Large deviation estimators and their efficiency

Hugo Touchette

Department of Mathematical Sciences
Stellenbosch University, South Africa

RESIM 2021 (2020)
Paris, France

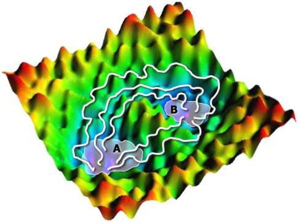
- Work with Arnaud Guyader (Sorbonne, Paris)
  Guyader & HT, J. Stat. Phys. **181**, 551, 2020

Rare events and large deviations

Equilibrium

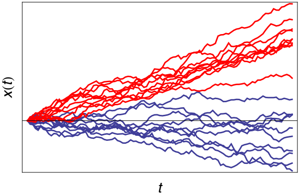
$$M_N = \frac{1}{N}\sum_{i=1}^{N} S_i$$

Dynamic transitions



$$X_0^\epsilon \in A \to X_T^\epsilon \in B$$

Integrated RVs



$$A_T = \frac{1}{T}\int_0^T f(X_t)dt$$

Large deviations

$$\text{Prob} \asymp e^{-nI}, \qquad n \to \infty$$

$n = N$ — Thermo limit

$n = 1/\epsilon$ — Low noise/temp

$n = T$ — Ergodic limit

Dynamical large deviations

- Markov process: X_t
- Observable: A_T

Direct problem

$$P(A_T = a) \asymp e^{-T I(a)}$$

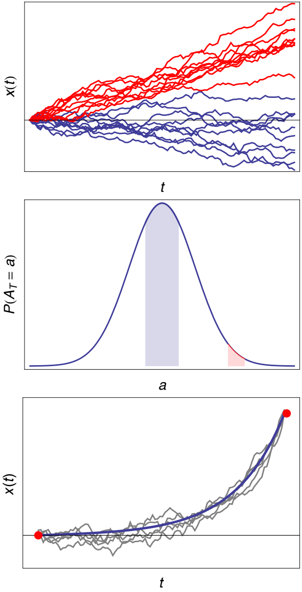
Dual problem

$$E[e^{TkA_T}] \asymp e^{T\lambda(k)}$$

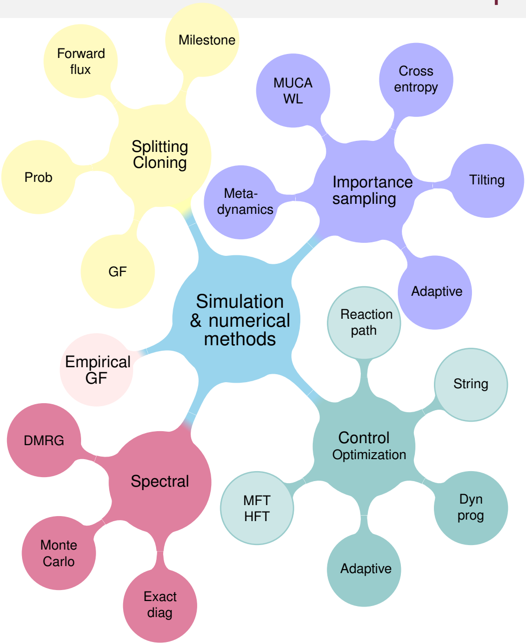
Prediction problem

How is fluctuation created?

- Reaction or optimal path
- Conditioning: $X_t | A_T = a$



Numerical and simulation techniques



- Numerical methods
- Sampling methods

- Thermo limit
- Low noise
- Long time

- Reversible processes
- Non-reversible

This talk

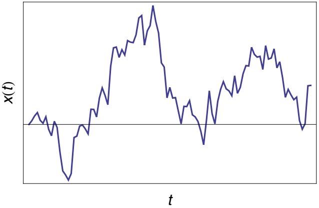
- Importance sampling
- Efficiency conditions

Long-time large deviations

[Review: HT Physica A 2018]

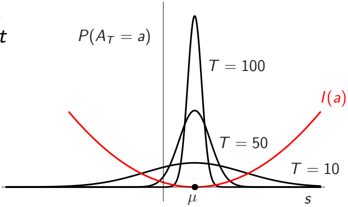
- Process:

$$dX_t = F(X_t)dt + \sigma dW_t$$



- Observable:

$$A_T = \frac{1}{T}\int_0^T f(X_t)\,dt + \frac{1}{T}\int_0^T g(X_t)\circ dX_t$$



- Large deviation principle:

$$P(A_T = a) \asymp e^{-T I(a)}, \quad T \to \infty$$

Examples

- Occupation, current, activity, etc.
- Stochastic thermo: Work, heat, entropy production

Spectral problem

Scaled cumulant generating fct

$$\lambda(k) = \lim_{T\to\infty} \frac{1}{T} \ln E[e^{TkA_T}]$$

Gärtner-Ellis Theorem

$\lambda(k)$ differentiable, then

1. $P(A_T = a) \asymp e^{-TI(a)}$
2. $I(a) = \sup_{k\in\mathbb{R}}\{ka - \lambda(k)\}$

Perron–Frobenius

$$\mathcal{L}_k r_k = \lambda(k) r_k$$

- Tilted (twisted) operator:

$$\mathcal{L}_k = F \cdot (\nabla + kg) + \frac{D}{2}(\nabla + kg)^2 + kf$$

- Dominant eigenvalue: $\lambda(k)$
- Dominant eigenfunction: $r_k(x)$

Prediction problem

[Chetrite & HT PRL 2013, AHP 2015, JSTAT 2015]

Driven process

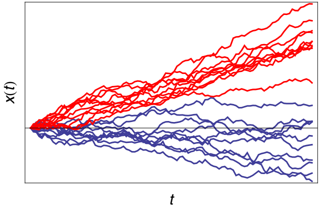
$$d\tilde{X}_t = \tilde{F}(\tilde{X}_t)dt + \sigma dW_t$$

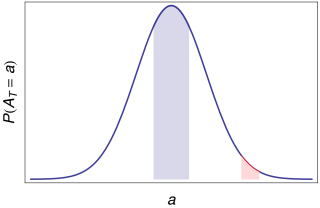
- Modified drift:

$$\tilde{F}(x) = F(x) + \sigma^2(k g + \nabla \ln r_k), \quad I'(a) = k$$

Interpretation

$$\underbrace{X_t \,|\, A_T = a}_{\text{conditioned}} \overset{T\to\infty}{\cong} \underbrace{\tilde{X}_t}_{\text{driven}}$$





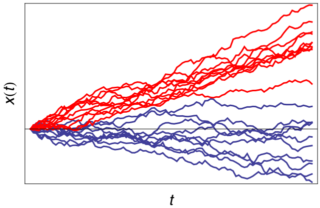
- Effective process creating fluctuation
- Generalization of reaction path / instanton

Direct sampling

$$P(A_T = a) \asymp e^{-TI(a)} \qquad A_T = A_T[x]$$

- Sample:

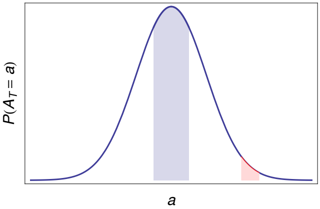
$$\left\{ \{x_t^{(j)}\}_{t=0}^{T} \right\}_{j=1}^{L} \quad \to \quad \{A_T^{(j)}\}_{j=1}^{L}$$



- Estimators:

$$\hat{P}_{T,L}(a) = \frac{1}{L}\sum_{j=1}^{L} \mathbf{1}_{[a,a+\Delta a]}(A_T^{(j)})$$

$$\hat{I}_{T,L}(a) = -\frac{1}{T}\ln \hat{P}_{T,L}(a)$$



Repeat for increasing L (sample size) and T (LD limit)

Importance sampling

Change process to hit rare event more often (reweighting)

- Modified process: $\tilde{X}_t \sim Q$
- Sample:

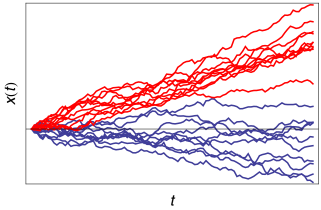
$$\left\{ \{\tilde{x}_t^{(j)}\}_{t=0}^T \right\}_{j=1}^L \quad \to \quad \{A_T^{(j)}\}_{j=1}^L$$

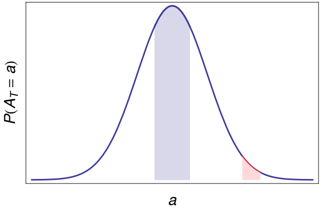
- Estimator:

$$\hat{P}_{T,L}(a) = \frac{1}{L}\sum_{j=1}^L \mathbf{1}_{[a,a+\Delta a]}(A_T^{(j)}) \underbrace{\frac{dP}{dQ}[\tilde{x}^{(j)}]}_{\text{Likelihood}}$$

- Unbiased:

$$P(A_T = a) = \underbrace{E_P[\mathbf{1}_a]}_{\text{direct}} = \underbrace{E_Q\left[\mathbf{1}_a \frac{dP}{dQ}\right]}_{\text{reweighted}}$$





Exponential tilting

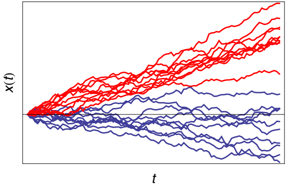
- Modified process:

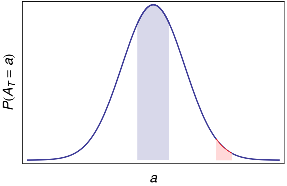
$$\underbrace{\color{red}{Q}[dx]}_{\text{modified}} = \underbrace{\frac{e^{TkA_T[x]}}{E_{\color{red}{Q}}[e^{TkA_T}]}}_{\text{exp reweighting}} \underbrace{\color{blue}{P}[dx]}_{\text{original}}$$

  - Esscher transform (1932)
  - Canonical ensemble

- Likelihood:

$$\frac{d\color{blue}{P}}{d\color{red}{Q}}[x] \asymp e^{-TA_T[x]+T\color{red}{\lambda(k)}}$$





- Exponential tilting is good IS measure
- Markovian for large T
- Equivalent to driven process
- **Problem:** Requires LD functions (no free lunch)

Problems

What is optimal IS process?

- X_t conditioned on $A_T = a$
- Zero variance
- **Problem:** Process can't be constructed in general

Good definition of optimal / efficient IS?

- Asymptotic efficiency (AE)
- Exponential tilting is AE
- **Problem:** Other efficient processes?

How to construct efficient processes?

- Driven process is AE
- **Problem:** Driven process based on r_k (no free lunch)

Asymptotic efficiency

- IS estimator:

$$\hat{P}_{T,L}(a) = \frac{1}{L}\sum_{j=1}^{L} \mathbf{1}_a(A_T^{(j)})\frac{dP}{dQ}[\tilde{x}^{(j)}], \qquad \tilde{x}^{(j)} \sim Q$$

- Variance:

$$\mathrm{Var}_Q(\hat{P}_{T,L}(a)) = \frac{E_Q[L_T^2\,\mathbf{1}_a(A_T)] - P_T(a)^2}{L}$$

Asymptotic efficiency (AE)

- Second moment rate:

$$E_Q[L_T^2\,\mathbf{1}_a(A_T)] \asymp e^{-TR_Q(a)}$$

- Bound: $R_Q(a) \leq 2I(a)$
- Q is AE if equality achieved

Efficiency conditions

Work with Arnaud Guyader (Paris) [Guyader & HT JSP 2020]

- Most works about exponential tilting
- Only sufficient conditions for AE

Find necessary and sufficient conditions for general Q to be AE

$$E_Q[L_T^2\, \mathbf{1}_a(A_T)], \qquad L_T = \frac{dP}{dQ}$$

1. AE determined by A_T and L_T
2. Likelihood exponential in T: $L_T = e^{-TW_T}$
3. (A_T, W_T) satisfies LDP under Q

$$E_Q[e^{-2TW_T}\, \mathbf{1}_a(A_T)] \asymp \underbrace{\int dw\, e^{-2Tw}\, e^{-TJ_Q(a,w)}}_{\text{Laplace integral}} \asymp e^{-TR_Q(a)}$$

Main result

[Guyader & HT JSP 2020]

- Action:

$$L_T = e^{-TW_T}, \qquad W_T = -\frac{1}{T}\log L_T$$

- Joint LDP:

$$Q(A_T = a, W_T = w) \asymp e^{-TJ_Q(a,w)}$$

- AE criterion:

$$\underbrace{\inf_w \{2w + J_Q(a,w)\}}_{R_Q(a)} \leq 2 \underbrace{\inf_w \{w + J_Q(a,w)\}}_{I(a)}$$

Theorem

Q is AE if and only if

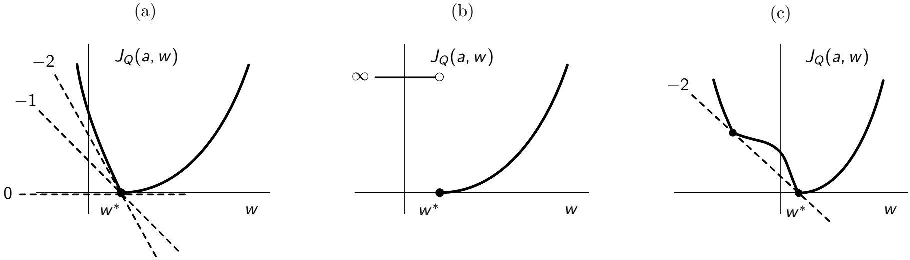
- There exists w^* such that $J_Q(a, w^*) = 0$ (Typicality condition)
- Left w-slope of $J_Q(a, w)$ at $w^* < -2$ (Steepness condition)

Interpretation

$$P(A_T = a) \asymp e^{-TI(a)}, \qquad Q(A_T = a, W_T = w) \asymp e^{-TJ_Q(a,w)}$$

$$E_Q[e^{-2TW_T}\,\mathbf{1}_a(A_T)] \asymp \int dw\, e^{-2Tw}\, e^{-TJ_Q(a,w)}$$

- **Typicality condition:** $A_T \to a$ under Q
- **Steepness condition:** Suppress fluctuations $W_T < w^*$
- Not AE if $J_Q(a,w)$ has smooth zero



Examples (1/2)

Exponential tilting

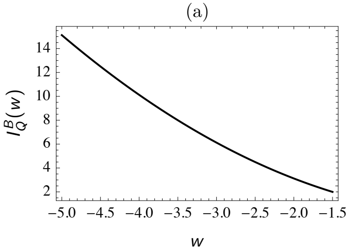
- Action: $W_T = kA_T + c$
- Rate function:

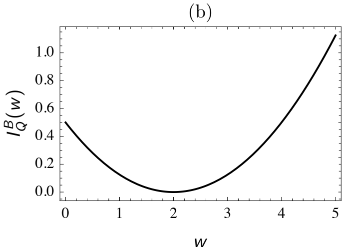
$$J_Q(a,w) = \begin{cases} 0 & w = ka + c \\ \infty & \text{otherwise} \end{cases}$$

Gaussian sums

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad X_i \sim \mathcal{N}(0,1)$$

- $P(S_n \geq 1) \asymp e^{-nI}$, $I = \frac{1}{2}$
- $\tilde{X}_i \sim \mathcal{N}(\mu, 1)$
  - $\mu < 1$: No zero. Not AE
  - $\mu = 1$: Zero + steep. AE
  - $\mu > 1$: Zero, not steep. Not AE





Examples (2/2)

Exponential sums

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad X_i \sim \text{Exp}(1), \qquad P(S_n \geq b) \asymp e^{-nI(b)}$$

$$I(b) = b - 1 - \log b$$

- Exponential tilting: $\tilde{X}_i \sim \text{Exp}(1/b)$
- Partial tilting: $\tilde{X}_i \sim \text{Exp}(1/b) \qquad i = 1, \ldots, n-1$
  $\tilde{X}_n \sim \text{Exp}(1)$
- AE if $b \in (1, 2]$, not AE if $b > 2$.

Other examples

- Markov chains
- Diffusions
- See [Guyader & HT JSP 2020]

Other estimators

$$\hat{P}_{T,L}(a) = \frac{1}{L}\sum_{j=1}^{L} \mathbf{1}_a(A_T^{(j)}) \qquad \longrightarrow \qquad \hat{I}_{T,L}(a) = -\frac{1}{T}\ln \hat{P}_{T,L}(a)$$

- Estimate rate function without histograms
- Optimal running cost:

$$I(a) = \lim_{T\to\infty} \inf_{\substack{\tilde{X}_t \\ A_T=a \text{ typical}}} \frac{1}{2T}\int_0^T [F(\tilde{X}_t) - \tilde{F}(\tilde{X}_t)]^2 dt$$

- Optimal $\tilde{X}_t$ = driven process [Chetrite & HT JSTAT 2015]

Trade off

- Use exponential LD structure
- Good variance
- **Problem:** Biased if not optimal

Conclusion

- New conditions for efficient sampling of large deviations
- Can be applied beyond exponential tilting
- $J_Q(a, w)$ explains bad IS cases
- Predicts estimator convergence

Open problems

- Errors when not AE (how bad when not optimal?)
- Bias bounds for control estimators

Ongoing works (from stat phys)

- Adaptive approximations of driven process [Ferré & HT JSP 2018]
- Spectral approximations [Garrahan (Nottingham)]
- Machine learning approaches (Tensor net, neural net, RL) [E (Princeton), Limmer (Berkeley)]

References

A. Guyader, H. Touchette
Efficient large deviation estimation based on importance sampling
J. Stat. Phys. **181**, 551, 2020

R. Chetrite, H. Touchette
Nonequilibrium Markov processes conditioned on large deviations
Ann. H. Poincaré **16**, 2005, 2015

R. Chetrite, H. Touchette
Variational and optimal control representations of conditioned and driven processes
J. Stat. Mech. 2015, P12001

appliedmaths.sun.ac.za/~htouchette

J. Bucklew
Introduction to Rare Event Simulation
Springer, 2004

Joint large deviations

- Observable: A_T
- Action: W_T
- Joint SCGF:

$$\lambda_Q(k,\gamma) = \lim_{T\to\infty} \frac{1}{T} \ln E_Q[e^{TkA_T + T\gamma W_T}]$$

- Markov processes: $\lambda_Q(k,\gamma) =$ dom eigenvalue

Gärtner–Ellis Theorem

If $\lambda_Q(k,\gamma)$ is differentiable, then

$$J_Q(a,w) = \inf_{k,\gamma}\{ka + \gamma w - \lambda_Q(k,\gamma)\}$$