Free energy techniques in Bayesian Statistics

Gabriel STOLTZ

stoltz@cermics.enpc.fr

(CERMICS, Ecole des Ponts & MICMAC team, INRIA Rocquencourt)

CECAM workshop, June 2012

Outline

Sampling a mixture model in Bayesian statistics

- Description of the model
- Choosing a reaction coordinate
- Free energy biased sampling

Convergence and efficicency of the Wang-Landau algorithm

- Description of our flavor of the Wang-Landau algorithm
- Convergence of the algorithm
- Assessing an improved convergence rate

[CLS12] N. Chopin, T. Lelièvre and G. Stoltz, *Statist. Comput.*, 2012
[FJLS12] G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre and G. Stoltz, in preparation

Sampling a mixture model in Bayesian statistics

Description of the mixture model

- Data set $\{y_i\}_{i=1,\dots,N_{\rm data}}$ approximated by mixture of K Gaussians

$$f(y\,|\,\theta) = \sum_{i=1}^{K} q_i \sqrt{\frac{\lambda_i}{2\pi}} \exp\left(-\frac{\lambda_i}{2}(y-\mu_i)^2\right)$$

- Parameters $\theta = (q_1, \dots, q_{K-1}, \mu_1, \dots, \mu_K, \lambda_1, \dots, \lambda_K)$ with

$$\mu_i \in \mathbb{R}, \quad \lambda_i \geqslant 0, \quad 0 \leqslant q_i \leqslant 1, \quad \sum_{i=1}^{K-1} q_i \leqslant 1$$

- Prior distribution $p(\theta)$: Random beta model

Aim

Find the values of the parameters (namely θ, and possibly K as well) describing correctly the data

[RG97] S. Richardson and P. J. Green. *J. Roy. Stat. Soc. B*, 1997.
[JHS05] A. Jasra, C. Holmes and D. Stephens, Statist. Science, 2005

Target distribution

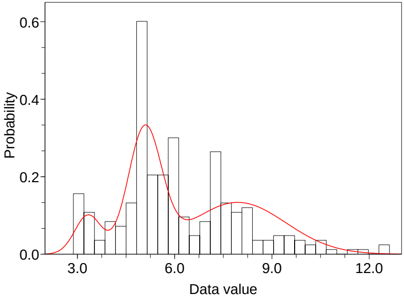
Prior distribution: additional variable $\beta \sim \Gamma(g, h)$

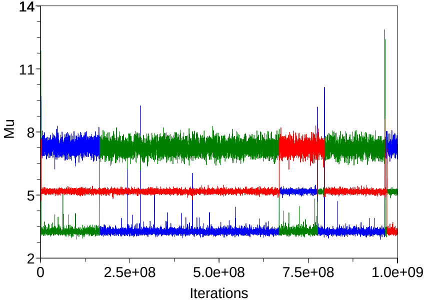
- uniform distribution of the weights q_i
- $\mu_k \sim \mathcal{N}\left(M, \frac{R^2}{4}\right)$ with $M =$ mean of data, $R = \max - \min$
- $\lambda_k \sim \Gamma(\alpha, \beta)$ with $g = 0.2$ and $h = 100g/\alpha R^2$

Posterior density $\pi(\theta) = \dfrac{1}{Z_K} p(\theta) \displaystyle\prod_{i=1}^{N_{\mathrm{data}}} f(y_i \,|\, \theta)$

- Initial conditions: equal weights, means and variances for the Gaussians
- Metropolis random walk with (anisotropic) Gaussian proposals
- Metastability: at least $K! - 1$ symmetric replicates of any mode, but there may be additional metastable states
- Metastability increased when N_{data} increases

Fish data



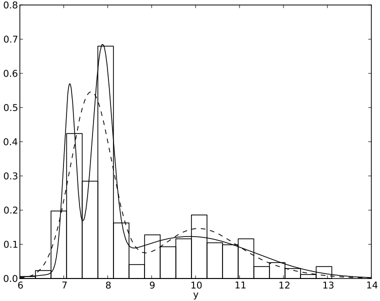


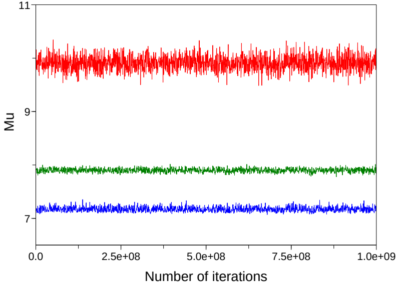
Left: Lengths of snappers ($N_{\rm data} = 256$), and a possible fit for $K = 3$ using the last configuration from the trajectory plotted in the right picture.

Right: Typical sampling trajectory, gaussian random walk with $(\sigma_q, \sigma_\mu, \sigma_v, \sigma_\beta) = (0.0005, 0.025, 0.05, 0.005)$.

[IS88] A. J. Izenman and C. J. Sommer, *J. Am. Stat. Assoc.*, 1988.
[BMY97] K. Basford *et al.*, *J. Appl. Stat.*, 1997

Hidalgo stamp data





Left: Thickness of Mexican stamps ($N_{\mathrm{data}} = 485$), and two possible fits for $K = 3$ ("genuine multimodality", solid line: dominant mode).

Right: Typical sampling trajectory, gaussian random walk with $(\sigma_{\boldsymbol{q}}, \sigma_{\mu}, \sigma_{\boldsymbol{v}}, \sigma_{\beta}) = (0.001, 0.05, 0.1, 0.005)$.

[TSM86] D. Titterington *et al.*, *Statistical Analysis of Finite Mixture Distributions*, 1986.
[FS06] S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching Models*, 2006.

Numerical procedure

Computation of the bias for a given reaction coordinate:

- Metropolis dynamics with Gaussian proposals
- Choose a reaction coordinate ξ
- ABF for q, μ, β, Self-Healing Umbrella Sampling for V
- Output: approximation $A(z)$ of the free energy associated with ξ

Biased dynamics:

- Sample $\pi_A(\theta) \propto \pi(\theta)\, \mathrm{e}^{A(\xi(\theta))}$
- Cauchy proposals with $\tau_\mu = R/1000$, $\tau_v = 2/R^2$, $\tau_\beta = 2\alpha R^2 \times 10^{-5}$

Reweighting procedure: $\mathbb{E}_\pi(\varphi) = \dfrac{\mathbb{E}_{\pi_A}\Big(\varphi \exp\{-A \circ \xi\}\Big)}{\mathbb{E}_{\pi_A}\Big(\exp\{-A \circ \xi\}\Big)}$

[DP01] E. Darve and A. Pohorille, *J. Chem. Phys.*, 2001
[HC04] J. Hénin and C. Chipot, *J. Chem. Phys.*, 2004
[MBCPS06] S. Marsili *et al.*, *J. Phys. Chem. B*, 2006

Adpative dynamics

Decompose the state space using slabs $\{\theta \, : \, \xi(\theta) \in (z_i, z_{i+1})\}$. Perfect convergence (statistical error, Δz,...) not required because of reweighting

Self-Healing Umbrella sampling: parameter free, no derivative needed. For $z \in (z_i, z_{i+1})$,

$$\exp\{-A_t(z)\} = \frac{1}{Z_t} \left(1 + \sum_{j=1}^{t-1} \mathbf{1}_{\{z_i \leqslant \xi(\theta_j) < z_{i+1}\}} \exp\left[- A_j \circ \xi(\theta_j) \right] \right),$$

Adaptive Biasing Force: discrete integration of the approximate mean force (hence smoother potential)

$$F_t(z) = \frac{\displaystyle \sum_{j=1}^{t-1} f(\theta_j) \, \mathbf{1}_{\{z_i \leqslant \xi(\theta_j) \leqslant z_{i+1}\}}}{\displaystyle \sum_{j=1}^{t-1} \mathbf{1}_{\{z_i \leqslant \xi(\theta_j) \leqslant z_{i+1}\}}}$$

Choice of the reaction coordinate

- How fast does the approximate free energy A_t converge to A?
- How efficient is the importance sampling/reweighting?

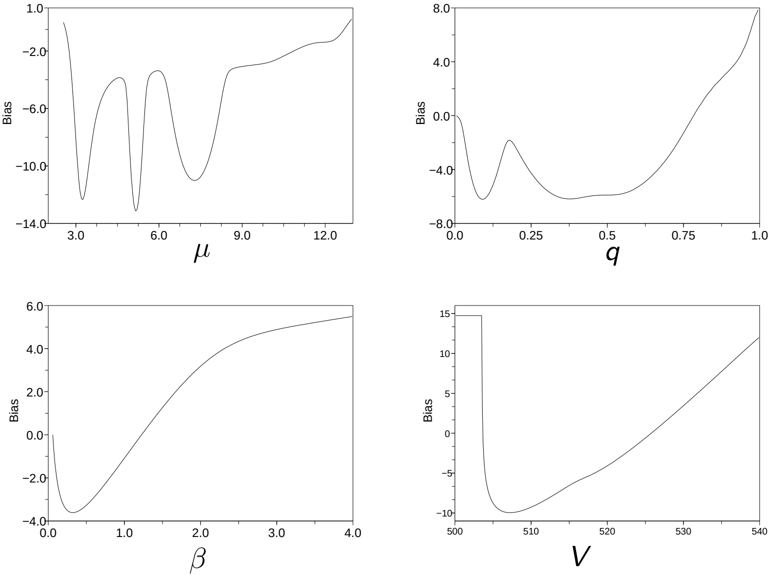
(1) How efficient is the MCMC sampling of the biased density?

(2) How representative are the points simulated from the biased distribution? (non-negligible weights)

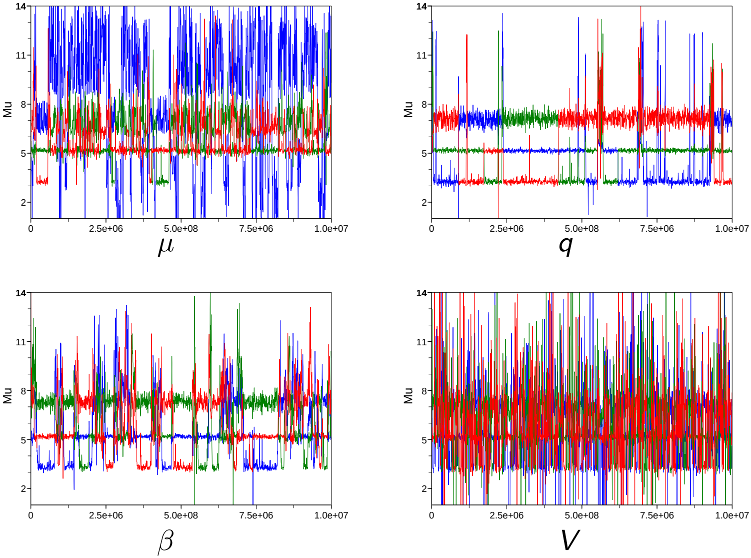
$$\mathrm{EF} = \frac{\left(\displaystyle\sum_{n=1}^{N} w(\theta_n)\right)^2}{T \displaystyle\sum_{n=1}^{N} w(\theta_n)^2} \simeq \frac{\left(\displaystyle\int_{z_{\min}}^{z_{\max}} \exp\left\{-\widehat{A}(z)\right\} dz\right)^2}{(z_{\max} - z_{\min}) \displaystyle\int_{z_{\min}}^{z_{\max}} \exp\left\{-2\widehat{A}(z)\right\} dz}$$

- How difficult is it to determine, a priori, an interval for the reaction coordinate values?

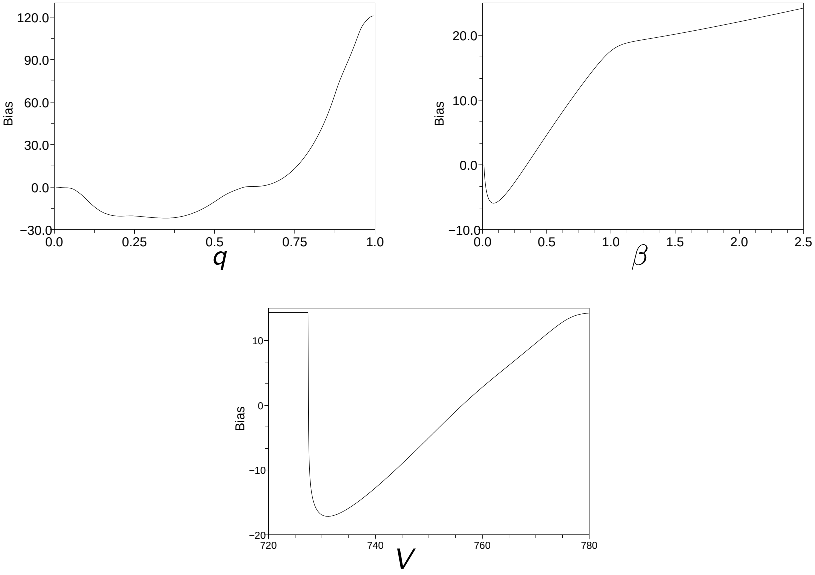
Free energies for various reaction coordinates (Fishery)



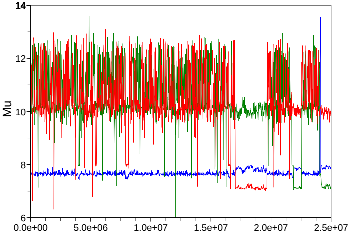
Trajectories of (μ_1, μ_2, μ_3), biased dynamics (Fishery)



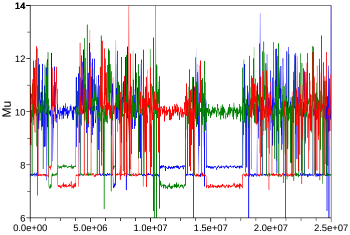
Free energies for various reaction coordinates (Hidalgo)



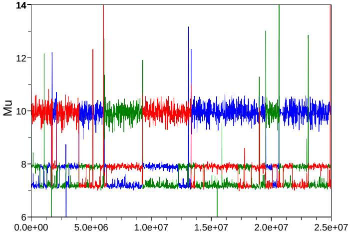
Trajectories of (μ_1, μ_2, μ_3), biased dynamics (Hidalgo)



q



β



V

Efficiency factors

Fishery data, $K = 3$

Reaction coordinate	β	potential	q_1	μ_1
EF (numerical)	0.17	0.16	0.48	0.04
EF (theoretical)	0.179	0.178	0.454	0.079

Fishery data, $\xi = \beta$

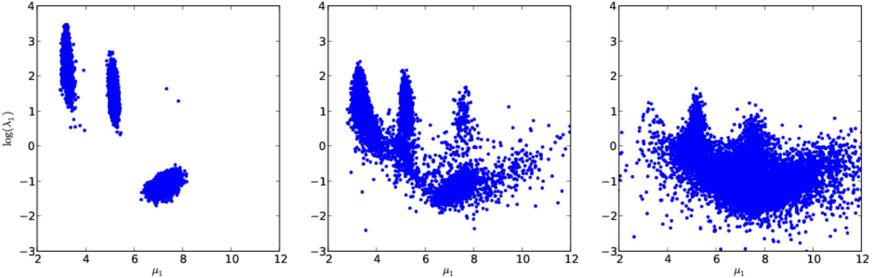
K	3	4	5	6
EF (numerical)	0.17	0.18	0.17	0.16
EF (theoretical)	0.179	0.195	0.180	0.171

Hidalgo, $K = 3$

Reaction coordinate	β	potential	q_1
EF (numerical)	0.02	0.24	0.23
EF (theoretical)	0.06	0.13	0.18

Why β works

Suggested range (statistical arguments): β is a small fraction of R^2, *e.g.* $z_{\min} = R^2/2000$ and $z_{\max} = R^2/20$



Simulated pairs $(\mu_1, \log\lambda_1)$ conditional $\beta \in [0, 0.5]$, $[1.5, 2]$ and $[3.5, 4]$

When β is large, large variances of the modes are favored by the prior distribution (density $\lambda^{\alpha-1}\exp(-\beta\lambda)$ and variance $v = \lambda^{-1}$)
$\rightarrow$ *the modes cover the full range of data and switchings are made easier*

Convergence of the Wang-Landau algorithm

Description of the Wang-Landau algorithm (1)

- **Partitioning** of the space X intro subsets X_i with weights

$$\theta_\star(i) \stackrel{\text{def}}{=} \int_{\mathsf{X}_i} \pi(x)\, dx$$

Typically, $\mathsf{X}_i = \xi^{-1}\Big([\alpha_{i-1}, \alpha_i)\Big)$ and $\pi(x) = \mathrm{e}^{-U(x)}$

- **Importance sampling** to reduce metastability issues: biased measure

$$\pi_\theta(x) = \left(\sum_{i=1}^d \frac{\theta_\star(i)}{\theta(i)} \right)^{-1} \sum_{i=1}^d \frac{\pi(x)}{\theta(i)} \, \mathbb{1}_{\mathsf{X}_i}(x)$$

for any $\theta \in \Theta = \left\{ \theta = (\theta(1), \cdots, \theta(d)) \;\middle|\; 0 < \theta(i) < 1, \ \sum_{i=1}^d \theta(i) = 1 \right\}$

[WL01] F. Wang and D. Landau, *Phys. Rev. Lett.* & *Phys. Rev. E*, 2001

Description of the Wang-Landau algorithm (2)

Linearized WL in the stochastic approximation setting

Given $X_0 \in \mathrm{X}$ and weights $\theta_0 \in \Theta$ (typically $\theta_0(i) = 1/d$),

(1) draw X_{n+1} from conditional distribution $P_{\theta_n}(X_n, \cdot)$ (Metropolis);

(2) assume that $X_{n+1} \in \mathrm{X}_i$. The weights are then updated as

$$\begin{cases} \theta_{n+1}(i) = \theta_n(i) + \gamma_{n+1}\ \theta_n(i)\,(1-\theta_n(i)) \\ \theta_{n+1}(k) = \theta_n(k) - \gamma_{n+1}\ \theta_n(k)\ \theta_n(i) \qquad \text{for } k \neq i. \end{cases} \tag{1}$$

Comparison with original Wang-Landau algorithm

- deterministic step-sizes γ_n, to be chosen appropriately
- no "flat histogram" criterion
- linearized weight update $\theta_{n+1}(i) = \theta_n(i) \dfrac{1 + \gamma_{n+1} \mathbb{1}_{I(X_{n+1})=i}}{1 + \sum_{j=1}^{d} \gamma_{n+1} \theta_n(j) \mathbb{1}_{I(X_{n+1})=i}}$

[AL10] Y. Atchade and J. Liu, *Stat. Sinica*, 2010
[Liang05] F. Liang, *J. Am. Stat. Assoc.*, 2005

Stochastic approximation framework

SSA reformulation

Define $\eta_{n+1} = H(X_{n+1}, \theta_n) - h(\theta_n)$ and $h(\theta) = \int_{\mathcal{X}} H(x,\theta)\, \pi_\theta(x)\, dx$. Then,

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\, h(\theta_n) + \gamma_{n+1}\eta_{n+1}.$$

Here, $H_i(x,\theta) = \theta(i)\left(\mathbb{1}_{\mathcal{X}_i}(x) - \theta(I(x))\right)$ and $h(\theta) = \left(\sum_{i=1}^{d} \frac{\theta_\star(i)}{\theta(i)}\right)^{-1} (\theta_\star - \theta)$

Idea of proofs:

- η_n is a "small, random" perturbation
- the mean-field function h ensures the convergence to $\theta_\star$ in the absence of noise: there is a Lyapunov function V such that $\langle \nabla V, h \rangle < 0$ when $\theta \neq \theta_\star$
- conditions on the step-sizes

Assumptions

• The density π with respect to the measure λ is such that $\sup_{\mathsf{X}} \pi < \infty$ and $\inf_{\mathsf{X}} \pi > 0$. In addition, $\theta_\star(i) > 0$.

• For any $\theta \in \Theta$, P_θ is a Metropolis-Hastings dynamics with invariant distribution π_θ and symmetric proposal distribution with density $q(x,y)$ satisfying $\inf_{\mathsf{X}^2} q > 0$.

• the sequence $(\gamma_n)_{n \geqslant 1}$ is a non-negative determinstic sequence such that

(a) $(\gamma_n)_n$ is a non-increasing sequence converging to 0;

(b) $\sup_n \gamma_n \leqslant 1$;

(c) $\sum_n \gamma_n = \infty$;

(d) $\sum_n \gamma_n^2 < \infty$;

(e) $\sum_n |\gamma_n - \gamma_{n-1}| < \infty$.

Examples of acceptable step-sizes: $\gamma_n = \dfrac{\gamma_*}{n^\alpha}$ with $\alpha \in (1/2, 1]$

Convergence of the Wang-Landau algorithm

The aim is to apply general convergence results in SSA.

Weak stability result

The weight sequence almost surely comes back to a compact subset of Θ

$$\limsup_{n\to\infty} \left(\min_{1\leqslant j\leqslant d} \theta_n(j) \right) > 0 \quad \text{a.s.}$$

Convergence result

The sequence $\{\theta_n\}$ almost surely converges to $\theta_\star$, and

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \xrightarrow{\text{a.s.}} \int f(x)\, \pi_{\theta_\star}(x)\, dx$$

Various ways to recover averages with respect to π (instead of $\pi_\star$).

[AMP05] C. Andrieu, E. Moulines, and P. Priouret, *SIAM J. Control Opt.*, 2005.

Efficiency of the Wang-Landau algorithm?

Various to quantify the efficiency

- convergence rate (asymptotic variance)
- exit times out of metastable states
- references: unbiased dynamics and dynamics at $\theta = \theta_\star$ fixed

Asymptotic variance:

- additional assumptions on the step-sizes (satisfied for $\gamma_n \sim \gamma_\star / n^\alpha$ when $\alpha \in (1/2, 1]$ or $\gamma_n = \gamma_* / n$ with $\gamma_* > d/2$)
- comparison with "ideal" dynamics $Y_{n+1} \sim P_{\theta_\star}(Y_n, \cdot)$ and

$$\widetilde{\theta}_{n+1} = \widetilde{\theta}_n + \gamma_{n+1} H(Y_n, \theta_\star),$$

- the sequences θ_n and $\widetilde{\theta}_n$ have the same asymptotic variances, which are of order $\mathrm{O}(\gamma_n)$
- with averaging: variance of order $1/n$ in all cases

First exit times: toy analytical example

Description of the dynamics

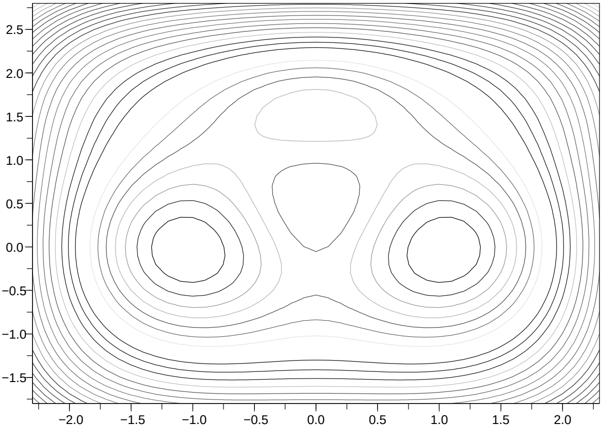
- three state model with $\theta_2^* = \dfrac{\varepsilon}{2-\varepsilon}$ and $\theta_1^* = \theta_3^* = \dfrac{1-\varepsilon}{2-\varepsilon}$
- in typical applications, $\varepsilon \sim \exp(\beta E_0)$
- proposal: equilikely jumps to nearest neighbor only ($1 \to 2$, $2 \to \{1,3\}$ and $3 \to 2$)

Scaling of the exit times

- first exit times $T_{1\to 3} = \min\left\{ n \,:\, I_n = 3 \text{ starting from } I_0 = 1 \right\}$
- non-adpative dynamics: $T_{1\to 3} \sim \dfrac{1}{\varepsilon}$
- Adaptive dynamics: $T_{1\to 3} \sim \begin{cases} |\ln \varepsilon|^{1/(1-\alpha)} & \text{for } \alpha \in (0,1) \\ \varepsilon^{-1/(1+\gamma_*)} & \text{for } \alpha = 1 \end{cases}$

First exit times: numerical results (1)

Entropic switch, Metropolis dynamics with isotropic Gaussian proposals



[PSLS03] S. Park, M. K. Sener, D. Lu, and K. Schulten, *J. Chem. Phys.*, 2003

First exit times: numerical results (2)

Table: Exponents of the scaling law $T_\beta \sim C_\alpha \beta^{\mu_\alpha}$ for $\gamma_n = n^{-\alpha}$ with $0 < \alpha < 1$.

α	μ_α	theoretical
0.125	1.11	1.14
0.25	1.30	1.33
0.375	1.55	1.60
0.5	2.02	2
0.625	2.72	2.67
0.75	4.06	4

Table: Exponents of the scaling law $T_\beta \sim \exp(\mu_{\gamma_*}\beta)$ for $\gamma_n = \gamma_*/n$.

γ_*	μ_{γ_*}	μ_{γ_*}/μ_0
0	2.32	1
1	1.74	0.75
2	1.51	0.65
4	1.25	0.54
8	0.92	0.40

Non-adaptive dynamics: $\alpha = 1$ and $\gamma_* = 0$

Conclusion: adaptive dynamics allow to go from exponential scalings of the exit times to power-law scalings.