Multilevel Sampling

Monte Carlo Methods for sequences of probability measures

Andreas Eberle

December 4, 2010

1 INTRODUCTION

THE PROBLEM : $\mu_0, \mu_1, \ldots, \mu_k$ probability measures on state space S.

E.g.: $S = V$ vertex set of graph, $S = \{0,1\}^V$, $S = \mathbb{R}^d$, $S = C([0,1], \mathbb{R}^d)$.

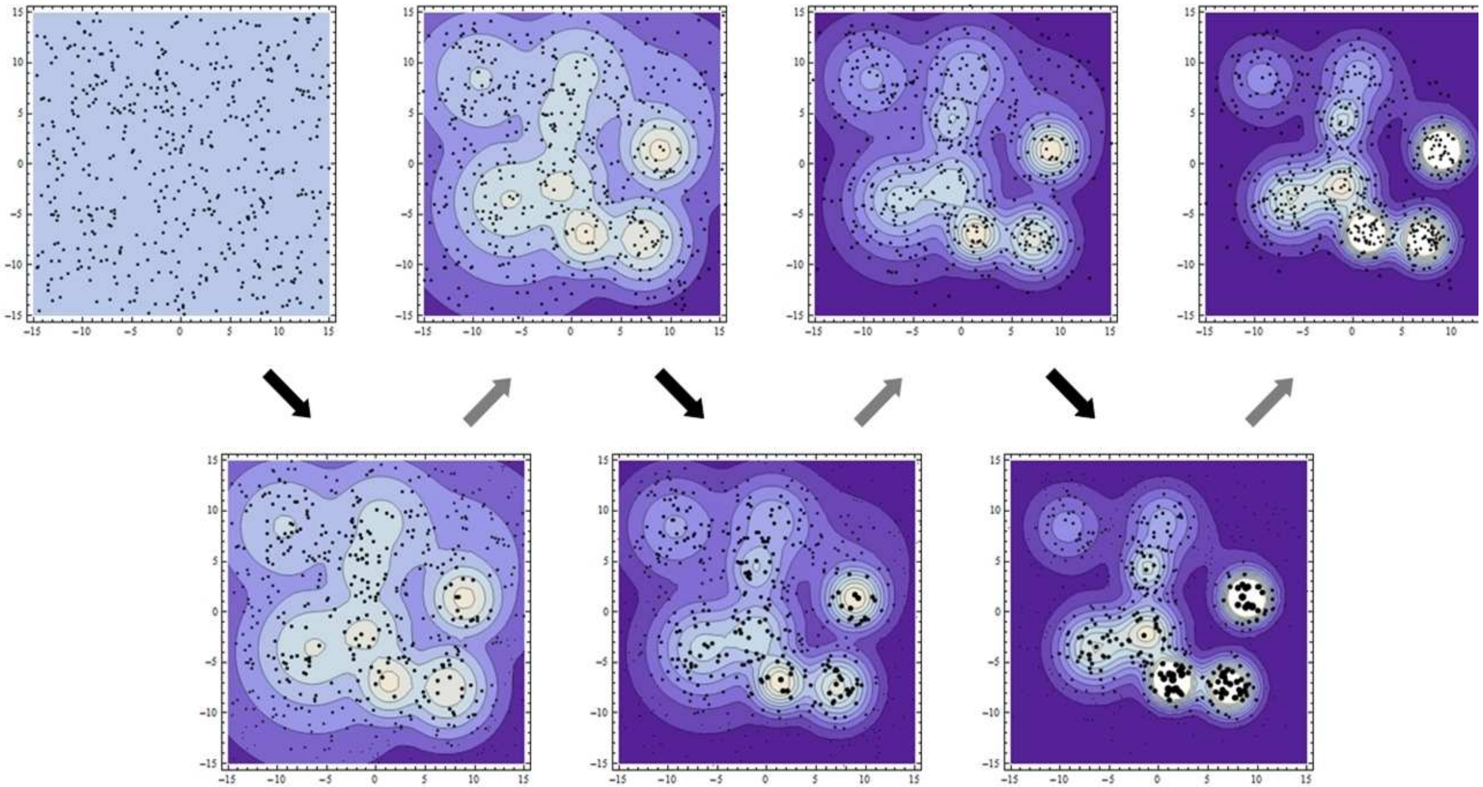
$$\mu_t(dx) \;=\; \mu_t(x)\,\nu(dx)\,; \qquad \mu_t(x) \;=\; \frac{1}{Z_t}\cdot \exp(-U_t(x))$$

- ν reference measure
- $U_t : S \to \mathbb{R}$ known/can be computed
- Z_t unknown normalization constant/partition function

AIM : Sequential Monte Carlo Estimation/Approximation of

$$\langle f, \mu_t\rangle \;:=\; \int_S f(x)\,\mu_t(dx)\,, \qquad t = 0, 1, \ldots, k\,,$$

for appropriate class of functions $f : S \to \mathbb{R}$.



MOTIVATION:

- *Dynamical:* E.g. non-linear filtering, Hidden Markov Models

  $X_0, X_1, \ldots, X_k$ Markov chain *(unobserved signal)*
  $Y_0, Y_1, \ldots, Y_k$ noisy perturbations *(observation)*

  $\mu_t =$ conditional distribution of X_t given $Y_0, \ldots, Y_t$

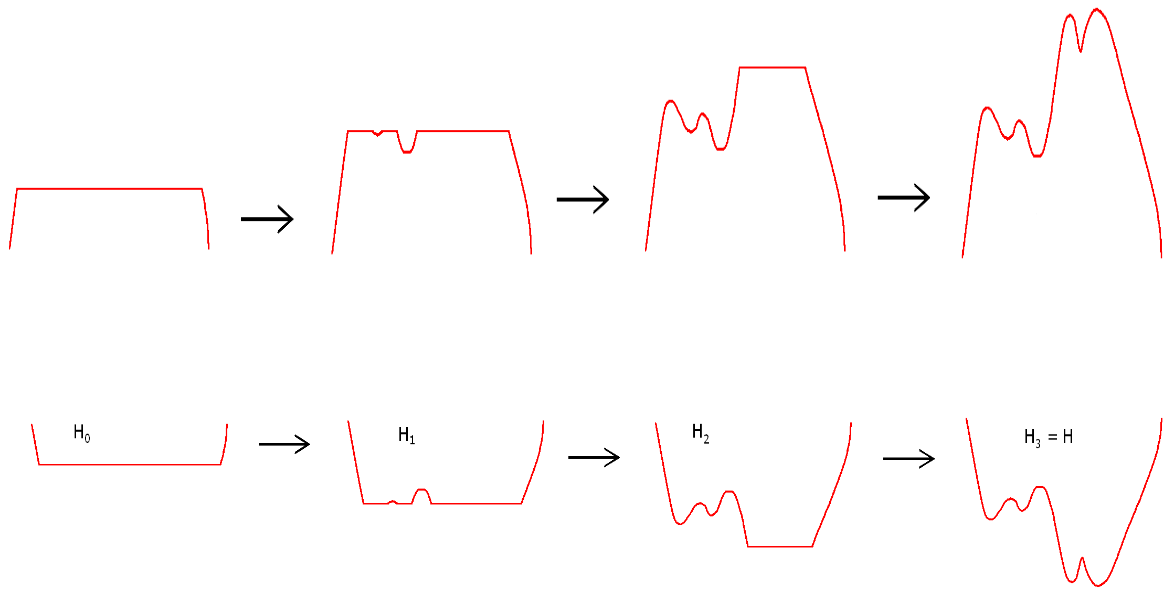
  $\rightsquigarrow$ Particle filters, see e.g. [Doucet, de Freitas, Gordon: *Sequential MC methods in practice*].

- *Static:* $\mu(dx) = Z^{-1} \exp(-U(x))\, \nu(dx)$ Target distribution

  - difficult to simulate because of multimodality, metastable states, singular density,...
  - $\rightsquigarrow$ choose interpolation $\nu = \mu_0, \mu_1, \ldots, \mu_k = \mu$, e.g. such that $\mu_k(x)/\mu_{k-1}(x) \leq 2$
  - $\rightsquigarrow$ *"resolution of singularity" by "homotopy method".*

EXAMPLES: Choice of interpolating probability measures

- *Annealing :* $\mu_t(x) \propto \exp(-\beta_t U(x))$, β_t cooling schedule, $\beta_0 = 0$, $\beta_k = 1$.
- *Equi-Energy Sampling :* $\mu_t(x) \propto \exp(-\beta_t \cdot \max(U(x), E_t))$

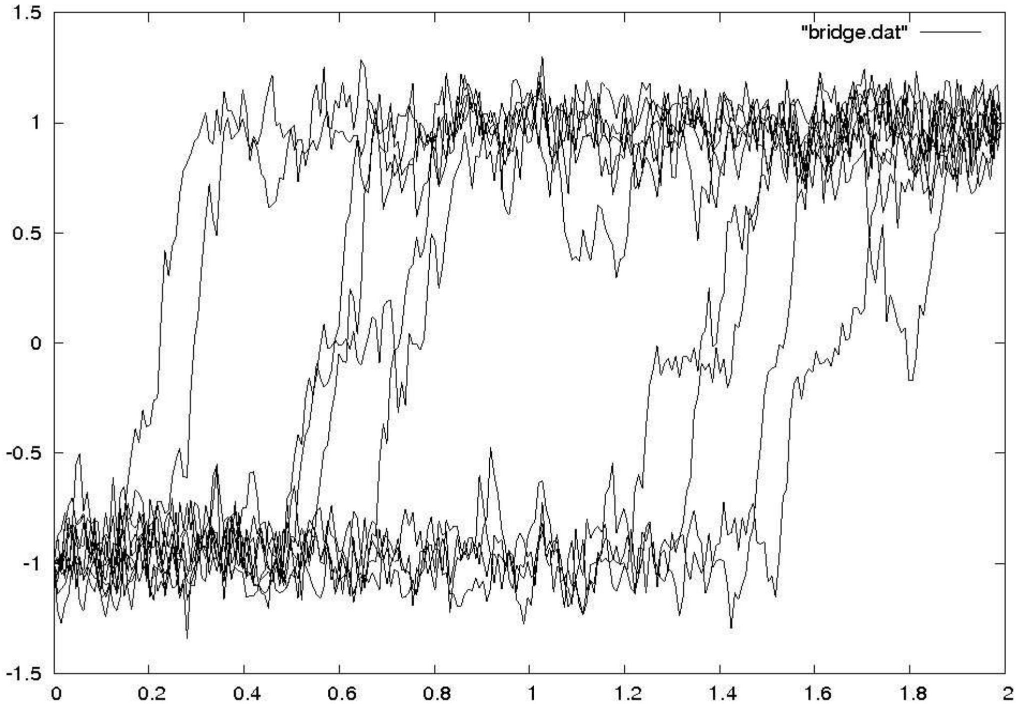


- *Spatial Coarse Graining :*

  e.g. μ measure on $S = C([0,1], \mathbb{R}^d)$, μ_t approximation on

$$S_k = \{\omega : [0,1] \to \mathbb{R}^d \ : \ \omega \text{ linear on } [(k-1)2^{-t}, k2^{-t}] \text{ for any } k\}.$$

  $\rightsquigarrow$ Transition Path Sampling



MARKOV CHAIN MONTE CARLO :

- *Energy Landscape:* $U : S \to \mathbb{R}_+$, $\mu_t(x) \propto \exp(-\beta_t \cdot \max(U(x), E_t))$
- *Markov Chain Monte Carlo Approach (MCMC):*
  - Explore energy landscape by reversible stochastic dynamics: (X^i) Markov chain with transition density $p_t(x,y)$ s.t.

$$\mu_t(x)\, p_t(x,y) \;=\; \mu_t(y)\, p_t(y,x) \qquad \text{Detailled Balance}$$

  - Estimate μ_t by $\hat{\mu}_t \;=\; \sum_{i=1}^{\lambda_t} \delta_{X^i}$.

  - Example [Metropolis et al. (1953), Hastings (1971)]

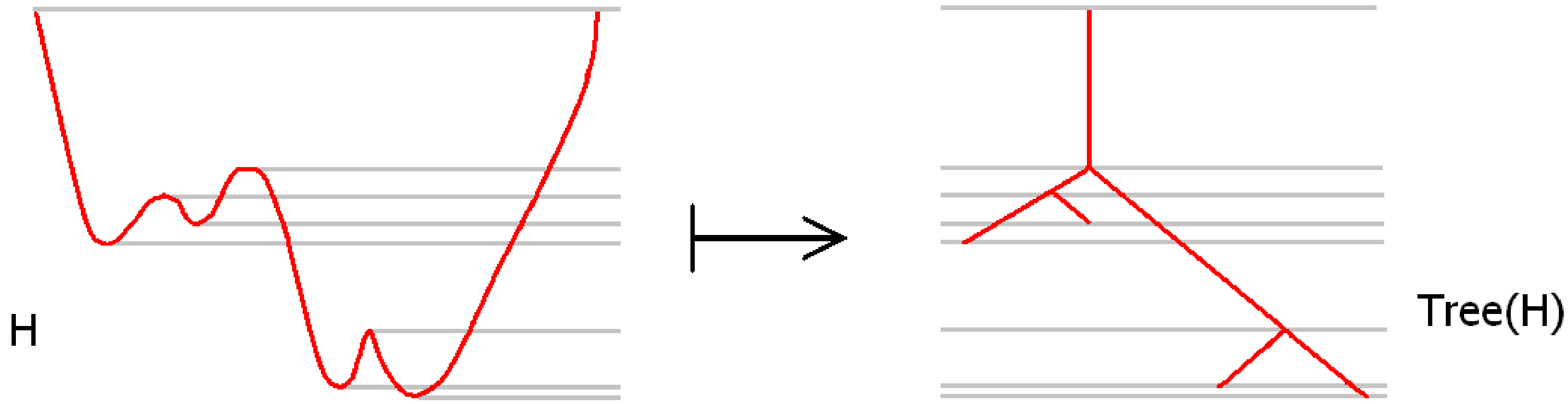
$$p_t(x,y) \;=\; q(x,y) \cdot \min\left(\frac{\mu_t(y)q(y,x)}{\mu_t(x)q(x,y)}, 1\right) \;, \quad q \text{ proposal density}$$

METASTABILITY PROBLEM :

- Local energy minima $\rightsquigarrow$ metastable states $\rightsquigarrow$ traps for Markov chain
- *Simulated Annealing with logarithmic cooling schedule:* Cool down so slowly that Markov chain escapes traps.

  $\rightsquigarrow$ not feasible in practice !
- *Realistic approach:* Cool down much faster.

  $\rightsquigarrow$ Markov chain eventually gets trappped



DISCONNECTIVITY TREE OF ENERGY FUNCTION:

S	$\longrightarrow$	Disconnectivity tree $\mathbb{T}$
Energy function $U : S \to \mathbb{R}_+$	$\longrightarrow$	Height function $h : \mathbb{T} \to \mathbb{R}_+$
Reference measure ν	$\longrightarrow$	Density of states $\Omega(dx)$ on $\mathbb{T}$
μ_t	$\longrightarrow$	$\bar{\mu}_t(dx) \propto e^{-\beta_t h(x)}\,\Omega(dx)$

- As t increases, the Markov chain gets trapped in deeper branches of the tree.
- The state space effectively splits into an increasing number of components (metastable states)

KEY PROBLEM:

- Are there feasible Markov chain based methods for Monte Carlo integral estimation w.r.t. the sequence μ_t ($t = 0, 1, \ldots, k$) in spite of trapping ?
- When do they apply ?
- Any quantitative error bounds for simple models ?

2 MONTE CARLO METHODS FOR SEQUENCES

- *Importance Sampling*
- *Markov Chain Monte Carlo (MCMC)* [Metropolis et al. 1953]

- *Simulated Annealing* [Kirkpatrick, Gelatt, Vecchi 1982]
- *Simulated Tempering* [Marinari, Parisi 1992]
- *Parallel Tempering* [Geyer 1991]
- *Equi-Energy Sampler* [Kou, Zhou, Wong 2006]
- *Sequential Monte Carlo Samplers* [Del Moral, Doucet, Jasra 2006]

REFERENCE: Liu, *Monte Carlo Methods in Scientific Computing*

IMPORTANCE SAMPLING:

ALGORITHM (Importance Sampling, Umbrella Sampling).

- Sample X^i $(1 \leq i \leq N)$ i.i.d. $\sim \nu$
- For $t := 0, 1, \ldots, k$ do
  - Compute importance weights $w_t^i := \mu_t(X^i) / \sum_{j=1}^{N} \mu_t(X^j)$
  - Estimate μ_t by $\hat{\mu}_t^N := \sum_{i=1}^{N} w_t^i \, \delta_{X^i}$

Remarks.

- Computation of importance weights can be implemented sequentially
- Law of Large Numbers $\Rightarrow$ $\langle f, \hat{\mu}_t^N \rangle$ is asymptotically unbiased estimator for $\langle f, \mu_t \rangle$
- $\mathrm{Var}(\langle f, \hat{\mu}_t^N \rangle) = O(N^{-1/2})$ for any $f \in L^2(\mu_t)$

Drawback.

- Variance can be very large if $\sup_x \mu_t(x) / \inf_x \mu_t(x)$ is large.

MCMC: t fixed, λ_t nr. of steps, $p_t(x,y)$ transition density, DB w.r.t. μ_t

ALGORITHM (Independent Monte Carlo Markov Chains).

- Sample X_t^i $(1 \leq i \leq N)$ i.i.d. $\sim \nu$
- For $m := 1$ to λ_t do
  - Sample Y_t^i condit. independent $\sim p_t(X_t^i, \cdot\,)$; replace X_t^i by Y_t^i.
- Estimate μ_t by $\hat{\mu}_t^N := N^{-1} \sum_{i=1}^N \delta_{X_t^i}$

TWO ERRORS.

$$\hat{\mu}_t^N - \mu_t \quad = \quad \underbrace{(\hat{\mu}_t^N - \nu_t)}_{\text{MC error}} + \underbrace{(\nu_t - \mu_t)}_{\text{Deviation from equilibrium}}, \quad \nu_t = \text{ distrib. of chain after } \lambda_t \text{ steps}$$

- $E\left[\left|\langle f, \hat{\mu}_t^N\rangle - \langle f, \nu_t\rangle\right|^2\right] \;=\; N^{-1} \cdot \mathrm{Var}(f;\nu_t)$
- $\langle f, \nu_t\rangle - \langle f, \mu_t\rangle \;\to\; 0$ as $\lambda_t \to \infty$ if ergodicity holds.

QUANTITATIVE BOUNDS FOR DISTANCE FROM EQUILIBRIUM:

a) W.r.t. χ^2 divergence:

$$\chi^2(\nu_t|\mu_t) = \sup_{\langle f^2,\mu_t\rangle\le 1} |\langle f,\nu_t\rangle - \langle f,\mu_t\rangle|^2 \le e^{-\lambda_t/C_t}\cdot\chi^2(\nu|\mu_t)$$

$$C_t^{-1} = \inf_{\langle f,\mu_t\rangle=0} \frac{\mathcal{E}_t(f,f)}{\langle f^2,\mu_t\rangle} \qquad \text{Spectral gap,}$$

$$\mathcal{E}_t(f,f) = \int\int \mu_t(x)p_t(x,y)\,(f(y)-f(x))^2\,\nu(dx)\,\nu(dy) \quad \text{Dirichlet form}$$

b) W.r.t. Relative Entropy / Kullback-Leibler divergence: (Chain in cont. time)

$$H(\nu_t|\mu_t) = \int \frac{d\nu_t}{d\mu_t}\log\frac{d\nu_t}{d\mu_t}\,d\mu_t \le e^{-\lambda_t/\gamma_t}\cdot H(\nu|\mu_t)$$

$$\gamma_t^{-1} = \inf_{\langle f^2,\mu_t\rangle=1} \frac{\mathcal{E}_t(f,f)}{\langle f^2\log f^2,\mu_t\rangle} \qquad \text{Logarithmic Sobolev constant}$$

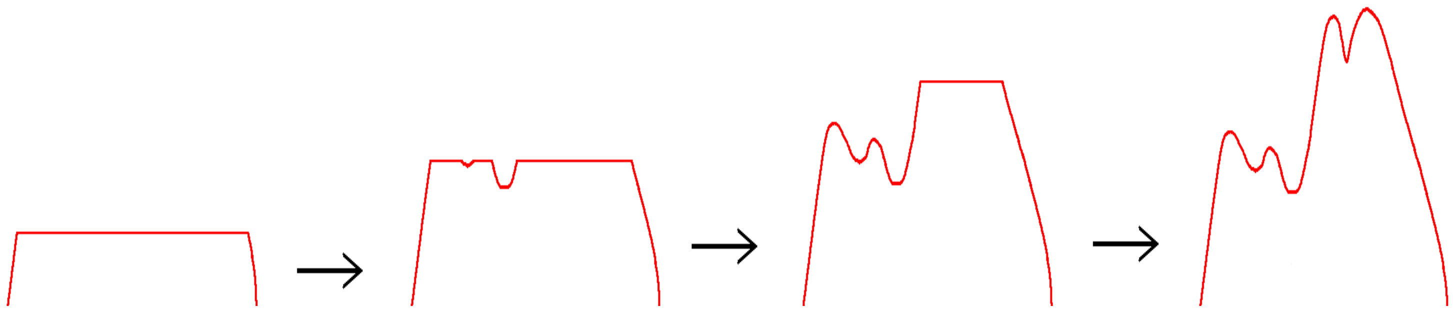
REFERENCE: Peres, *Markov Chains and mixing times*

SIMULATED TEMPERING:

- MCMC on $S \times \{0, 1, \ldots, k\}$ with stationary distribution

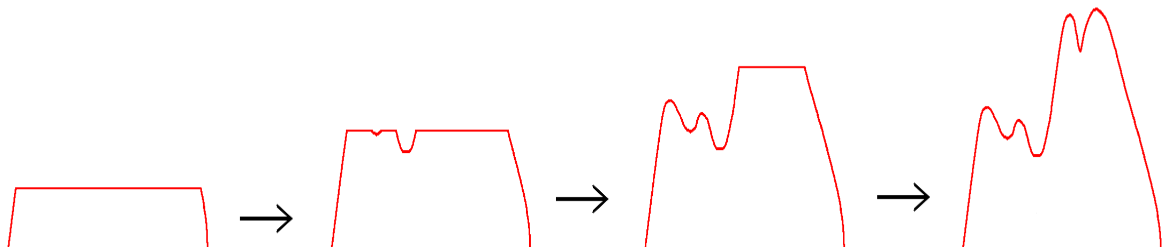
$$\bar{\mu}(x,t) \;\propto\; a_t \cdot \exp(-U_t(x))\,, \qquad \bar{\mu}(\,\cdot\,|t) = \mu_t\,.$$

- $a_t = Z_t^{-1} \;\Rightarrow\; \bar{\mu}(S,t) = (k+1)^{-1}$ Uniform distribution in t
- In practice use estimate $\hat{Z}_t$



PARALLEL TEMPERING: MCMC on $S^{\{0,1,\ldots,k\}}$ with equilibrium

$$\tilde{\mu}(x_0, x_1, \ldots, x_k) \;=\; \prod_{t=0}^{k} \mu_t(x_t)\,.$$



Transition step.

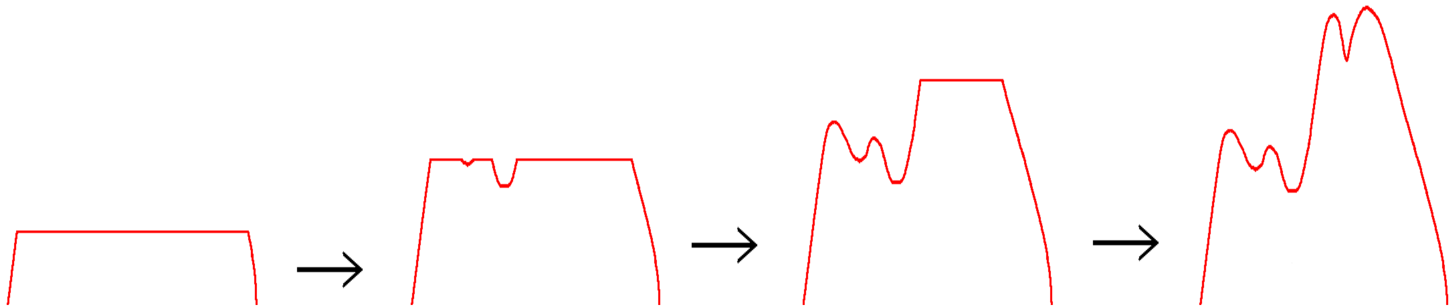
- Sample $t \sim \mathrm{Unif}\{0, 1, \ldots, k\}$
- *With probability* $1 - \varepsilon$ *:* Sample $Y_t \sim p_t(X_t, \cdot\,)$; replace X_t by Y_t.
- *With probability* ε *:* Sample $U \sim \mathrm{Unif}(0, 1)$;
  if $t > 0$ and $U < \min\left(1, \frac{\mu_t(X_{t-1})\mu_{t-1}(X_t)}{\mu_t(X_t)\mu_{t-1}(X_{t-1})}\right)$ then **exchange** X_t and X_{t-1}.

[Madras and Zheng], [Bhatnagar and Randall]

- Rapid Mixing on Mean Field Ising Model

$$C_t \;=\; O(N^{\alpha})\,, \qquad N = \text{number of spins}\,,$$

- BUT: Torpid Mixing on Mean-Field Potts Model.

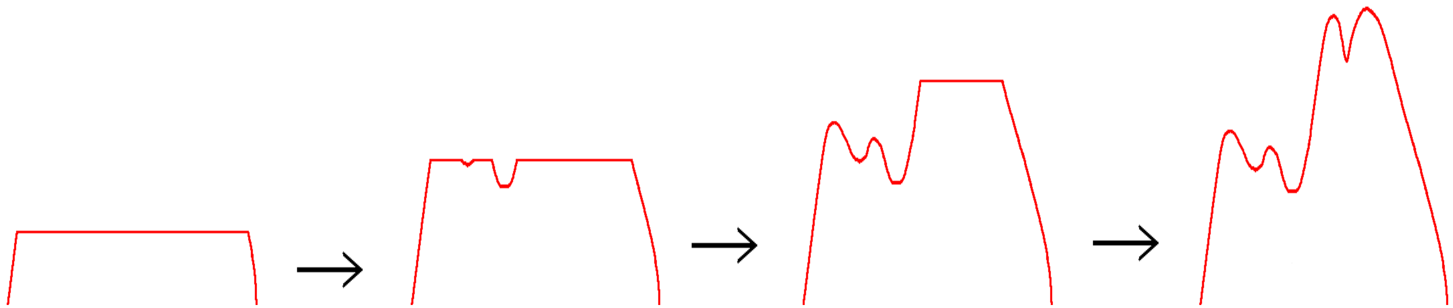


EQUI-ENERGY SAMPLER: [Kou, Zhou, Wong, *Annals of Statistics* 2006]

ALGORITHM (EE-Sampler). Fix $\lambda_0, \lambda_1, \cdots, \lambda_k \in \mathbb{N}$, $\delta_0, \delta_1, \cdots, \delta_k > 0$.

- Initialization for $t = 0$:
  - Sample $X_0^0, X_0^1, \ldots, X_0^{\lambda_0} \sim \mu_0$; set $S_0 := \{X_0^0, \ldots, X_0^{\lambda_0}\}$.
- Step: For $t := 1$ to k do
  - Sample $X_t^0 \sim \mathrm{Unif}(S_{t-1})$
  - For $i := 1$ to λ_t do
    - $*$ *With probability* $1-\varepsilon$ *:* Sample X_t^i condit. indep. $\sim p_t(X_t^{i-1}, \cdot\,)$
    - $*$ *With probability* ε *: PROPOSE EQUI-ENERGY MOVE*
      - · Sample $Y \sim \mathrm{Unif}\{x \in S_{t-1} : |U_t(x) - U_t(X_t^{i-1})| < \delta_t\}$;
      - · Sample $U \sim \mathrm{Unif}(0,1)$;
        if $U < \min\left(1, \frac{\mu_t(Y)\mu_{t-1}(X_t^{i-1})}{\mu_t(X_t^{i-1})\mu_{t-1}(Y)}\right)$ then set $X_t^i := Y$;
        else set $X_t^i := X_t^{i-1}$.
  - Set $S_t := \{X_t^0, \ldots, X_t^{\lambda_t}\}$.

EQUI-ENERGY SAMPLER



- Sequential algorithm
- Adaptive MCMC method : Detailed Balance w.r.t. μ_t holds only in the limit as $\lambda_i \to \infty$ for $i < t$!

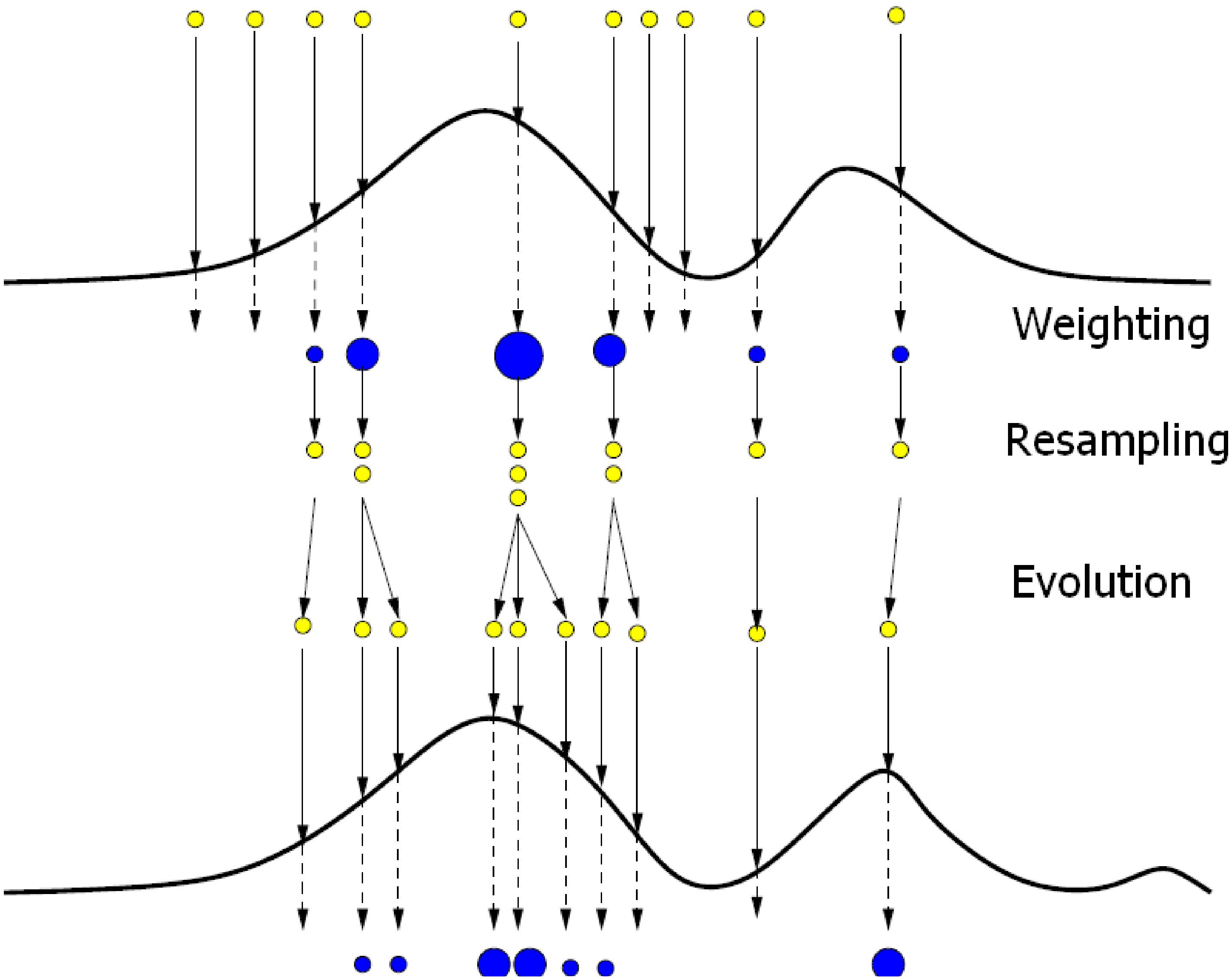
3 SEQUENTIAL MONTE CARLO SAMPLERS

P. Del Moral, A. Doucet, A. Jasra, J. R. Statist. Soc. B **68** (2006)

$\mu_0, \mu_1, \ldots, \mu_k$ probability measures, $\lambda_0, \lambda_1, \ldots, \lambda_k \in \mathbb{N}$,
$p_1, p_2, \ldots, p_k$ Markov kernels s.t. $\mu_i p_i = \mu_i$

ALGORITHM (SMCMC with multinomial resampling).

- Initialization:
  - Sample X_0^i $(1 \le i \le N)$ i.i.d. $\sim \mu_0$, set $\eta_0^N := N^{-1} \sum_{i=1}^N \delta_{X_0^i}$
- Step: For $t := 1$ to k do
  - **SIR**: Sample X_t^i i.i.d. $\sim \sum_{i=1}^N w_t^i \cdot \delta_{X_{t-1}^i}$, $w_t^i \propto \mu_t(X_{t-1}^i)/\mu_{t-1}(X_{t-1}^i)$
  - **MCMC**: For $m := 1$ to λ_t do
    - $*$ Sample Y_t^i condit. indep. $\sim p_t(X_t^i, \cdot)$; set $X_t^i := Y_t^i$
  - Set $\eta_t^N := N^{-1} \sum_{i=1}^N \delta_{X_t^i}$


Weighting
Resampling
Evolution

Related methods:

- *Parallel tempering*, Geyer (1991)
- *Equi-energy sampler*, Kou, Zhou, Wong, Annals of Statistics **34** (2006)

Good performance in various simulations, e.g. on Gaussian mixture models.

- Can we understand mathematically ?
- Can we even prove feasible quantitative bounds ?
- How to choose the interpolating probability measures ?
- How many MCMC moves per importance sampling/resampling step are required ?
- Dependence of error on structure of energy landscape ?

4 SMCMC IN CONTINUOUS TIME

[A.E., C. Marinelli 2009, 2010]

Aim : Sequential estimation / approximation of probability measures

$$\mu_t(x) \;\propto\; \exp\left(- \int_0^t H_s(x)\, ds \right)\; \mu(x)$$

on a finite state space S. W.l.o.g.

$$\langle H_s \,,\, \mu_s \rangle \;=\; 0\,.$$

Estimator :

$$\mu_t \;\approx\; \eta_t^N \;:=\; \frac{1}{N} \sum_{i=1}^N \delta_{X_t^i}$$

THE PARTICLE SYSTEM $(X_t^i)_{1\le i\le N}$ **:**

- Independent Markov chain moves with generator $\lambda_t \cdot \mathcal{L}_t$
- X_t^i replaced by X_t^j with rate $\frac{1}{N}(H_t(X_t^i) - H_t(X_t^j))^+$

i.e., $(X_t^1, \dots, X_t^N)$ is the Markov process on S^N with generator

$$\begin{aligned} \mathcal{L}_t^N \varphi(x_1, \dots, x_N) &= \lambda_t \sum_{i=1}^N \mathcal{L}_t^{(i)} \varphi(x_1, \dots, x_N) \\ &\quad + \frac{1}{N} \sum_{i,j=1}^N (H_t(x_i) - H_t(x_j))^+ \cdot \big(\varphi(x^{i\to j}) - \varphi(x)\big) \end{aligned}$$

$\lambda_t > 0$ constants, $\mathcal{L}_t$ generator of a Markov process on S satisfying

$$\mu_t(x)\mathcal{L}_t(x,y) = \mu_t(y)\mathcal{L}_t(y,x) \quad \text{detailed balance,}$$

$\mathcal{L}_t^{(i)}$ action of $\mathcal{L}_t$ on i th component.

SCALING LIMIT :

$$\frac{\partial}{\partial t}\mu_t = -H_t\,\mu_t$$
$$= \lambda_t \mathcal{L}_t^* \mu_t - H_t \mu_t + \langle H_t, \mu_t \rangle\, \mu_t$$

η_t^N is a discretization of this equation:

$$\frac{\partial}{\partial t}\mathbb{E}\left[\langle f, \eta_t^N \rangle\right] = \mathbb{E}\left[\langle f\,,\, \lambda_t \mathcal{L}_t^* \eta_t^N - H_t \eta_t^N + \langle H_t, \eta_t^N \rangle\, \eta_t^N \rangle\right]$$

LLN / Scaling limit:

$$\eta_t^N \approx \mathbb{E}[\eta_t^N] \approx \mu_t \qquad \text{for large} N.$$

5 QUANTITATIVE BOUNDS

- **LLN, CLT, EXPRESSION FOR ASYMPTOTIC VARIANCE:**
  (for closely related particle system)

  P. Del Moral, L. Miclo. Branching and Interacting Particle Systems Approx. of Feynman-Kac Formulae (2000)

  M. Rousset. On the control of an interacting particle approximation of Schrödinger ground states. *SIAM J. Math. An.* **38(3)** (2006)

- **CLTs IN DISCRETE TIME:**

  P. Del Moral. Feynman-Kac Formulae, Springer 2004

  N. Chopin. CLT for sequential SMC methods and its application to Bayesian inference. *Annals of Statistics* **32 (6)** (2004)

  H. R. Künsch. Recursive Monte Carlo Filters: Algorithms and Theoretical Analysis. *Annals of Statistics* **33(5):** 1983-2021, 2004.

PROBLEMS:

- Implicit expression for asymptotic variance
  $\rightsquigarrow$ *need L^p bounds for Feynman-Kac propagators*

  A.E., C. Marinelli. L^p estimates for Feynman-Kac propagators with time dependent reference measures, *J.Math.Anal.Appl. 2009*.

- Feasible bound for fixed number of particles ?
  Under global mixing conditions:

  A.E., C. Marinelli. Quantitative approximations of evolving probability measures and sequential MCMC methods, Preprint 2010.

AN UNBIASSED ESTIMATOR:

$$\nu_t^N \;:=\; \exp\left(-\int_0^t \langle H_s, \eta_s^N\rangle\right)\; \eta_t^N$$

THEOREM.

$$E\left[\langle f, \nu_t^N\rangle\right] \;=\; \langle f,\, \mu_t\rangle \qquad \forall\; t \geq 0,\;\; f : S \longrightarrow \mathbb{R}.$$

FEYNMAN-KAC PROPAGATOR:

Define $q_{s,t}f$ as solution of backward equation

$$\frac{\partial}{\partial s} q_{s,t}f = -\lambda_s \mathcal{L}_s q_{s,t}f - H_s q_{s,t}f, \qquad q_{t,t}f = f,$$

Feynman-Kac representation:

$$q_{s,t}f(x) = \mathbb{E}_{s,x}\Big[e^{-\int_s^t H_r(X_r)\,dr} f(X_t)\Big],$$

where $(X_t, \mathbb{P}_{s,x})$ is Markov process with gen. $\lambda_t \mathcal{L}_t$ and init. cond. $X_s = x$.

Fix $q > 6$ and $p \in \left(\frac{4q}{q-2},\, q\right)$.

THEOREM. Suppose that

$$N \geq \max(120 \cdot K_t,\ 80) \qquad \text{and}$$

$$\lambda_s \geq \max\left(\frac{p}{4}A_s + \frac{p(p+3)}{4}tB_s\ ,\ 35 \cdot \textsf{osc}(H_s) \cdot C_s\right) \qquad \forall s \in [0,t].$$

Then

$$\mathrm{Var}\ \left(\langle f, \nu_t^N\rangle\right)^{1/2} \leq \left(\mathrm{Var}_{\mu_t}(f) + V_t(f) + \|f\|^2_{L^p(\mu_t)}\right)^{1/2} N^{-1/2} + 40 \cdot K_t \cdot \|f\|_{L^p(\mu_t)}\ N^{-1}$$

where

$$V_t(f) = -\int_0^t \langle H_s(q_{s,t}f)^2,\ \mu_s\rangle + 2\int\int |H_s(x)|\left(q_{s,t}f(y) - q_{s,t}f(x)\right)^2 \mu_s(dx)\mu_s(dy) \leq 13 \cdot K_t \cdot \|f\|_{L^p(\mu_t)}$$

CONSTANTS:

$$K_t \;=\; \int_0^t \|H_s\|_{L^q(\mu_s)}\; ds$$

$$A_t \;=\; \sup_{\langle f, \mu_t\rangle = 0} \frac{\int H_t\, f^2\, d\mu_t}{\mathcal{E}_t(f,f)}$$ H–Poincaré constant

$$B_t \;=\; \sup_{\langle f, \mu_t\rangle = 0} \frac{\left|\int H_t f\, d\mu_t\right|^2}{\mathcal{E}_t(f,f)}$$ modified H–Poincaré

$$C_t \;=\; \sup_{\langle f^2, \mu_t\rangle = 1} \frac{\int f^2\, \log f^2\, d\mu_t}{\mathcal{E}_t(f,f)}$$ Log-Sobolev constant

COROLLARY. Similar estimates hold for

$$E\left[\left|\langle f, \eta_t^N\rangle - \langle f, \mu_t\rangle\right|\right]$$

EXAMPLE. (Moving Gaussians)

$\mu_t = N(m_t, \sigma_t^2)$ on $[-r, r]^d$, $\mathcal{L}_t =$ Ornstein-Uhlenbeck generator

Quantitative bounds depending on $\dot{m}_t/m_t$ and $\dot{\sigma}_t/\sigma_t$.

6 OUTLOOK

OPEN PROBLEMS:

- Generalization to discrete time and continuous space

  see PhD thesis of Nikolaus Schweizer

- Non-asymptotic bounds under **local** mixing conditions.

  First step: Asymptotic bounds:

  A.E., C. Marinelli. Stability of nonlinear flows of probability measures related to sequential MCMC methods.

  Second step: Non-asymptotic analysis on trees:

  PhD thesis of Nikolaus Schweizer

“Recipes” for applications

- Try to guarantee osc$(H_t) \leq 1$, or at least osc$(H_t^-) \leq 1$
- Use enough MCMC steps such that there is sufficient mixing in each metastable state
- Quality of error estimates depends (among other things) on structure of disconnectivity tree