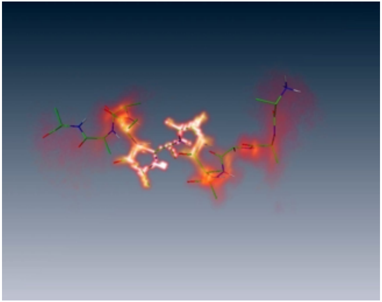
Importance sampling of rare events using cross-entropy minimization

Carsten Hartmann (BTU Cottbus-Senftenberg),

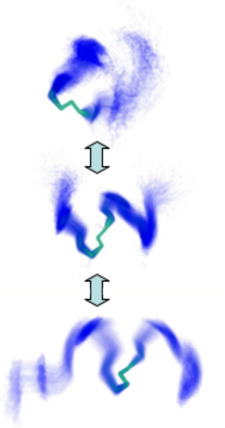
jointly with Christof Schütte (ZIB) & Wei Zhang (FU Berlin),
and Omar Kebiri (U Tlemcen)

Institut Henri Poincaré, Paris, 26/04/2017

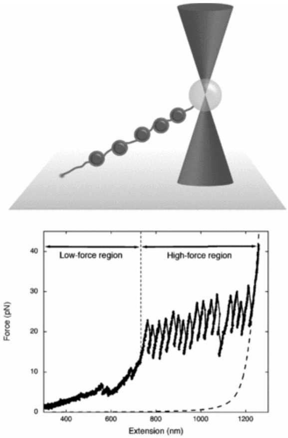
Motivation: conformation dynamics of biomolecules



1.3μs MD simulation of *dodeca-alanin* at $T = 300K$
(GROMOS96, visualization: Amira@ZIB)



Motivation: single molecule experiments



- Probing of equilibrium properties by **nonequilibrium experiments**:

$$F = -\log \mathbb{E}\left[e^{-W}\right].$$

(includes rates, statistical weights, etc.)

- Perturbation drives the system out of equilibrium with **likelihood quotient**

$$\varphi = \frac{d\mu_0}{d\mu}.$$

- Experimental and numerical realization: AFM, SMD, TMD, Metadynamics, ...

[Schlitter, J Mol Graph, 1994], [Schulten & Park, JCP, 2004], [H. et al, Proc Comput Sci, 2010]

Set-up: estimation problem

Given an "equilibrium" diffusion process $X = (X_t)_{t\geq 0}$ on $\mathbb{R}^n$,

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t\,, \quad X_0 = x\,,$$

we want to **estimate path functionals** of the form

$$\psi(x) = \mathbb{E}\left[e^{-W(X)}\right]$$

Example: mean passage time to a set $C \subset \mathbb{R}^n$

Let $W = \alpha\tau_C$. Then, for sufficiently small $\alpha > 0$,

$$-\alpha^{-1}\log\psi = \mathbb{E}[\tau_C] + \mathcal{O}(\alpha)$$

Guiding example: bistable system

- Overdamped Langevin equation

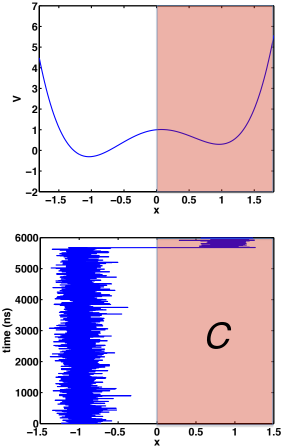
$$dX_t = -\nabla V(X_t)dt + \sqrt{2\epsilon}dB_t \,.$$

- Standard estimator of MGF $\psi = \psi_\epsilon$

$$\hat{\psi}_\epsilon^N = \frac{1}{N}\sum_{i=1}^{N} e^{-\alpha\tau_C^i} \,.$$

- Small noise asymptotics (Kramers)

$$\lim_{\epsilon\to 0} \epsilon \log \mathbb{E}[\tau_C] = \Delta V \,.$$



[Freidlin & Wentzell, 1984], [Berglund, Markov Processes Relat Fields 2013]

Guiding example, cont'd

- **Relative error** of the MC estimator

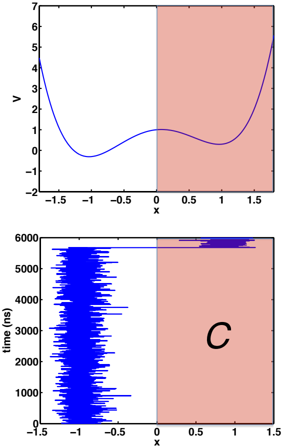
$$\delta_\epsilon = \frac{\sqrt{\mathrm{Var}[\hat{\psi}_N^\epsilon]}}{\mathbb{E}\left[\hat{\psi}_N^\epsilon\right]}$$

- Varadhan's large deviations principle

$$\mathbb{E}\left[(\hat{\psi}_\epsilon^N)^2\right] \gg \left(\mathbb{E}\left[\hat{\psi}_\epsilon^N\right]\right)^2, \ \epsilon \text{ small.}$$

- Unbounded relative error as $\epsilon \to 0$

$$\limsup_{\epsilon \to 0} \delta_\epsilon = \infty$$



[Asmussen et al, Encyclopedia of Operations Research and Management Sciences, 2012]

Outline

Adaptive importance sampling of rare events

Optimal controls from cross-entropy minimization

Numerical strategies for high-dimensional problems

Adaptive importance sampling of rare events

Optimal controls from cross-entropy minimization

Numerical strategies for high-dimensional problems

Guiding example, cont'd

- Mean first passage time for small ϵ

$$\mathbb{E}[\tau_C] \asymp \exp(\Delta V/\epsilon)$$

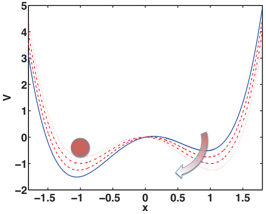
- **Adaptive tilting** of the potential

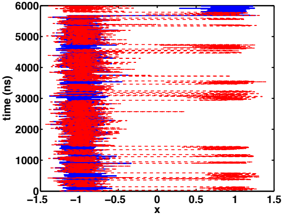
$$U(x,t) = V(x) - u_t x$$

  decreases the energy barrier.

- Controlled Langevin equation

$$dX_t^u = (u_t - \nabla V(X_t^u))\, dt + \sqrt{2\epsilon} dB_t\, .$$





Estimation problem revisited

Given a "nonequilibrium" (tilted) diffusion process $X^u = (X_t^u)_{t\geq 0}$,

$$dX_t^u = (b(X_t^u) + \sigma(X_t^u)u_t)dt + \sigma(X_t^u)dB_t\,, \quad X_0^u = x\,,$$

estimate a **reweigthed version** of ψ:

$$\mathbb{E}\left[e^{-W(X)}\right] = \mathbb{E}^{\mu}\left[e^{-W(X^u)}\varphi(X^u)\right]$$

with equilibrium/nonequilibrium likelihood ratio $\varphi = \frac{d\mu_0}{d\mu}$.

Remark: We allow for W's of the general form

$$W(X) = \int_0^\tau f(X_s, s)\,ds + g(X_\tau)\,,$$

for suitable functions f, g and an a.s. finite stopping time $\tau < \infty$.

Can we systematically **speed up the sampling** while **controlling the variance** by tilting the energy landscape?

Variational characterization of free energies

Theorem (Donsker & Varadhan)

For any bounded and measurable function W it holds

$$-\log \mathbb{E}\left[e^{-W}\right] = \inf_{\mu \ll \mu_0} \left\{ \mathbb{E}^{\mu}[W] + KL(\mu, \mu_0) \right\}$$

where $KL(\mu, \mu_0) \geq 0$ is the KL divergence between μ and μ_0.

Sketch of proof: Let $\varphi = \frac{d\mu_0}{d\mu}$. Then

$$\begin{aligned} -\log \int e^{-W} d\mu_0 &= -\log \int e^{-W + \log \varphi} d\mu \\ &\leq \int (W - \log \varphi)\, d\mu \end{aligned}$$

with **equality iff** $W - \log \varphi$ **is constant** (μ-a.s.).

[Boué & Dupuis, LCDS Report #95-7, 1995], [Dai Pra et al, Math Control Signals Systems, 1996]

Same same, but different...

Variational characterization of free energies, cont'd

Theorem

Technical details aside, let u^* be a minimizer of the cost functional

$$J(u) = \mathbb{E}\left[W(X^u) + \frac{1}{4}\int_0^{\tau^u} |u_s|^2\, ds\right]$$

under the controlled dynamics

$$dX_t^u = (b(X_t^u) + \sigma(X_t^u)u_t)dt + \sigma(X_t^u)dB_t\,, \quad X_0^u = x\,.$$

The minimizer is unique with $J(u^*) = -\log\psi(x)$. Moreover,

$$\psi(x) = e^{-W(X^{u^*})}\varphi(X^{u^*}) \quad \text{(a.s.)}\,.$$

[H & Schütte, JSTAT, 2012], [H et al, Entropy, 2014]

Guiding example, cont'd

- Exit problem: $f = \alpha$, $g = 0$, $\tau = \tau_C$:

$$J(u^*) = \min_u \mathbb{E}\left[\alpha\tau_C^u + \frac{1}{4}\int_0^{\tau_C^u} |u_s|^2\, ds\right]$$

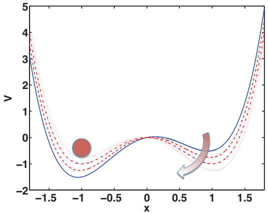
- Recovering **equilibrium statistics** by

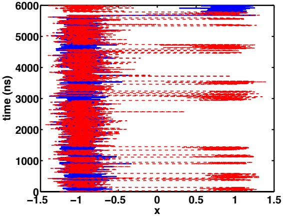
$$\mathbb{E}[\tau_C] = \frac{d}{d\alpha}\bigg|_{\alpha=0} J(u^*)$$

- Optimally tilted potential

$$U^*(x,t) = V(x) - u_t^*\, x$$

with stationary feedback $u_t^* = c(X_t^{u^*})$.





Some remarks . . .

Duality between estimation and control

The optimal control is a feedback control in gradient form ,

$$u_t^* = -2\sigma(X_t^{u^*})^T \nabla F(X_t^{u^*}, t)\,,$$

with the bias potential being the value function

$$F(x, t) = \min\{J(u) \colon X_t^u = x\}\,.$$

(In many interesting cases, $F = F(x)$ will be stationary.)

No-free-lunch theorem: The bias potential is given by

$$F = -\log\psi,$$

i.e., u^* depends on the quantity we want to estimate.

[H & Schütte, JSTAT, 2012], [H et al, Entropy, 2014]; cf. [Fleming, SIAM J Control, 1978]

More on the duality between estimation and control

The Legendre-type variational principle for the free energy furnishes an equivalence between the **dynamic programming equation**

$$-\frac{\partial F}{\partial t} + \min_{c \in \mathbb{R}^k} \left\{ LF + (\sigma c) \cdot \nabla F + \frac{1}{2}|c|^2 + f \right\} = 0 \; + \; b.c.$$

for F and the **Feynman-Kac formula** for $e^{-F} = \mathbb{E}[e^{-W}]$:

$$\left(\frac{\partial}{\partial t} - L \right) e^{-F} = 0 \,,$$

with L being the infinitesimal generator of $X_t^{u=0}$.

Related work on asymptotics (non-exhaustive)

- Exponential change of measure and large deviations statistics: [Siegmund, Ann. Stat., 1976], [Heidelberger, ACM Trans. Modeling Comp. Simulation, 1995], ...
(cf. also [Glasserman & Kou, Ann. Appl. Prob., 1997], [Glasserman & Wang, Ann. Appl. Prob., 1997])
- Adaptive IS based on HJB and Isaac equations: [Dupuis & Wang, Stochastics, 2004], [Dupuis & Wang, Math Oper Res, 2007], [Vanden-Eijnden & Weare, CPAM, 2012], ...
- Extensions to multiscale systems: [Spiliopoulos et al, Winter Simulation Conference, 2013], Spiliopoulos et al., SIAM MMS, 2012], [Zhang et al, Prob. Theor. Rel. F. 2017] ...
- For an overview see: [Asmussen & Glynn, Springer, 2007], [Rubinstein & Kroese, Wiley, 2007]

Adaptive importance sampling of rare events

Optimal controls from cross-entropy minimization

Numerical strategies for high-dimensional problems

Two key facts about our control problem

Fact #1

If $\sigma\sigma^T$ has a uniformly bounded inverse, then the optimal control can be represented as a **feedback law** of the form

$$u_t^* = \sigma(X_t^u) \sum_{i=1}^{\infty} c_i \nabla \phi_i(X_t^u, t) \,,$$

with coefficients $c_i \in \mathbb{R}$ and basis functions $\phi_i \in C^{1,0}(\mathbb{R}^n, [0, \infty))$.

Fact #2

Letting μ denote the probability (path) measure on $C([0,\infty))$ associated with the **tilted dynamics** X^u, it holds that

$$J(u) - J(u^*) = KL(\mu, \mu^*)$$

with $\mu^* = \mu(u^*)$ and

$$KL(\mu, \mu^*) = \begin{cases} \int \log\left(\frac{d\mu}{d\mu^*}\right) d\mu & \text{if } \mu \ll \mu^* \\ \infty & \text{otherwise} \end{cases}$$

the **Kullback-Leibler divergence** between μ and μ^*.

Cross-entropy method for diffusions

Idea: seek a minimizer of J among all controls of the form

$$\hat{u}_t = \sigma(X_i^u) \sum_{i=1}^{M} c_i \nabla \phi_i(X_t^u, t)\,, \quad \phi_i \in C^{1,0}(\mathbb{R}^n, [0, \infty))\,.$$

and minimize the Kullback-Leibler divergence

$$S(\mu) = KL(\mu, \mu^*)$$

over all candidate probability measures of the form $\mu = \mu(\hat{u})$.

Remark: unique minimizer is given by $d\mu^* = \psi^{-1} e^{-W} d\mu_0$.

Unfortunately, . . .

Cross-entropy method for diffusions, cont'd

...that doesn't work without knowing the normalization factor ψ.

Feasible cross-entropy minimization

Minimization of the auxiliary functional $KL(\mu^*, \cdot)$ is equivalent to **cross-entropy minimization**: minimize

$$CE(\mu) = -\int \log \mu \, d\mu^*$$

over all admissible $\mu = \mu(\hat{u})$, with $d\mu^* \propto e^{-W} d\mu_0$.

Note: $KL(\mu, \mu^*)=0$ iff $KL(\mu^*, \mu) = 0$, which holds iff $\mu = \mu^*$.

[Rubinstein & Kroese, Springer, 2004], [Zhang et al, SISC, 2014]

Some remarks

- The cross-entropy functional can be recast as

$$CE(\mu) = -\int (\log \mu(\hat{u})) e^{-W(X^{\hat{u}})} \varphi(\hat{u})\, d\mu(\hat{u})$$

  where both φ and μ (more precisely: its Wiener measure density) can be computed from **Girsanov's Theorem**.
- The **necessary optimality conditions** are of the form

$$Ac = \zeta$$

  with unknowns $c = (c_1, \ldots, c_M)$ and coefficients $A = (A_{ij})$, $\zeta = (\zeta_1, \ldots, \zeta_M)$ that are computable by Monte Carlo.
- In practice, annealing and clever choice of basis functions ϕ_i (e.g. global or local) greatly **enhances convergence**.

[H et al, Nonlinearity, 2016], [Badowski, PhD thesis, 2016], [Kappen et al, J Stat Phys, 2017]

Example I (guiding example)

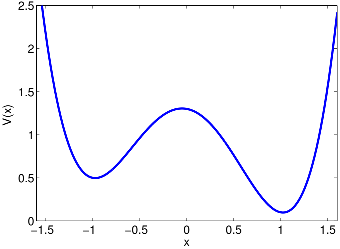
Computing the mean first passage time ($n = 1$)

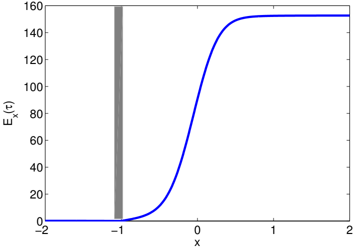
Minimize

$$J(u;\alpha) = \mathbb{E}\left[\alpha\tau + \frac{1}{4}\int_0^\tau |u_t|^2\,dt\right]$$

with $\tau = \inf\{t > 0\colon X_t \in [-1.1, -1]\}$ and the dynamics

$$dX_t^u = (u_t - \nabla V(X_t^u))\,dt + 2^{-1/2}\,dB_t$$





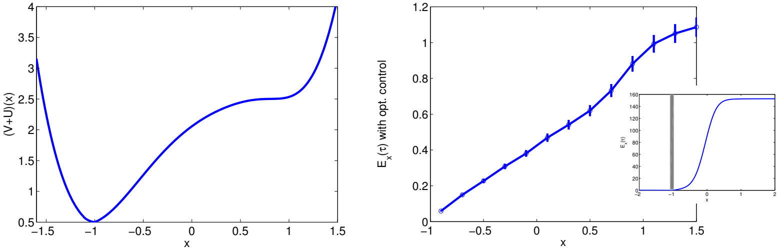
Skew double-well potential V and MFPT of the set $S = [-1.1, -1]$ (FEM reference solution).

[H & Schütte, JSTAT, 2012]

Computing the mean first passage time, cont'd

Cross-entropy minimization using a parametric ansatz

$$c(x) = \sum_{i=1}^{10} \alpha_i \nabla \phi_i(x)\,, \quad \phi_i : \text{ equispaced Gaussians}$$



Biasing potential $V + 2F$ and unbiased estimate of the limiting MFPT.

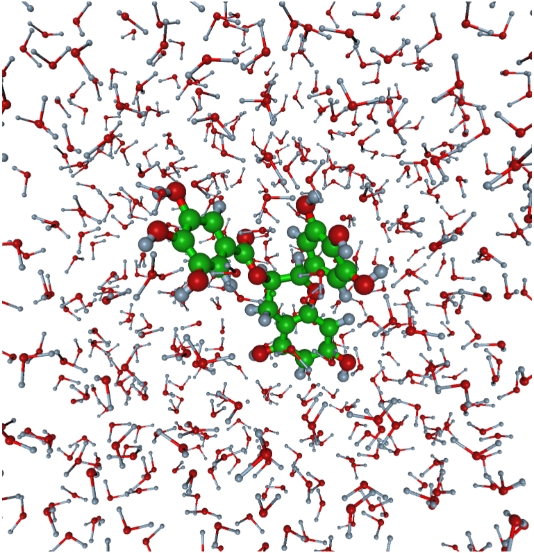
cf. [Lorenz Richter, MSc thesis, 2016], [Arampatatzis et al, JCP, 2016]

Adaptive importance sampling of rare events

Optimal controls from cross-entropy minimization

Numerical strategies for high-dimensional problems

The bad news



The good news: suboptimal controls from averaging

Averaged control problem: minimize

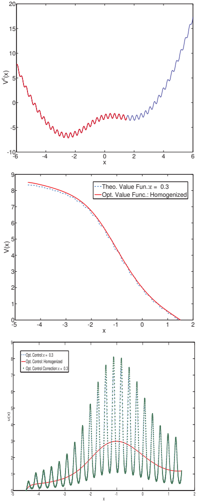
$$I(v) = \mathbb{E}\left[\bar{W}(\xi^v) + \frac{1}{4}\int_0^{\tau^v} |v_s|^2\, ds\right]$$

subject to the averaged dynamics

$$d\xi_t^u = (\Sigma(\xi_t^v)v_t - B(\xi_t^v))dt + \Sigma(\xi_t^v)dW_t$$

Control approximation strategy

$$u_t^* \approx c(\xi(X_t^{u^*}), t) = \nabla\xi(X_t^{u^*})v_t^*\,.$$



[H et al, Nonlinearity, 2016]; cf. [Legoll & Lelièvre, Nonlinearity, 2010]

The good news, cont'd

Uniform bound of the relative error using "averaged" optimal controls

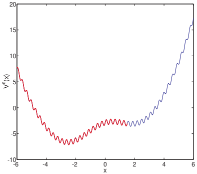
$$\delta_{\text{rel}} \le CN^{-1/2}\,\eta^{1/8}\,, \quad \eta = \frac{\tau_{\text{fast}}}{\tau_{\text{slow}}}$$

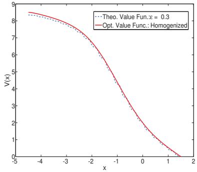
Issues for **highly oscillatory controls**:

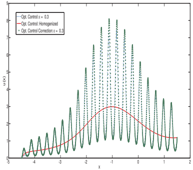
(a) Weak convergence of controls

$$u^\eta \rightharpoonup u \quad \not\Rightarrow \quad J(u^\eta) \to J(u)$$

(b) Log efficient estimators based on HJB subsolutions due to Dupuis, Spiliopoulos and Wang







[H et al, J Comp Dyn, 2014], [Zhang et al, PTRF, 2017], cf. [Spiliopoulos et al, MMS, 2012]

Remark: homogenization of forward-backward SDE

- **FBSDE** representation for a finite stopping time $\tau = T$:

$$dX_s^\eta = b^\eta(X_s^\eta)ds + \sigma^\eta(X_s^\eta)dW_s\,,\ X_t^\eta = x$$
$$dY_s^\eta = h^\eta(X_s^\eta, Y_s^\eta, Z_s^\eta)ds + Z_s^\eta \cdot dW_s\,,\ Y_T^\eta = g(X_T^\eta)\,,$$

where $t \leq s \leq T$ and

$$F^\eta(x,t) = Y_t^\eta \quad \text{(as a function of the initial value } x\text{)}$$

- **Homogenization result:** strong convergence of control value (so far in the Gaussian case only)

$$\sup\{|Y_t^\eta - Y_t| \colon 0 \leq t \leq T\} \leq C\eta^{1/4}$$

- Semi-explicit discretization of BSDE by **least-squares MC**.

[Bender & Steiner, in: Numer Meth Finance, 2012], [Kebiri et al, Preprint, 2017]

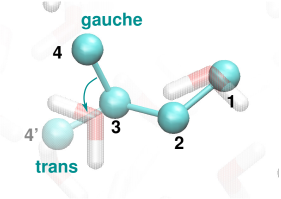
Example II (suboptimal control)

Conformational transition of butane in water ($n = 16224$)

Probability of making a **gauche-trans transition** before time T:

$$-\log \mathbb{P}(\tau_C \le T) = \min_u \mathbb{E}\left[\frac{1}{4}\int_0^\tau |u_t|^2\,dt - \log \mathbf{1}_{\partial C}(X_\tau)\right],$$

with $\tau = \min\{\tau_C, T\}$ and τ_C denoting the first exit time from the gauche conformation "C" with smooth boundary ∂C



T [ps]	$\mathbf{P}(\tau \le T)$	Error	Var	Accel. $\mathcal{I}$
0.1	4.30×10^{-5}	0.77×10^{-5}	3.53×10^{-6}	42.5
0.2	1.21×10^{-3}	0.11×10^{-3}	2.50×10^{-4}	26.0
0.5	6.85×10^{-3}	0.38×10^{-3}	2.88×10^{-3}	13.0
1.0	1.74×10^{-2}	0.08×10^{-2}	1.21×10^{-2}	7.0

IS of butane in a box of 900 water molecules (SPC/E, GROMOS force field) using cross-entropy minimization

[Zhang et al, SISC, 2014]

Take-home message

- Nonasymptotic adaptive importance sampling scheme based on **equivalent (dual) optimal control problem**.
- **Variational problem:** find the optimal perturbation by cross-entropy minimization.
- Method features short trajectories with **minimum variance estimators** of the rare event statistics.

- **Next steps:** adaptivity, non-parametric framework, ...

Thank you for your attention!

Acknowledgement:

Omar Kebiri
Lara Neureither
Lorenz Richter
Christof Schütte
Han Wang
Wei Zhang

German Science Foundation (DFG)
Einstein Center for Mathematics Berlin (ECMath)