Nested Stochastic Simulation Algorithm for KMC with Multiple Time-Scales

Eric Vanden-Eijnden

Courant Institute

Join work with Weinan E and Di Liu

Chemical Systems:

Evolution of an isothermal, spatially homogeneous mixture of chemically reacting molecules contained in a fixed volume V.

N_S species of molecules S_i, $i = 1, \ldots, N_S$ involved in
M_R reactions R_j, $j = 1, \ldots, M_R$.

Let x_i be the number of molecules of species S_i.

Each reaction R_j is characterized by a rate function $a_j(x)$ and a state change (or stochiometric) vector ν_j:

$$R_j = (a_j, \nu_j), \qquad R = \{R_1, \ldots, R_{M_R}\}.$$

Continuous-time Markov Chain:

Given state $x = (x_1, \ldots, x_{N_S})$, *the occurrences of the reactions on an infinitesimal time interval* dt *are independent of each other and the probability for reaction* R_j *to happen during this time interval is given by* $a_j(x)dt$. *The state of the system after reaction* R_j *is* $x + \nu_j$.

Equivalently: Given that the state of the system is $X_t = x$ at time t;

1. The probability that the next reaction happens after time $t+s$ is $e^{-a(x)s}$ where $a(x) = \sum_{j=1}^{M_R} a_j(x)$.

2. Given that a reaction happens at time $t + s$, the probability that it be reaction j is $a_j(x)/a(x)$.

Gillespie's Stochastic Simulation Algorithm (SSA, aka BKL):

Assume that the system is at state X_n at time t_n, then:

1. Generate two independent random numbers r_1 and r_2 with uniform distribution on the unit interval $(0, 1]$. Let

$$\delta t_{n+1} = -\frac{\ln r_1}{a(X_n)},$$

and k_{n+1} be the natural number such that ($a_0 = 0$)

$$\frac{1}{a(X_n)} \sum_{j=0}^{k_{n+1}-1} a_j(X_n) < r_2 \le \frac{1}{a(X_n)} \sum_{j=0}^{k_{n+1}} a_j(X_n),$$

2. Update the time and the state of the system by

$$t_{n+1} = t_n + \delta t_{n+1}, \qquad X_{n+1} = X_n + \nu_{k_{n+1}}.$$

Then repeat.

Exact realization of the process X_t

Systems with two disparate time scales.:

Assume that the reactions can be grouped as

$$R^s = \{(a^s(x), \nu^s)\}, \qquad \text{(Slow reactions)}$$
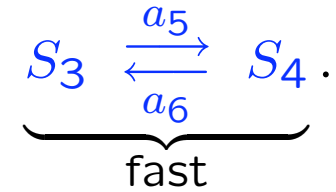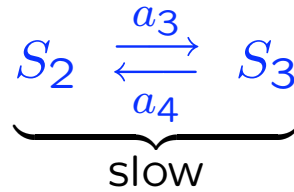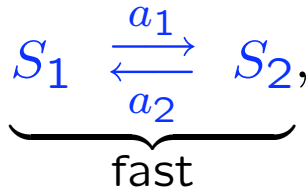$$R^f = \{(\varepsilon^{-1}a^f(x), \nu^f)\} \quad \text{(Fast reactions)}$$

where $\varepsilon \ll 1$ represents the ratio of time scales of the system.

Then: *The time-step between reactions is $O(\varepsilon)$ and with probability $1 - O(\varepsilon)$ a fast reaction happens.*

*Difficult to simulate the evolution up to
the $O(1)$ time-scale of the slow reactions!*

Several recent works on this topic e.g. by E. L. Haseltine and J. B. Rawlings, *J. of Chem. Phys.*, **117** 6959 (2002); C. V. Rao and A. P. Akin *J. of Chem. Phy.* **118**, 4999–5010 (2003); Y. Cao, D. Gillespie, and L. Petzold, *J. Chem. Phys.* **122**, 014116 (2005); A. Samant and D. G. Vlachos *J. Chem. Phys.* **123**, 144114 (2005).

Simple example:

$$S_1 \underset{a_2}{\overset{a_1}{\rightleftarrows}} S_2, \qquad\qquad S_2 \underset{a_4}{\overset{a_3}{\rightleftarrows}} S_3 \qquad\qquad S_3 \underset{a_6}{\overset{a_5}{\rightleftarrows}} S_4.$$

$$\underbrace{}_{\text{fast}} \qquad\qquad \underbrace{}_{\text{slow}} \qquad\qquad \underbrace{}_{\text{fast}}$$

with

$$a_1 = 10^5 x_1, \qquad \nu_1 = (-1, +1, \ 0, \ 0),$$
$$a_2 = 10^5 x_2, \qquad \nu_2 = (+1, -1, \ 0, \ 0),$$
$$a_3 = x_2, \qquad\quad\ \nu_3 = (\ 0, -1, +1, \ 0),$$
$$a_4 = x_3, \qquad\quad\ \nu_4 = (\ 0, +1, -1, \ 0),$$
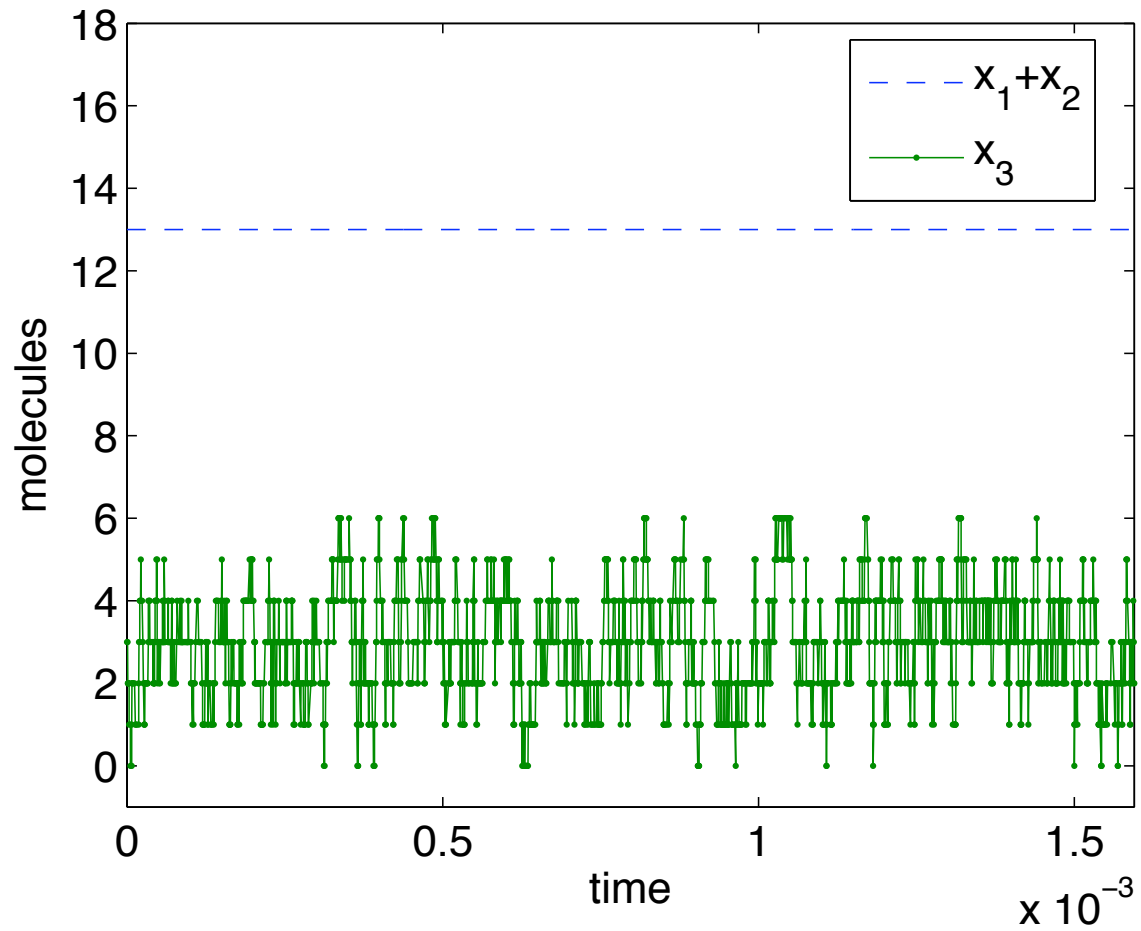$$a_5 = 10^5 x_3, \qquad \nu_5 = (\ 0, \ 0, -1, +1),$$
$$a_6 = 10^5 x_4, \qquad \nu_6 = (\ 0, \ 0, +1, -1).$$

i.e. the first and third reactions are faster than the second one.

Every species is involved in at least one fast reaction so there is no slow species.

But the variables $y_1 = x_1 + x_2$ and $y_2 = x_3 + x_4$ are conserved during the fast reactions and only evolve during the slow reaction.

Evolution of slow variable $y_1 = x_1 + x_2$ and
fast variable x_3 on the intermediate time scale.

Solution: Nested Stochastic Simulation Algorithm:

Inner SSA: Run N independent replicas of SSA with the fast reactions $R^f = \{(\varepsilon^{-1}a^f, \nu^f)\}$ only, for a time interval of $T_0 + T_f$. During this calculation, compute the modified slow rates from

$$\tilde{a}^s_j = \frac{1}{N} \sum_{k=1}^N \frac{1}{T_f} \int_{T_0}^{T_f + T_0} a^s_j(X^k_\tau) d\tau, \qquad j = 1, \cdots, M_{R_S}$$

where X^k_τ is the result of the k-th replica of this auxiliary virtual fast process at virtual time τ whose initial value is $X^k_{t=0} = X_n$.

Outer SSA: Run one step of SSA for the modified slow reactions $\tilde{R}^s = (\tilde{a}^s, \nu^s)$ to generate (t_{n+1}, X_{n+1}) from (t_n, X_n).

Then repeat.

Totally seamless!

Justification: Averaging theorem for Markov chains.

Consider the $u(x,t) = \mathbb{E}_x f(X_t)$, where X_t is the state variable at time t, and $\mathbb{E}_x$ denotes expectation conditional on $X_{t=0} = x$.

Then

$$\sup_{0 \leq t \leq T} |u(x,t) - \bar{u}(x,t)| = O(\varepsilon)$$

Here $\bar{u}(x,t) = \mathbb{E}_x f(\bar{X}_t)$, where $\bar{X}_t$ is the process generated by:

$$\bar{R} = \left\{ \left(\bar{a}_j(x) = \sum_{x'} a_j^s(x') \mu_{y(x)}(x'), \nu_j^s \right) \right\}_{j=1,...,M_{R_S}}$$

where $\mu_{y(x)}(x')$ is the equilibrium distribution of the *virtual fast process*, i.e. the original process where only the fast reactions are kept and $X_{t=0} = x$.

$u(x,t) = \mathbb{E}_x f(X_t)$ satisfies the *backward Kolomogorov equation*:

$$\frac{\partial u(x,t)}{\partial t} = \sum_{j=1}^{M_R} a_j(x) \left(u(x + \nu_j, t) - u(x,t) \right)$$

$$\equiv (Lu)(x,t) = \varepsilon^{-1}(L_1 u)(x,t) + (L_0 u)(x,t)$$

Singular perturbation analysis of this equation (Khasminskii,...).

Key observations: (i) the slow variables coincide with the maximum ergodic components of the process associated with L_1; (ii) they are given by
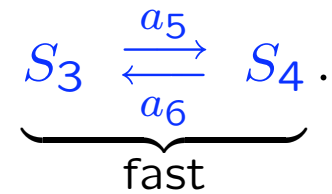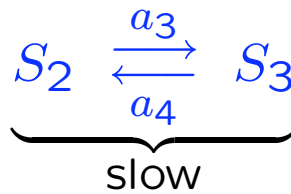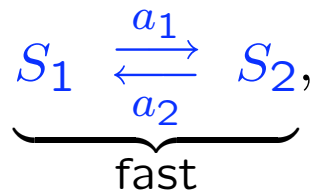
$$y_k = b_k \cdot x \quad \text{for} \quad k = 1, \ldots, K,$$

where $\{b_1, \ldots, b_K\}$ form a basis of the linear subspace such that

$$b \cdot \nu_j^f = 0 \quad \text{for all } \nu_j^f;$$

and (iii) if x and x' belong to the same ergodic component, then $x + \nu_j^s$ and $x' + \nu_j^s$ also belong to the same ergodic component for all ν_j^s.

Simple example revisited

$$S_1 \overset{a_1}{\underset{a_2}{\rightleftarrows}} S_2, \qquad S_2 \overset{a_3}{\underset{a_4}{\rightleftarrows}} S_3 \qquad S_3 \overset{a_5}{\underset{a_6}{\rightleftarrows}} S_4 .$$

$$\underbrace{}_{\text{fast}} \qquad \underbrace{}_{\text{slow}} \qquad \underbrace{}_{\text{fast}}$$

Slow variables:

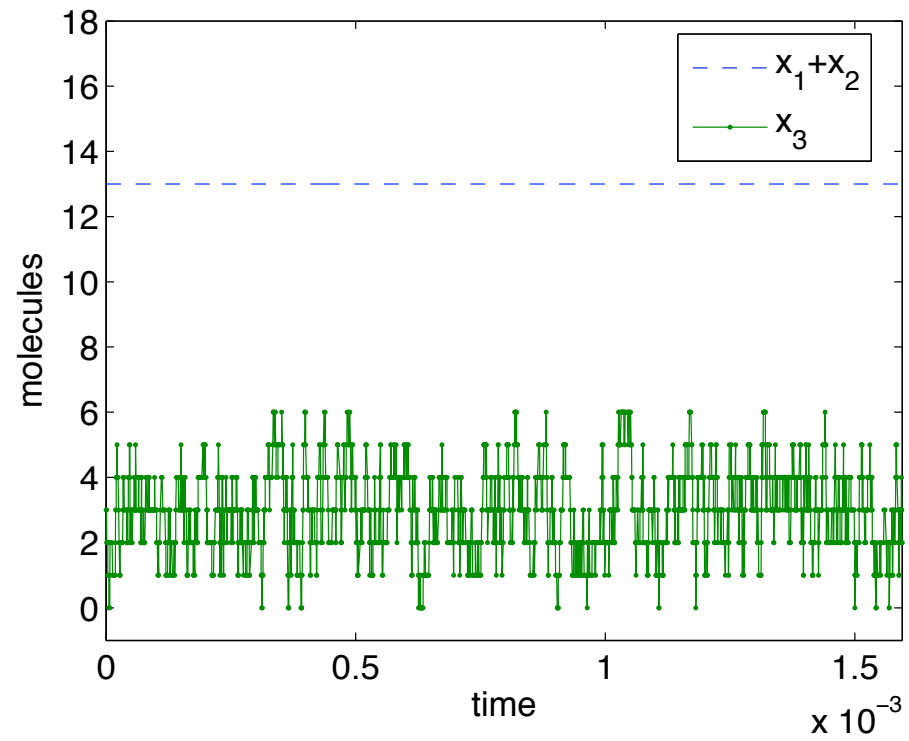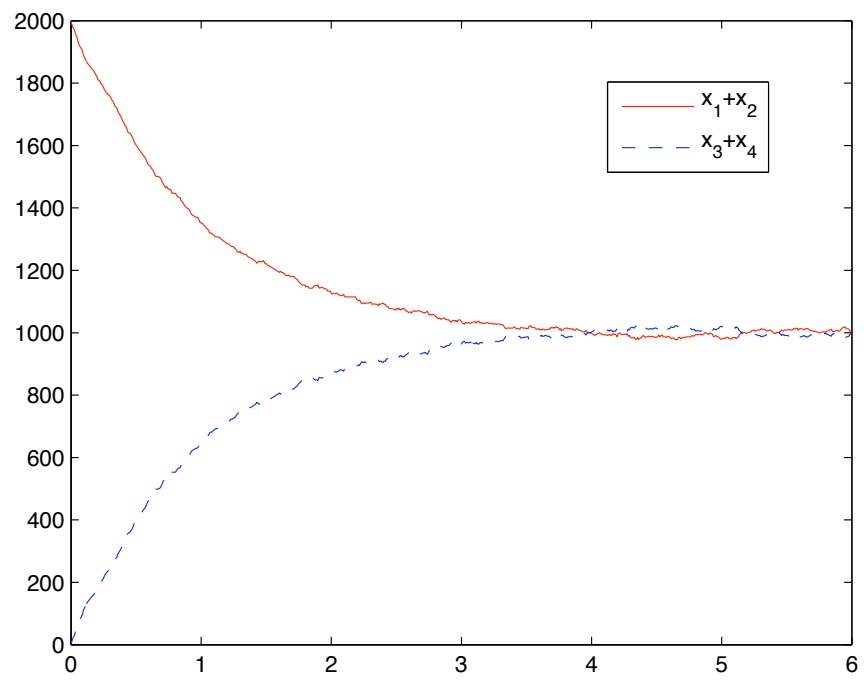$$y_1 = x_1 + x_2, \qquad y_2 = x_3 + x_4$$

Equilibrium distribution of the virtual fast process:

$$\mu_{y_1, y_2}(x_1, x_2, x_3, x_4) = \frac{y_1! \; y_2!}{x_1! \; x_2! \; x_3! \; x_4!} (1/2)^{y_1} (1/2)^{y_2} \delta_{x_1 + x_2 = y_1} \delta_{x_3 + x_4 = y_2}.$$

Effective dynamics:

$$\bar{a}_3^s = P x_2 = \frac{x_1 + x_2}{2} = \frac{y_1}{2}, \qquad \bar{\nu}_3^s = (-1, +1),$$

$$\bar{a}_4^s = P x_3 = \frac{x_3 + x_4}{2} = \frac{y_2}{2}, \qquad \bar{\nu}_4^s = (+1, -1).$$

Error estimate:

For any $T > 0$, there exist constants C and α independent of (N, T_0, T_f) such that,

$$\sup_{0 \leq t \leq T} \mathbb{E}\left|v(x,t) - u(x,t)\right| \leq C\left(\varepsilon + \frac{e^{-\alpha T_0/\varepsilon}}{1 + T_f/\varepsilon} + \frac{1}{\sqrt{N(1 + T_f/\varepsilon)}}\right).$$

Efficiency:

Given an error tolerance λ:

$$\text{cost} = O(N(1 + T_0/\varepsilon + T_f/\varepsilon)) = O\left(\frac{1}{\lambda^2}\right) \qquad \text{(nested SSA)}$$

$$\text{cost} = O\left(\frac{1}{\varepsilon}\right) \qquad \text{(direct SSA)}$$

Example: Heat shock response of E. Coli

Mechanism of protection that the E. Coli bacteria uses to fight against denaturation (unfolding) of its constituent proteins induced by the increase of temperature.
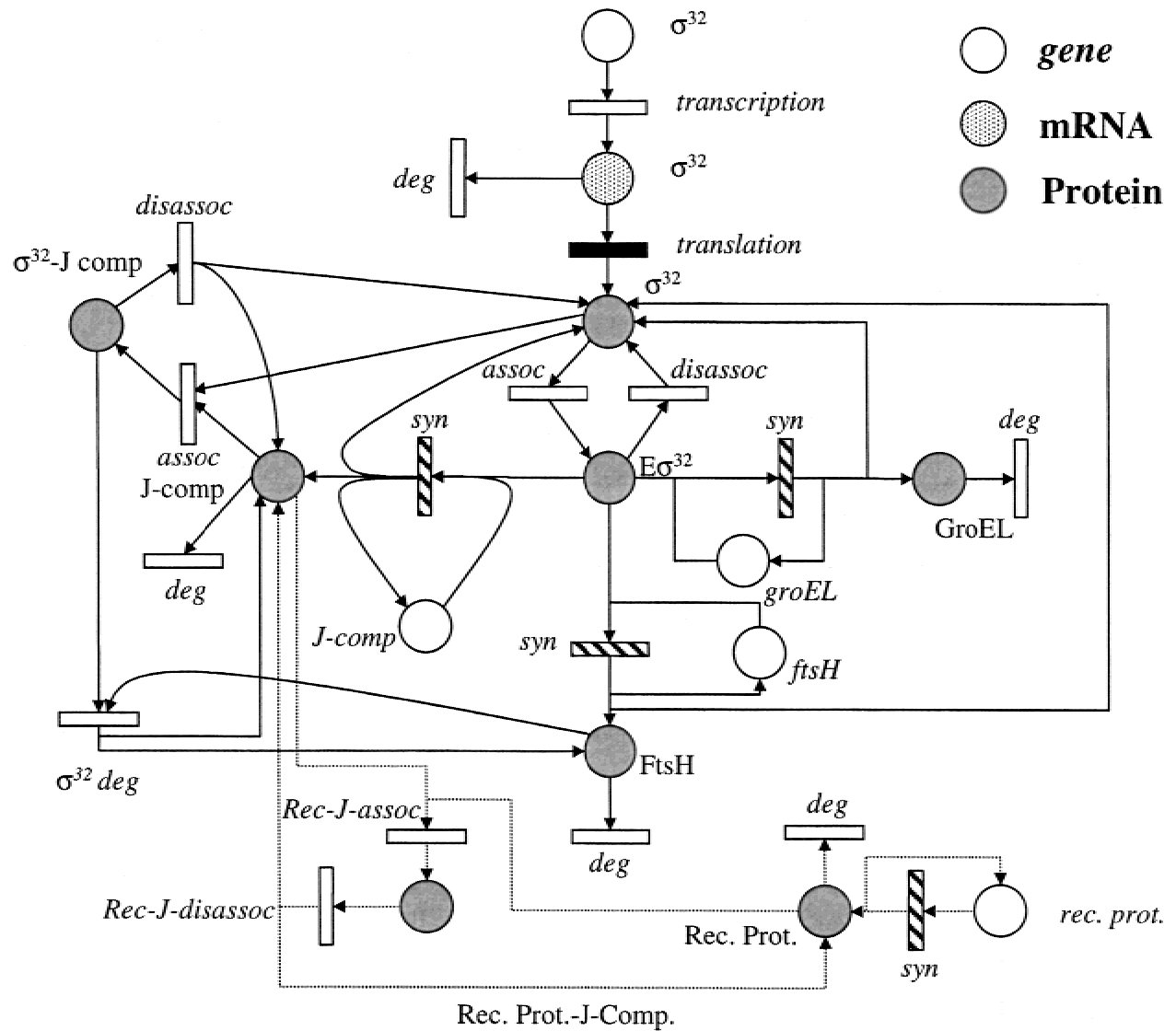
Stochastic petri net for heat shock response (from Srivastava *et al.*)

R. Srivastava, M. Peterson and W. Bently, *Biotech. Bioeng.* **75**, 120 (2001)

14 species,
17 reactions

Species	Initial value
DNA.σ^{32}	1
mRNA.σ^{32}	17
σ^{32}	15
RNAPσ^{32}	76
DNA.DnaJ	1
DNA.FtsH	0
DNA.GroEL	1
DnaJ	4640
FtsH	200
GroEL	4314
DnaJ.UnfoldedProtein	5×10^6
Protein	5×10^6
σ^{32}.DnaJ	2959
UnfoldedProtein	2×10^5

Reaction	Rate constant	Rates magnitude
DNA.$\sigma^{32} \rightarrow$ mRNA.σ^{32}	1.4×10^{-3}	1.4×10^{-3}
mRNA.$\sigma^{32} \rightarrow \sigma^{32} +$ mRNA.σ^{32}	0.07	1.19
mRNA.$\sigma_{32} \rightarrow$ degradation	1.4×10^{-6}	2.38×10^{-5}
$\sigma_{32} \rightarrow$ RNAPσ^{32}	0.7	10.5
RNAP$\sigma^{32} \rightarrow \sigma^{32}$	0.13	9.88
$\sigma^{32} +$ DnaJ $\rightarrow \sigma^{32}$.DnaJ $(\star\star)$	3.62×10^{-3}	25.2
DnaJ $\rightarrow$ degradation $(\star\star)$	6.4×10^{-10}	2.97×10^{-6}
σ^{32}.DnaJ$\rightarrow \sigma^{32} +$ DnaJ	4.4×10^{-4}	1.30
DNA.DnaJ + RNAP$\sigma^{32} \rightarrow$ DnaJ + DNA.DnaJ + σ^{32}	8	3.71
DNA.FtsH + RNAP.$\sigma^{32} \rightarrow$ FtsH + DNA.FtsH + σ^{32}	4.88×10^{-2}	0
FtsH $\rightarrow$ degradation	7.4×10^{-11}	1.48×10^{-8}
GroEL $\rightarrow$ degradation	1.8×10^{-8}	7.76×10^{-5}
σ^{32}.DnaJ + FtsH $\rightarrow$ DnaJ + FtsH	1.42×10^{-5}	8.4
DNA.GroEL + RNAP$\sigma^{32} \rightarrow$ GroEL + DNA.GroEL +σ^{32}	0.063	4.78
Protein $\rightarrow$ UnfoldedProtein $(\star)$	0.2	10^6
DnaJ+ UnfoldedProtein $\rightarrow$ DnaJ.UnfoldedProtein $(\star)$	0.108	10^7
DnaJ.UnfoldedProtein $\rightarrow$ DnaJ+ UnfoldedProtein $(\star)$	0.2	10^6

Nested SSA allows a speed-up by a factor of 100 in this example without any loss of accuracy
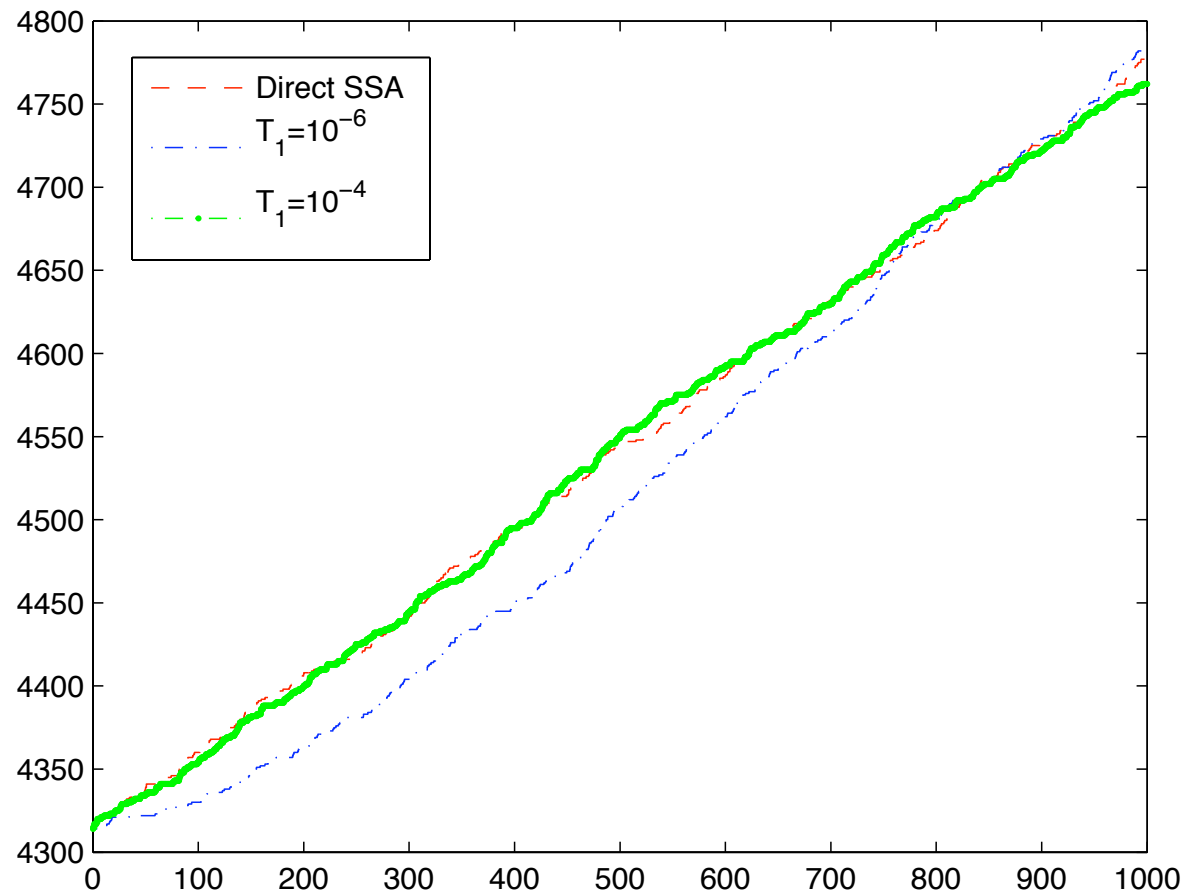
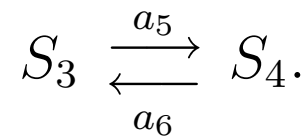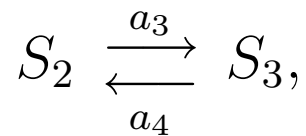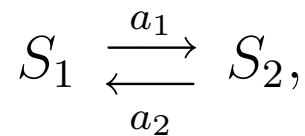Shown: Growth rate of GroEL (a protein measuring stress response)

$(N, T_f/10^{-6})$	$(1, 1)$	$(1, 4)$	$(1, 16)$	$(1, 64)$	$(1, 256)$	$(1, 1024)$
CPU	0.62	1.32	2.98	9.56	35.81	142.08
$\overline{\sigma^{32}}$	4.60	8.66	13.60	14.52	14.98	15.00
$\mathrm{var}(\sigma^{32})$	4.41	8.11	12.22	13.13	13.73	14.66

TABLE III: Efficiency of nested SSA when $N = 1$. Since we used $N_0 = 1000$ realizations of the Outer SSA to compute $\overline{\sigma^{32}}$ and $\mathrm{var}(\sigma^{32})$, the statistical errors on these quantities is about 0.1.

$(N, T_f/10^{-6})$	$(1, 1)$	$(2, 2)$	$(4, 4)$	$(8, 8)$	$(16, 16)$	$(32, 32)$
CPU	0.64	1.38	3.17	10.13	36.94	142.65
$\overline{\sigma_{32}}$	4.60	9.06	13.85	14.57	15.04	14.90
$\mathrm{var}(\sigma_{32})$	4.41	8.68	13.07	13.63	14.01	14.38

TABLE IV: Efficiency of nested SSA with multiple replicas in the Inner SSA. Again the statistical errors on $\overline{\sigma^{32}}$ and $\mathrm{var}(\sigma^{32})$ is about 0.1.

Generalization: Nested SSA with more than two levels:

$$S_1 \underset{a_2}{\overset{a_1}{\rightleftarrows}} S_2, \qquad S_2 \underset{a_4}{\overset{a_3}{\rightleftarrows}} S_3, \qquad S_3 \underset{a_6}{\overset{a_5}{\rightleftarrows}} S_4.$$

$a_1 = 2 \cdot 10^{10} x_1, \qquad \nu_1 = (-1, +1,\ 0,\ 0),$

$a_2 = 10^{10} x_2, \qquad \nu_2 = (+1, -1,\ 0,\ 0),$

$a_3 = 10^5 x_2, \qquad \nu_3 = (\ 0, -1, +1,\ 0),$

$a_4 = 2 \cdot 10^5 x_3, \qquad \nu_4 = (\ 0, +1, -1,\ 0),$

$a_5 = x_3\ , \qquad \nu_5 = (\ 0,\ 0, -1, +1),$

$a_6 = x_4, \qquad \nu_6 = (\ 0,\ 0, +1, -1).$
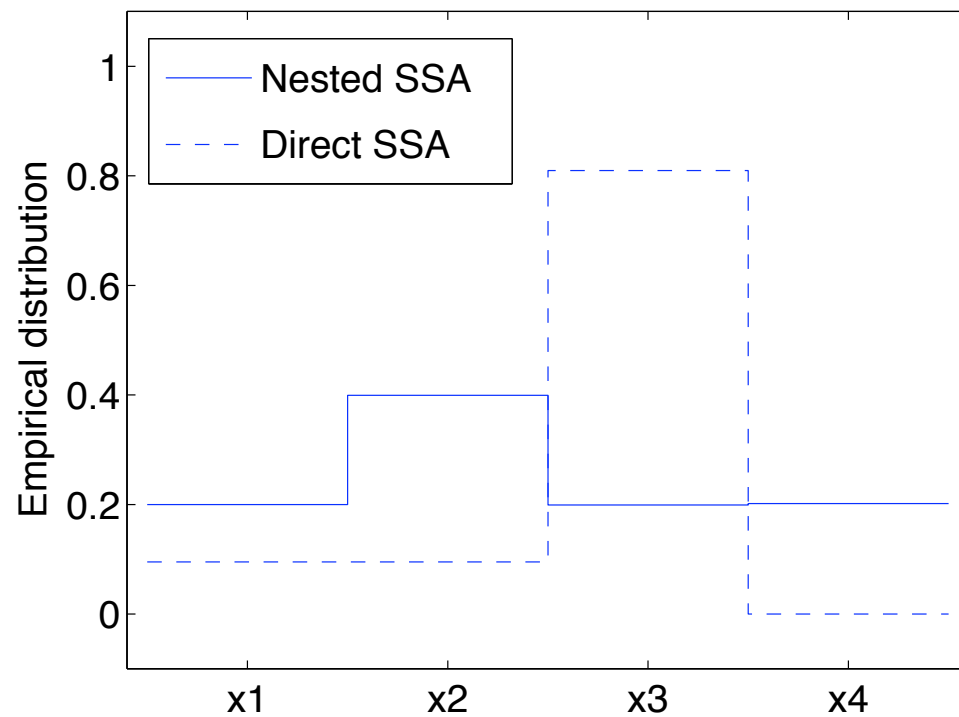
Generalization: Adaptive nested SSA:
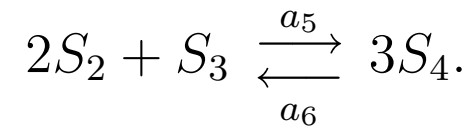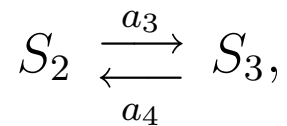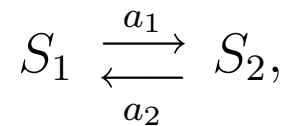
$$S_1 \underset{a_2}{\overset{a_1}{\rightleftarrows}} S_2, \qquad S_2 \underset{a_4}{\overset{a_3}{\rightleftarrows}} S_3, \qquad 2S_2 + S_3 \underset{a_6}{\overset{a_5}{\rightleftarrows}} 3S_4.$$

$$a_1 = x_1, \qquad\qquad\qquad \nu_1 = (-1, \ +1, \ 0, \ 0),$$
$$a_2 = x_2, \qquad\qquad\qquad \nu_2 = (+1, \ -1, \ 0, \ 0),$$
$$a_3 = 10^4 x_2, \qquad\qquad\quad \nu_3 = (0, \ -1, \ +1, \ 0),$$
$$a_4 = 10^4 x_3, \qquad\qquad\quad \nu_4 = (0, \ +1, \ -1, \ 0),$$
$$a_5 = 2x_2(x_2 - 1)x_3, \qquad \nu_5 = (\ 0, -2, -1, +3),$$
$$a_6 = 2x_4(x_4 - 1)(x_4 - 2), \qquad \nu_6 = (\ 0, +2, +1, -3).$$

Initial data: $\qquad (x_1, x_2, x_3, x_3) = (100, 3, 3, 3).$
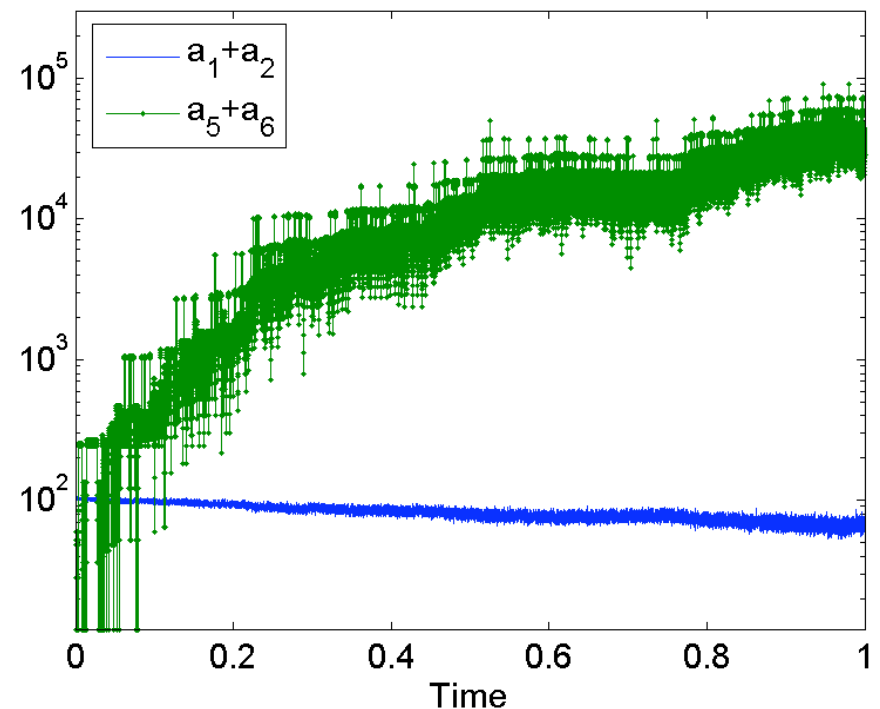
Direct SSA: $\overline{x_1} = 27.62 \pm 0.2,$ $\mathrm{var}(x_1) = 20.97 \pm 0.2.$

: 8781.83 seconds of CPU time

Adaptive Nested SSA:

$T_f/10^{-5}$	1	2	4	8	16	32	64
CPU	13.6	18.6	28.0	47.2	86.2	163.0	316.2
$\overline{x_1}$	27.50	27.55	27.44	27.51	27.55	27.55	27.61
$\mathrm{var}(x_1)$	20.58	20.65	20.82	20.57	20.84	20.58	21.01

Table 4

Efficiency and accuracy of the adaptive nested SSA.

Generalization to other types of KMC

Generator

$$(Lf)(x) = \sum_{j \in \mathcal{J}} \lambda_j(x)(f(x + z_j(x)) - f(x))$$

Ergodic components of fast process: for any test function f : $\mathcal{X} \to \mathbb{R}$, we have

$$\text{if } x \in \mathcal{X}_y \quad \text{then} \quad \lim_{T \to \infty} \frac{1}{T} \int_0^T \mathbb{E}_x f(X_t^f) dt = \sum_{x \in \mathcal{X}} \mu_y(x) f(x),$$

where $\mathbb{E}_x$ denotes the expectation conditional on $X_{t=0}^f = x$

Assumption: Slow reactions induce an injection on the ergodic components $\mathcal{X}_y$, i.e. for each e_j, there exists a unique $\bar{e}_j$ such that for all $x \in \mathcal{X}$,

$$\text{if } x \in \mathcal{X}_y \text{ then } x + e_j \in \mathcal{X}_{y_j} \text{ with } y_j = y + \bar{e}_j.$$