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

1.3 $\mu$s 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, ...
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

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

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 \( \hat{\psi} = \psi_\epsilon \)
  \[ \hat{\psi}_\epsilon^N = \frac{1}{N} \sum_{i=1}^{N} e^{-\alpha \tau_i^C}. \]
- 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
  \[
  \delta_\epsilon = \sqrt{\frac{\text{Var}[\hat{\psi}_N^\epsilon]}{\mathbb{E}[\hat{\psi}_N^\epsilon]}}
  \]

- Varadhan’s large deviations principle
  \[
  \mathbb{E}[(\hat{\psi}_\epsilon^N)^2] \gg (\mathbb{E}[\hat{\psi}_\epsilon^N])^2, \quad \epsilon \text{ small.}
  \]

- Unbounded relative error as \( \epsilon \to 0 \)
  \[
  \limsup_{\epsilon \to 0} \delta_\epsilon = \infty
  \]

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
Mean first passage time for small $\epsilon$

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

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 $X^u = (X^u_t)_{t \geq 0}$,

$$dX^u_t = (b(X^u_t) + \sigma(X^u_t)u_t)dt + \sigma(X^u_t)dB_t, \quad X^u_0 = x,$$

estimate a reweighthed version of $\psi$:

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

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?
Theorem (Donsker & Varadhan)

For any bounded and measurable function $W$ it holds

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

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

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

$$- \log \int e^{-W} d\mu_0 = - \log \int e^{-W + \log \varphi} d\mu \leq \int (W - \log \varphi) d\mu$$

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

Same same, but different...
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

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

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

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

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

\[ F(x, t) = \min\{ J(u) : X_t^u = x \}. \]

(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.

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

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

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 \( X_t^{\mu=0} \).
Related work on asymptotics (non-exhaustive)


- 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 $\phi_i \in C^{1,0}(\mathbb{R}^n, [0, \infty))$. 
Letting $\mu$ 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 $\mu$ and $\mu^*$. 
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, ...
... that doesn’t work without knowing the normalization factor $\psi$.

### 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 $\mu = \mu^*$.

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 $\varphi$ and $\mu$ (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 $\phi_i$ (e.g. global or local) greatly **enhances convergence**.

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 \(\tau = \inf\{ t > 0 : X_t \in [-1.1, -1]\}\) 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]
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

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

subject to the averaged dynamics

$$d\xi_t^u = (\Sigma(\xi^\nu) \nu_t - B(\xi^\nu)) \, dt + \Sigma(\xi^\nu) \, dW_t$$

Control approximation strategy

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

[H et al, Nonlinearity, 2016]; cf. [Legoll & Lelièvre, Nonlinearity, 2010]
Uniform bound of the relative error using “averaged” optimal controls

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

Issues for highly oscillatory controls:
(a) Weak convergence of controls

\[ u^n \rightharpoonup u \not\Rightarrow J(u^n) \to J(u) \]

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

Remark: homogenization of forward-backward SDE

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

  $dX_t^\eta = b^\eta(X_t^\eta)dt + \sigma^\eta(X_t^\eta)dW_t, \quad X_t^\eta = x$

  $dY_t^\eta = h^\eta(X_t^\eta, Y_t^\eta, Z_t^\eta)dt + Z_t^\eta \cdot dW_t, \quad Y_T^\eta = g(X_T^\eta)$,

  where $t \leq s \leq T$ and

  \[ F^\eta(x, t) = Y_t^\eta \] (as a function of the initial value $x$)

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

  \[ \sup \{|Y_t^\eta - Y_t|: 0 \leq t \leq T\} \leq C_\eta^{1/4} \]

- 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 \(T\):

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

with \(\tau = \min\{\tau C, T\}\) and \(\tau C\) denoting the first exit time from the gauche conformation “C” with smooth boundary \(\partial C\)

<table>
<thead>
<tr>
<th>(T) [ps]</th>
<th>(\mathbb{P}(\tau \leq T))</th>
<th>Error</th>
<th>Var</th>
<th>Accel. (I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>(4.30 \times 10^{-5})</td>
<td>(0.77 \times 10^{-5})</td>
<td>(3.53 \times 10^{-6})</td>
<td>42.5</td>
</tr>
<tr>
<td>0.2</td>
<td>(1.21 \times 10^{-3})</td>
<td>(0.11 \times 10^{-3})</td>
<td>(2.50 \times 10^{-4})</td>
<td>26.0</td>
</tr>
<tr>
<td>0.5</td>
<td>(6.85 \times 10^{-3})</td>
<td>(0.38 \times 10^{-3})</td>
<td>(2.88 \times 10^{-3})</td>
<td>13.0</td>
</tr>
<tr>
<td>1.0</td>
<td>(1.74 \times 10^{-2})</td>
<td>(0.08 \times 10^{-2})</td>
<td>(1.21 \times 10^{-2})</td>
<td>7.0</td>
</tr>
</tbody>
</table>

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)