), reweighted (blue)



$t_f = 2$



$t_f = 5$

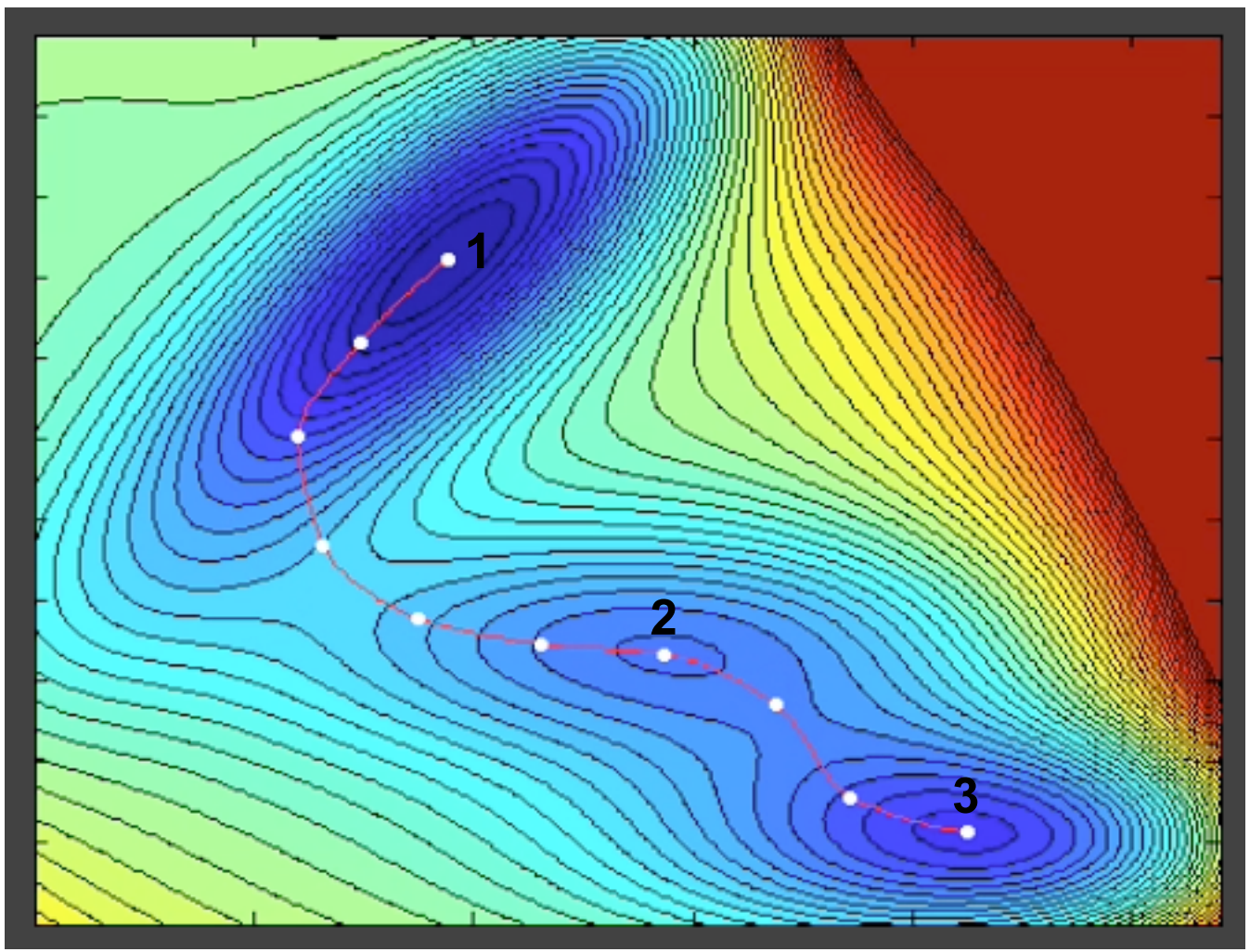


$t_f = 10$

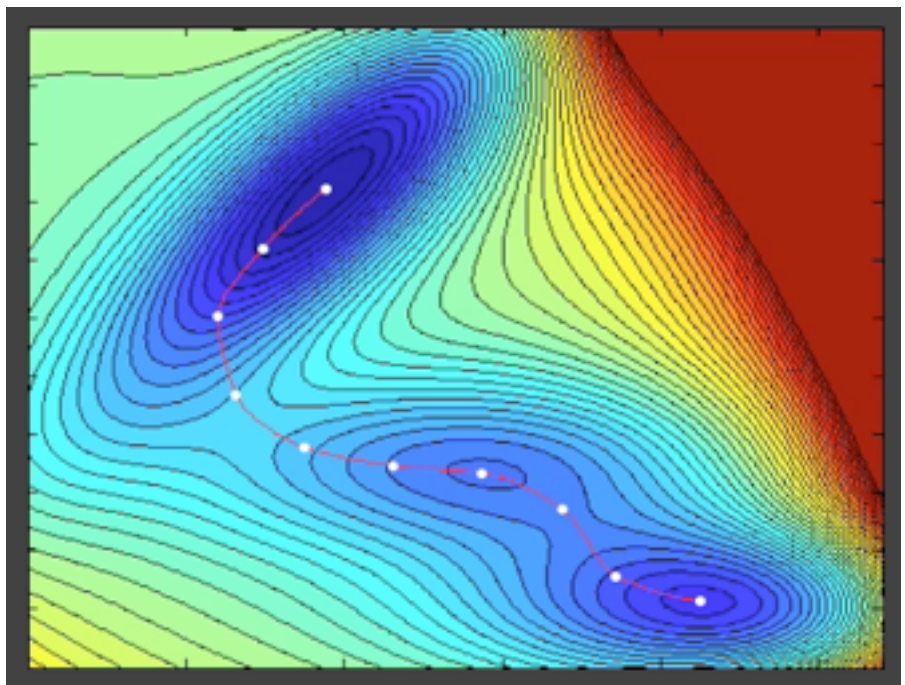
The Mueller potential

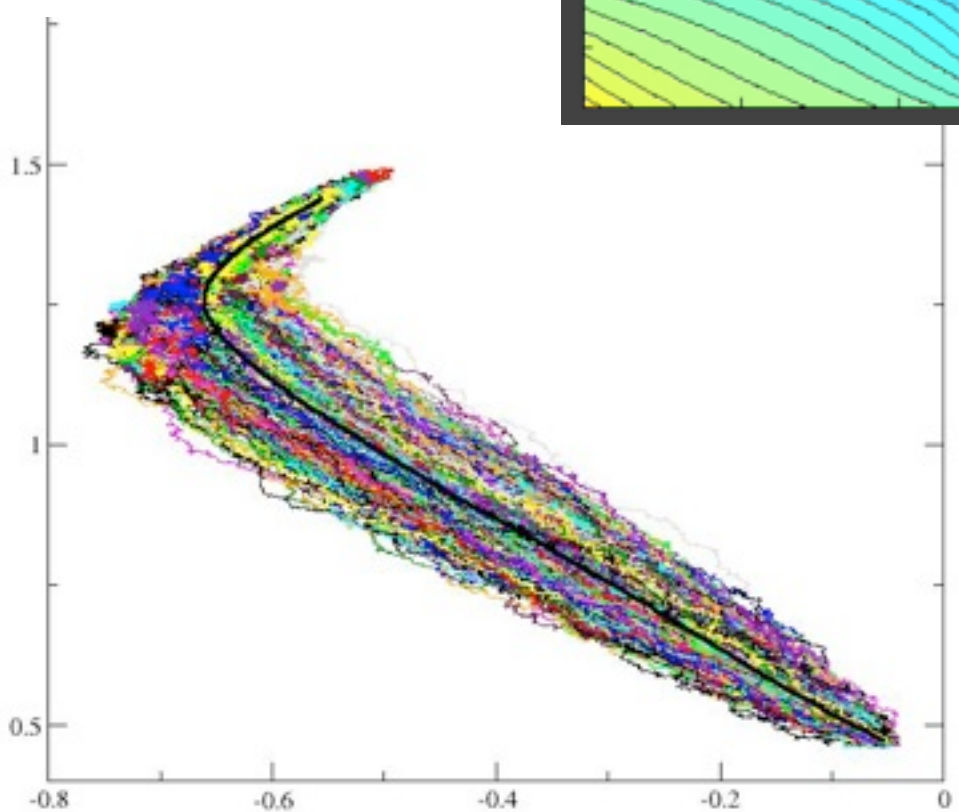
$$V(x,y) = \sum_{i=1}^{4} A_i \exp\left[a_i(x-x_i^0)^2 + b_i(x-x_i^0)(y-y_i^0) + c_i(y-y_i^0)^2\right]$$

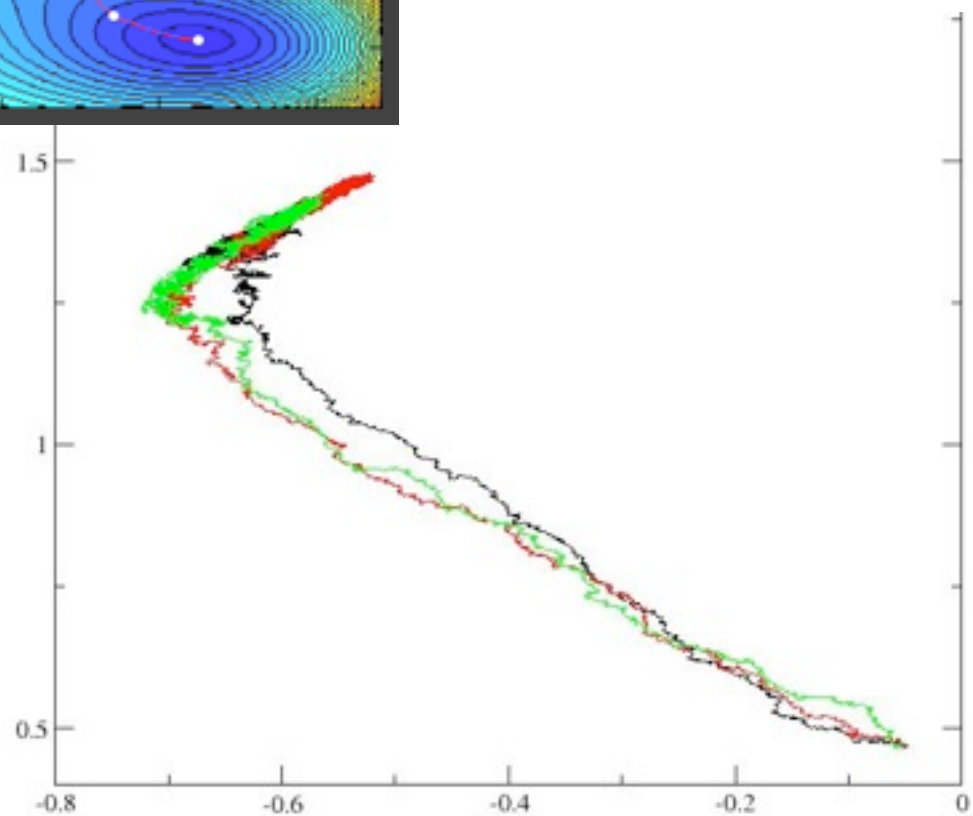
where $A = (-200, -100, -170, 15)$, $a = (-1, -1, -6.5, 0.7)$, $b = (0, 0, 11, 0.6)$,



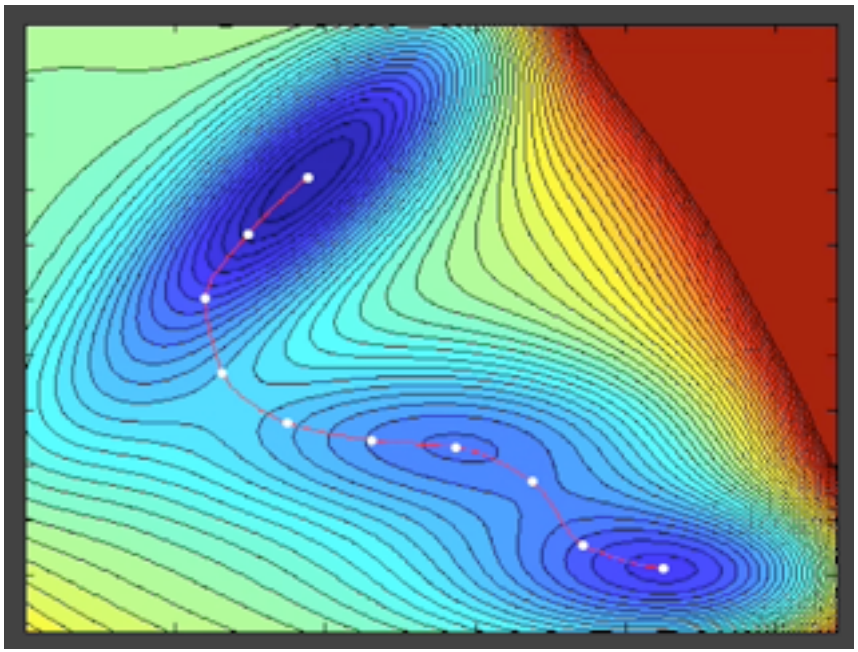
Transition 1-2

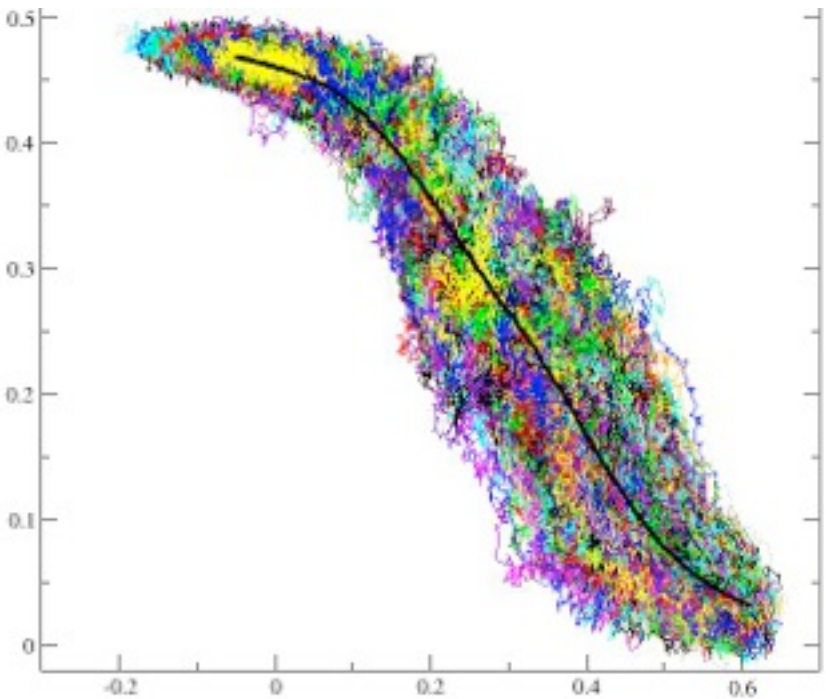


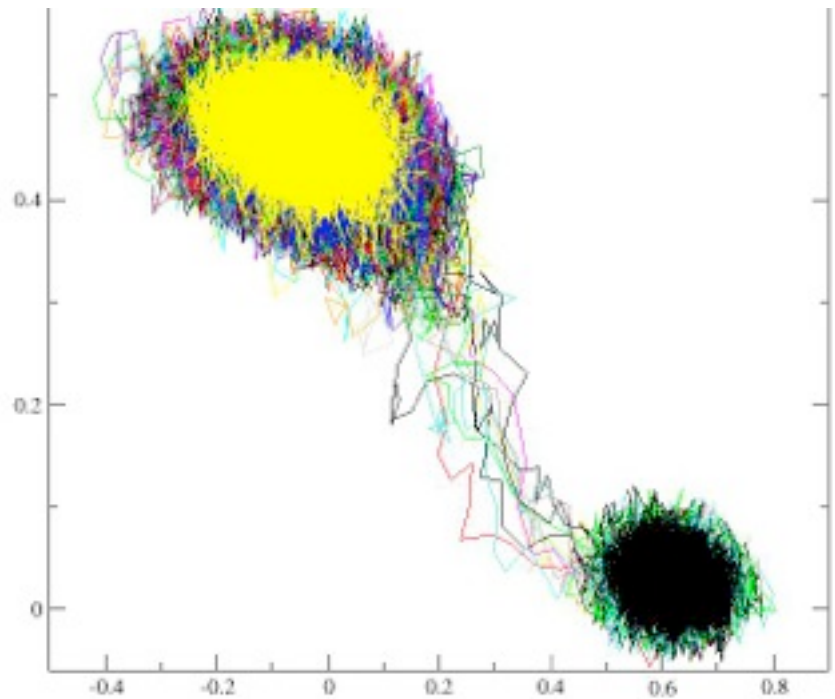




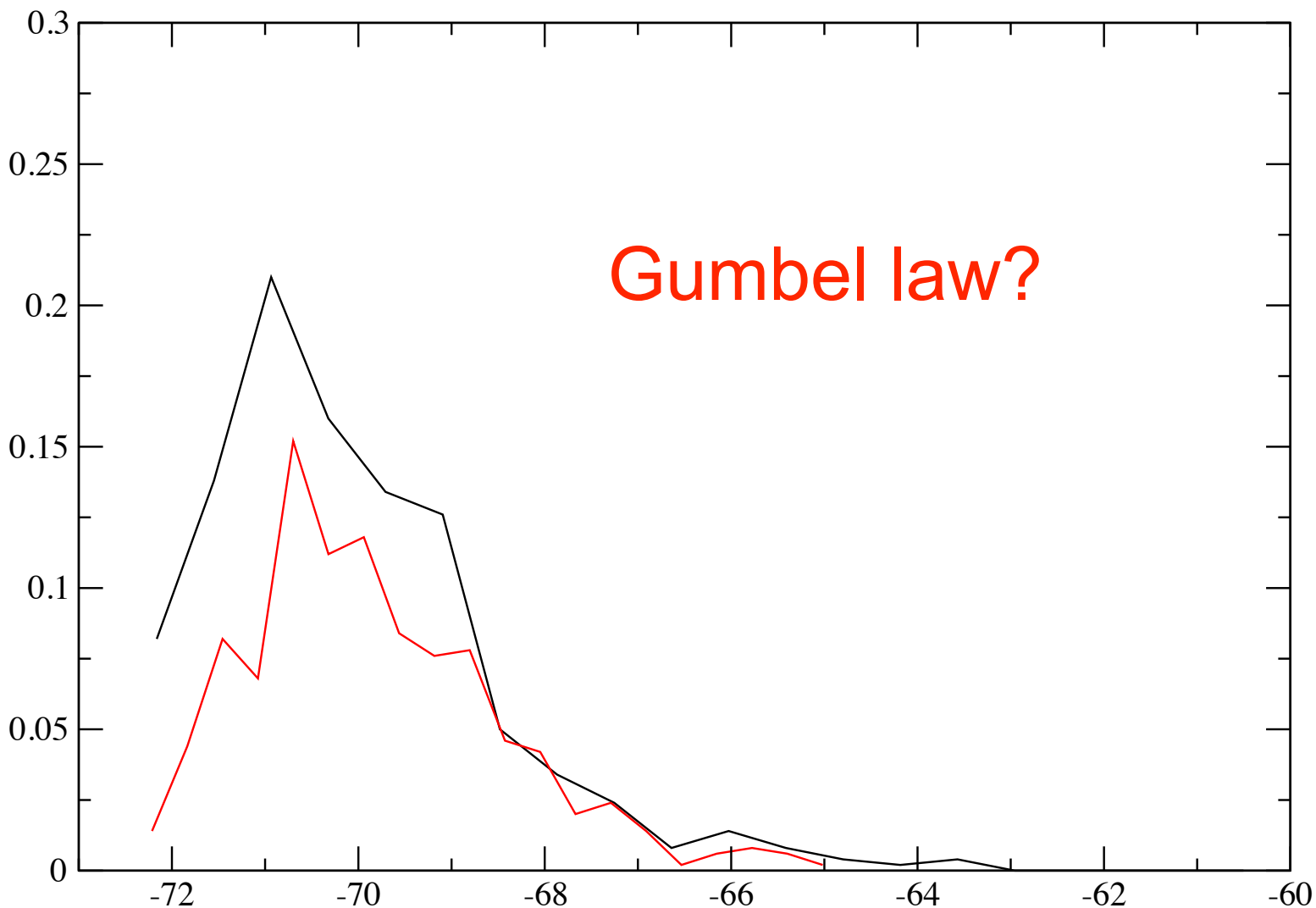
Transition 2-3



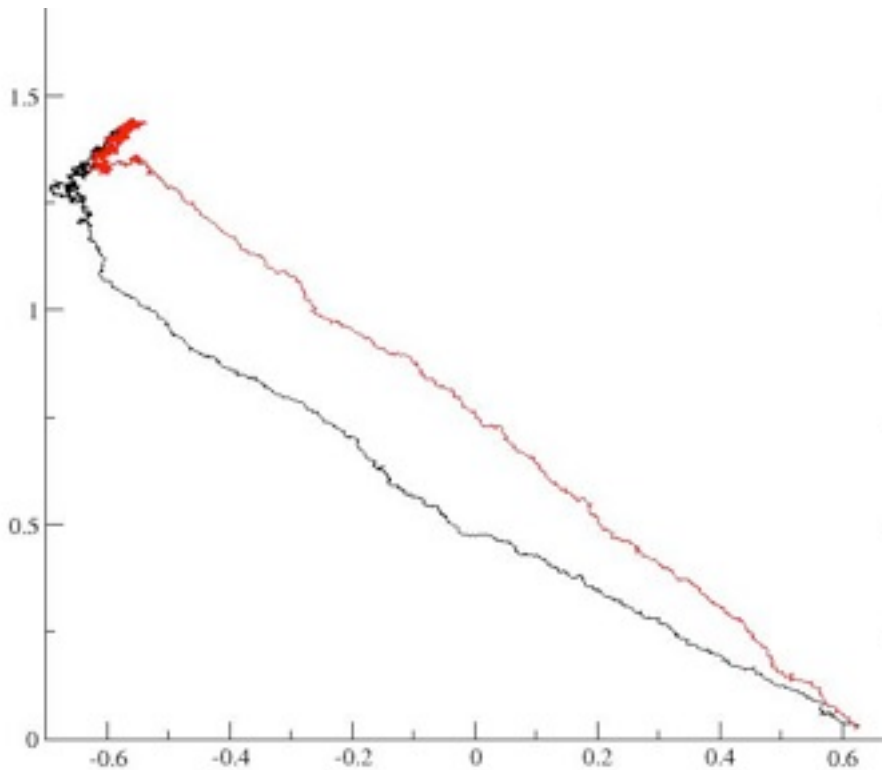


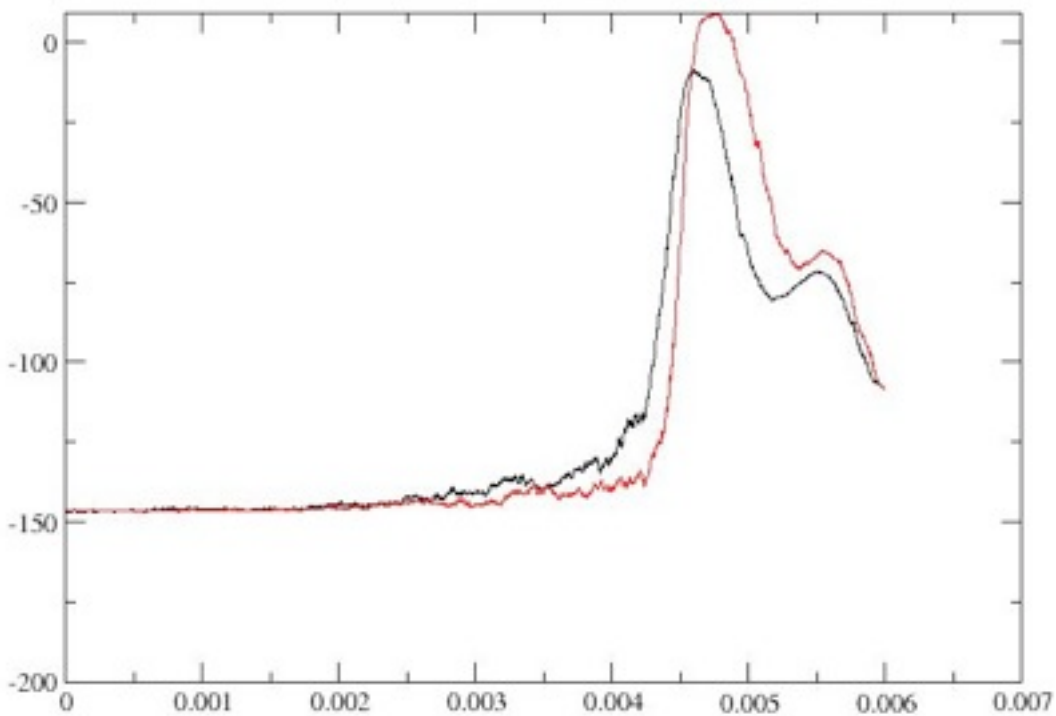


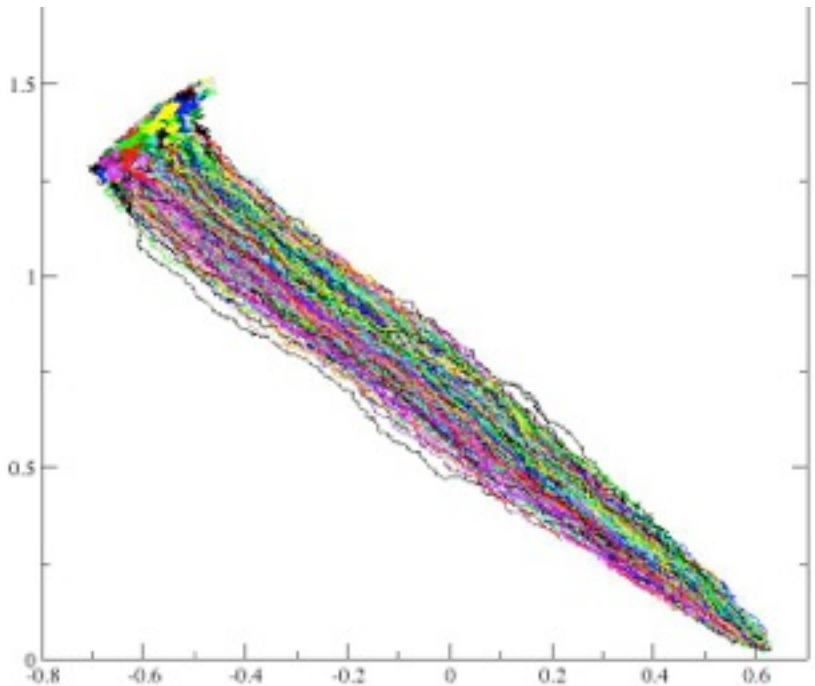
Histogram of barrier heights

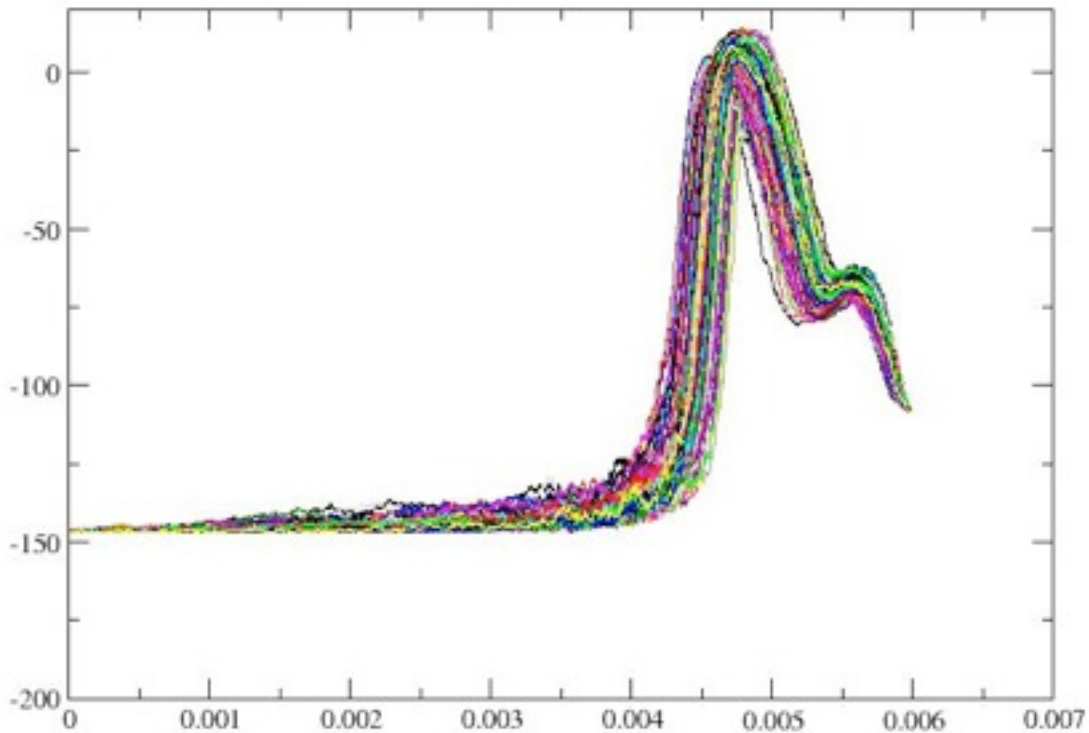


Transition 1-3






1.5
1
0.5
0
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6


0
-50
-100
-150
-200
0
0.001
0.002
0.003
0.004
0.005
0.006
0.007

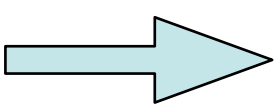
Work in progress: exact numerical solution of Langevin Bridge

- Need to solve the 2 equations

$$\frac{dx}{dt} = -\frac{D}{k_B T}\frac{\partial U}{\partial x} + 2D\frac{\partial \ln Q}{\partial x} + \eta(t)$$

$$\frac{\partial Q}{\partial t} = -D\frac{\partial^2 Q}{\partial x^2} + D\beta\frac{\partial U}{\partial x}\frac{\partial Q}{\partial x}$$

- but need $Q(x,t)$ only along trajectories



equation for $Q(x(t),t)$

$$x_{k+1} = x_k - D\beta dt \frac{\partial U(x_k)}{\partial x_k} + 2Ddt \frac{\partial \log Q(x_k, k)}{\partial x_k} + \sqrt{2Ddt}\zeta_k$$

$$\begin{aligned} Q(x_{k+1}, k+1) = \quad & Q(x_k, k) + (x_{k+1} - x_k)\frac{\partial Q(x_k, k)}{\partial x_k} + \frac{1}{2}(x_{k+1} - x_k)^2 \frac{\partial^2 Q(x_k, k)}{\partial x_k^2} \\ - \quad & Ddt \frac{\partial^2 Q(x_k, k)}{\partial x_k^2} + D\beta dt \frac{\partial U(x_k)}{\partial x_k} \frac{\partial Q(x_k, k)}{\partial x_k} \end{aligned}$$

$$< \zeta_k >= 0$$

$$< \zeta_k \zeta_k >= \delta_{kl}$$

So if we know a path up to time *k*, we can increment *x* and then *Q*.

To compute the derivatives of *Q*, we need to grow a family of many paths in parallel, and look for points close enough to compute derivatives.

There remains some difficulties (instabilities)

Other numerical approach

- Equation to solve:

$$\frac{dx}{dt} = -\frac{D}{k_B T}\frac{\partial U}{\partial x} + 2D\frac{\partial \ln Q}{\partial x} + \eta(t)$$

- Start with $Q_0(x,t)$
- Generate *M* trajectories

$$x_\alpha^{(0)}(t),\ \{\alpha = 1, ..., M\}$$

- From these trajectories, generate $Q_1(x,t)$

- Iterate procedure

Conclusion

- No need of reaction coordinates
- Method is efficient, and fast : completely parallelizable
- All trajectories are statistically independent
- Possibility to reweight the trajectories
- Possibility to include the solvent
- Can be generalized to discrete systems.