Proceedings of the Edinburgh Mathematical Society (2002) **45**, 701–716 ©
DOI:10.1017/S0013091500000882 Printed in the United Kingdom

METASTABILITY IN THE CLASSICAL, TRUNCATED BECKER–DÖRING EQUATIONS

DUGALD B. DUNCAN[1] AND RACHEL M. DUNWELL[2]

[1] *Department of Mathematics, Heriot-Watt University,*
Edinburgh EH14 4AS, UK (D.B.Duncan@hw.ac.uk)
[2] *Bindura University College, Post Bag 1020,*
Bindura, Zimbabwe

(Received 21 August 2000)

Abstract We show that in the classical (fixed-monomer-concentration) Becker–Döring equations truncated at finite cluster size, the slow evolution (metastability) of solutions can be explained in terms of the eigensystem of this linear ordinary differential equation (ODE) system. In particular, for a common choice of coagulation–fragmentation rate constants there is an extremely small non-zero eigenvalue which is isolated from the rest of the spectrum. We give estimates and bounds on the size of this eigenvalue, the gap between it and the second smallest, and the size of the largest eigenvalue. The bounds on the smallest eigenvalue are very sharp when the system size and/or monomer concentration are large enough.

Keywords: coagulation–fragmentation; metastability; eigenvalue estimates;
asymptotic behaviour of solution

AMS 2000 *Mathematics subject classification:* Primary 34A30; 15A18; 65F15

1. Introduction

The process of coagulation and fragmentation of clusters of particles is important in physics, astronomy, polymer physics, atmospheric physics and colloid chemistry. In 1935, Becker and Döring [**2**] introduced a model of the dynamics of cluster formation in a system composed of identical particles, where clusters can only gain or lose single particles called monomers and are uniformly distributed in space. Their model has fixed monomer concentration and is an infinite system of ordinary differential equations (ODEs) for the concentrations of clusters of different sizes. It was originally formulated as a model of condensation.

One of the main subjects of interest in studies of the Becker–Döring model and its variants is the slow evolution (metastability) of solutions. The metastable time-scale determines how long processes like condensation and polymerization reactions take. Penrose [**8**] summarizes the history of, and problems encountered in, work on Becker–Döring metastability and goes on to give rigorous estimates for the time-scale of metastable solutions. Refinements of that work can be found in [**6**, **7**] and § 2 below. A detailed numeri-

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

cal analysis is given in [**3**], and various reduced models which reproduce the metastable behaviour are derived and tested in [**5**].

In this paper we consider the truncated version of classical (fixed-monomer-concentration) Becker–Döring equations given by

$$\left.\begin{aligned} c_1 &= z, \\ \dot{c}_r &= J_{r-1} - J_r, \quad r = 2, \dots, N-1, \\ \dot{c}_N &= J_{N-1}, \end{aligned}\right\} \tag{1.1}$$

where $c_r(t)$ denotes the number of r-particle clusters per unit volume at time t, initial data $c_r(0) \geqslant 0$, the monomer concentration $z > 0$ is a constant independent of t, and the flux

$$J_r = a_r c_1 c_r - b_{r+1} c_{r+1}, \quad r = 1, \dots, N-1, \tag{1.2}$$

is the net rate at which r-clusters are converted to $(r+1)$-clusters. Parameters $a_r > 0$ and $b_{r+1} > 0$, for $r = 1, \dots, N-1$, are, respectively, the coagulation and fragmentation rate constants and $b_1 = 0$ since monomers cannot fragment. A typical choice of parameters is

$$a_r \equiv 1, \qquad b_{r+1} = \exp(r^{2/3} - (r-1)^{2/3}), \tag{1.3}$$

for $r \geqslant 1$.

The problem defined by (1.1) is a linear system of ODEs and can be written in the vector form

$$\dot{\boldsymbol{c}} = M\boldsymbol{c} + (a_1 z^2, 0, \dots, 0)^{\mathrm{T}} \tag{1.4}$$

where the coefficient matrix M is the $N-1 \times N-1$ tridiagonal

$$M = \begin{pmatrix} -a_2 z - b_2 & b_3 & & & & \\ \ddots & \ddots & \ddots & & & \\ & \ddots & \ddots & \ddots & & \\ & z a_{r-1} & -a_r z - b_r & b_{r+1} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots \\ & & & & a_{N-1} z & -b_N \end{pmatrix}, \tag{1.5}$$

and for convenience $\boldsymbol{c} = (c_2, c_3, \dots, c_N)^{\mathrm{T}}$. There is a unique equilibrium solution given, component-wise, by

$$\tilde{c}_r = Q_r z^r, \quad r = 1, \dots, N, \tag{1.6}$$

where

$$Q_1 = 1 \quad \text{and} \quad Q_r = \prod_{k=2}^{r} \frac{a_{k-1}}{b_k}, \quad r = 2, \dots, N.$$

With rate constants (1.3),

$$Q_r = \exp(-(r-1)^{2/3}).$$

The formula (1.6) is found by noting that $J_r \equiv 0$ for equilibrium in (1.1) and using the definition (1.2) for J_r.

In 1979, Penrose and Leibowitz [**9**] proposed an alternative Becker–Döring model with fixed system density instead of fixed monomer concentration, giving rise to a nonlinear system of ODEs. Since then there has been a great deal of work on the metastable behaviour (slow evolution) of solutions of the original and Penrose–Leibowitz versions (see, for example, [**1**, **6**–**9**] and [**3**] for detailed numerical analysis). In fact there are more than two variants of the basic equations. These studies have considered either fixed monomer concentration or fixed system density (Cases A and B of [**8**]) and either infinite or finite system size. The truncation to finite system size also takes different forms. One way is to set all concentrations $c_r(t) \equiv 0$, for $r > N$, giving

$$\dot{c}_N = J_{N-1} - z a_N c_N \tag{1.7}$$

(see, for example, [**7**, **8**]), and another is to set all fluxes $J_r(t) \equiv 0$, for $r \geqslant N$, giving the form used in (1.1) (see, for example, [**1**, **3**, **6**]).

The equilibrium solution (1.6) is common to all of these versions of the Becker–Döring equations except (1.7) and plays a key role in determining if metastability is present or not. With rate constants (1.3), metastable behaviour is possible in the finite-dimensional cases when $\tilde{c}_r$ has a minimum turning point at $r = r^* \in (1, N)$. Roughly speaking, the problem represents a very slow phase transition from clusters smaller than r^* to those larger than it. This can only happen for $z > 1$, since $\tilde{c}_r$ is monotonic decreasing for all r if and only if $z \leqslant 1$. Results for the linear (fixed-monomer-concentration) infinite-dimensional case are the same, while those for the nonlinear infinite-dimensional case are similar, but more subtle. The references above give a full discussion of this. The main point for the problem here is that metastability is only expected when $z > 1$.

In this paper we show that solutions of the linear Becker–Döring Equation (1.1) can evolve extremely slowly and give estimates for the time-scales involved. In § 2 we construct upper and lower bounds on a class of solutions as they evolve in time, and use these to obtain an estimate for the speed of approach to equilibrium of these solutions. In § 3 we use analysis borrowed from numerical linear algebra to give some very accurate estimates of the size of the important eigenvalues of the coefficient matrix (1.5) and hence obtain estimates of the time-scales in the solution of (1.1). We finish with some illustrative examples in § 4 and conclusions in § 5.

2. Upper and lower bounds on the solution

Here we show that a class of solutions of (1.1) is trapped between upper and lower bounds, and use this information to estimate how fast this class of solutions approaches equilibrium. This follows work in [**6**, **8**] for other versions of the Becker–Döring equations. Many of the results in this section rely on the non-negativity property of the Becker–Döring equations (1.1): solutions with initial data $c_r(0) \geqslant 0$, for $r = 1, \dots, N$, satisfy $c_r(t) \geqslant 0$ for all $t \geqslant 0$. This is physically reasonable since the c_r are concentrations. Recall that ODEs of the form $\dot{\boldsymbol{c}} = M\boldsymbol{c} + \boldsymbol{b}$ have this property when all the elements of $\boldsymbol{b}$ and all the off-diagonal elements of M are non-negative.

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

We start by showing that a wide class of solutions of (1.1) approach the equilibrium solution from below.

Lemma 2.1. *If $c^a(t)$ and $c^b(t)$ are solutions of (1.1) with initial data $c^a_r(0) \leqslant c^b_r(0)$, for $r = 1, \dots, N$, then $c^a_r(t) \leqslant c^b_r(t)$ for all $t \geqslant 0$. In particular, if solution $c(t)$ has initial data $c_r(0) \leqslant \tilde{c}_r$ (the equilibrium solution (1.6)), then $c_r(t) \leqslant \tilde{c}_r$ for all $t \geqslant 0$.*

Proof. Since $c^a(t)$ and $c^b(t)$ are solutions of (1.1), $u_r(t) = c^b_r(t) - c^a_r(t)$ satisfies $\dot{\boldsymbol{u}} = M\boldsymbol{u}$, where matrix M is defined by (1.5) and $\boldsymbol{u} = (u_2, \dots, u_N)^{\mathrm{T}}$. This ODE system has the non-negativity property, and so the result follows immediately. The second part follows by setting $c^a(t) = c(t)$ and $c^b(t) = \tilde{c}$. □

The next results are for solutions with initial data of the form $c_r(0) = M_r(z, y)$, for $z, y > 0$, where the function

$$M_r(z, y) \stackrel{\text{def}}{=} Q_r z^r \left(1 - y \sum_{k=1}^{r-1} A_k(z)\right) \tag{2.1}$$

with

$$A_k(z) \stackrel{\text{def}}{=} \frac{1}{a_k Q_k z^{k+1}}.$$

A key observation is that M_r satisfies the identity

$$z a_r M_r(z, y) - b_{r+1} M_{r+1}(z, y) = y$$

for each $r = 1, \dots, N-1$. The function M_r plays an important part in the analysis of linear and nonlinear Becker–Döring equations in [**6**–**8**] and it is used again in the next section.

Lemma 2.2. *Let $c^m(t)$ denote the solution of (1.1) determined by the initial conditions $c^m_r(0) = M_r(z, \mathcal{J}(0))$ with $0 \leqslant \mathcal{J}(0) \leqslant 1/\sum_{k=1}^{N-1} A_k(z)$. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t} c^m_r(t) \geqslant 0,$$

$$0 \leqslant c^m_r(0) \leqslant c^m_r(t) \leqslant M_r(z, \mathcal{J}(t)) \leqslant \tilde{c}_r$$

and

$$\mathcal{J}(0)\mathrm{e}^{-b_N t} \leqslant J_r(t), \qquad \dot{c}^m_N(t) \leqslant \mathcal{J}(0)$$

for all $t \geqslant 0$ and all relevant cluster sizes r, where

$$\mathcal{J}(t) \stackrel{\text{def}}{=} \frac{\tilde{c}_N - c^m_N(t)}{\tilde{c}_N \sum_{k=1}^{N-1} A_k(z)}.$$

Proof. Differentiating (1.4) gives $\ddot{\boldsymbol{c}}^m = M\dot{\boldsymbol{c}}^m$, an ODE for $\dot{\boldsymbol{c}}^m = (\dot{c}^m_2, \dots, \dot{c}^m_N)^{\mathrm{T}}$ which has the non-negativity property. It has non-negative initial data $\dot{c}^m_N(0) = \mathcal{J}(0) \geqslant 0$, $\dot{c}^m_r(0) = 0$, for $r = 2, \dots, N-1$, and so the result $\dot{c}^m_r \geqslant 0$ follows and immediately gives $c^m_r(t) \geqslant c^m_r(0)$.

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

By inspection of the definition (2.1), the upper bound on $\mathcal{J}(0)$ ensures that $0 \leqslant c_r^m(0)$. It is also clear that $M_r(z, \mathcal{J}(t)) \leqslant \tilde{c}_r$ if and only if $\mathcal{J}(t) \geqslant 0$ if and only if $c_N^m(t) \leqslant Q_N z^N$, while Lemma 2.1 guarantees that the last inequality is satisfied.

To show that $v_r(t) \overset{\text{def}}{=} M_r(z, \mathcal{J}(t)) - c_r^m(t) \geqslant 0$, substitute into (1.1) to get

$$\dot{v}_r = za_{r-1}v_{r-1} - (b_r + za_r)v_r + b_{r+1}v_{r+1} + \left(\frac{\tilde{c}_r \displaystyle\sum_{k=1}^{r-1} A_k(z)}{\tilde{c}_N \displaystyle\sum_{k=1}^{N-1} A_k(z)} \right) \frac{\mathrm{d}}{\mathrm{d}t} c_N^m, \quad r = 2, \dots, N-1,$$

with $v_1 = 0$ and $v_N = 0$. The solution $v(t)$ is non-negative since the ODEs have the non-negativity property (the inhomogeneous terms, which come from the time derivative of $M_r(z, \mathcal{J}(t))$, are non-negative) and the initial data are non-negative.

To show that $J_r(t) \leqslant \mathcal{J}(0)$, first note that $c(t)$ is a solution of (1.1) if and only if $J(t)$ is a solution of

$$\left.\begin{aligned} \dot{J}_1 &= -b_2 J_1 + b_2 J_2, \\ \dot{J}_r &= za_r J_{r-1} - (za_r + b_{r+1}) J_r + b_{r+1} J_{r+1}, \quad 2 \leqslant r \leqslant N-2, \\ \dot{J}_{N-1} &= za_{N-1} J_{N-2} - (za_{N-1} + b_N) J_{N-1}. \end{aligned}\right\} \tag{2.2}$$

Next define $K_r(t) = \mathcal{J}(0) - J_r(t)$, substitute into (2.2) and show that if $K_r(0) \geqslant 0$ then $K_r(t) \geqslant 0$ for all $t \geqslant 0$ using the same argument as for the previous part.

To show that $J_r(t) \geqslant \mathcal{J}(0)\mathrm{e}^{-b_N t}$, define $L_r(t) = J_r(t) - \mathcal{J}(0)\mathrm{e}^{-b_N t}$, substitute into (2.2) and follow the same argument as for the previous part.

The results for $\dot{c}_N^m$ follow directly from those for J_{N-1}. □

The previous results lead directly to the following estimate of the time taken to approach the equilibrium solution. The distance from equilibrium of a solution $c(t)$ of (1.1) is measured by

$$d_{\mathrm{eq}}(t) = \sum_{r=1}^{N} r|\tilde{c}_r - c_r(t)|$$

using a weighted 1-norm which measures in terms of number of particles per unit volume. This norm is common in studies of Becker–Döring equations (see, for example, [**1**]).

Theorem 2.3. *If initial conditions for (1.1) are chosen so that $c_r(0) \leqslant M_r(z, \mathcal{J}(0))$ (see Lemma 2.2 for definitions) and $J_r(0) \equiv za_r c_r(0) - b_{r+1} c_{r+1}(0) \geqslant 0$, then the distance $d_{\mathrm{eq}}(t)$ of the solution $c(t)$ from equilibrium satisfies*

$$1 - \frac{t}{Q_N z^N \sum_{k=1}^{N-1} A_k(z)} \leqslant \frac{d_{\mathrm{eq}}(t)}{d_{\mathrm{eq}}(0)} \leqslant 1 \quad \textit{and} \quad \frac{\mathrm{d}}{\mathrm{d}t} d_{\mathrm{eq}} \leqslant 0.$$

Thus $t = Q_N z^N \sum_{k=1}^{N-1} A_k(z)$ is a lower bound on the time-scale of the approach to equilibrium for a range of initial data.

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

Proof. Lemmas 2.1 and 2.2 together give

$$c_r(t) \leqslant c_r^m(t) \leqslant M_r(z, \mathcal{J}(t)) \leqslant \tilde{c}_r \tag{2.3}$$

for all $t \geqslant 0$, using the functions defined in Lemma 2.2. Hence

$$d_{\text{eq}}(t) = \sum_{r=2}^{N} r|\tilde{c}_r - c_r(t)| = \sum_{r=2}^{N} r(\tilde{c}_r - c_r(t)).$$

Differentiating with respect to t and using Equation (1.1) gives

$$\frac{\mathrm{d}}{\mathrm{d}t} d_{\text{eq}} = -\sum_{r=2}^{N} r\dot{c}_r(t) = -2J_1 - \sum_{r=2}^{N-1} J_r \leqslant 0,$$

since the $J_r \geqslant 0$. They satisfy the ODEs (2.2) in Lemma 2.2 which have the non-negativity property and non-negative initial data.

The result $d_{\text{eq}}(t)/d_{\text{eq}}(0) \leqslant 1$ follows directly from $\dot{d}_{\text{eq}} \leqslant 0$, while result (2.3) is used for the inequality in

$$\begin{aligned} d_{\text{eq}}(t) &= \sum_{r=2}^{N} r(Q_r z^r - c_r(t)) \\ &\geqslant \sum_{r=2}^{N} r(Q_r z^r - M_r(z, \mathcal{J}(t))) \\ &= \mathcal{J}(t)\biggl(\sum_{r=2}^{N} rQ_r z^r \sum_{k=1}^{r-1} A_k(z)\biggr) \\ &= \mathcal{J}(t) d_{\text{eq}}(0)/\mathcal{J}(0). \end{aligned}$$

The final stage is to use $\dot{c}_N^m \leqslant \mathcal{J}(0)$ from Lemma 2.2 to get $c_N^m(t) - c_N^m(0) \leqslant \mathcal{J}(0)t$, so that

$$\begin{aligned} \frac{\mathcal{J}(t)}{\mathcal{J}(0)} &= \frac{Q_N z^N - c_N^m(t)}{Q_N z^N - c_N^m(0)} = 1 - \frac{c_N^m(t) - c_N^m(0)}{Q_N z^N - c_N^m(0)} \\ &\geqslant 1 - \frac{\mathcal{J}(0)t}{\mathcal{J}(0) Q_N z^N \sum_{k=1}^{N-1} A_k(z)} \end{aligned}$$

and the result is proved.

The set of initial data allowed by the theorem is not empty, since, for example, $c_r(0) = \mu M_r(z, \mathcal{J}(0))$, for $r = 2, \dots, N$, satisfies its conditions for all $\mu \in [0, 1]$. □

Remark 2.4. The theorem covers the special case of pure monomer initial data: $c_r = 0$, for $r = 2, \dots, N$, and $c_1 = z$. Results for this case are given in § 4.

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

Remark 2.5. With parameters (1.3), the time-scale of the approach to equilibrium given in the previous theorem is

$$Q_N z^N \sum_{k=1}^{N-1} A_k(z) \geqslant Q_N z^N A_1(z) = \exp((N-2)\log z - (N-1)^{2/3}),$$

which is exponentially large in system size N for monomer concentrations $z > 1$.

In the next section we take a different approach to estimating the time-scales of the problem, and connect our main result from Theorem 2.3 above with the results derived below for the eigenvalues of the coefficient matrix (1.5) (see Remark 3.1). Results from both sections are examined and plotted for a range of values of system size N and monomer concentration z in § 4.

3. Eigensystem of the Becker–Döring equation

The main aim of this section is to explain the behaviour of solutions of the linear ODE system (1.1) in terms of the eigensystem of its coefficient matrix M given above in (1.5).

One of the important relationships used below is the similarity transform

$$M = DSD^{-1}, \tag{3.1}$$

where D is the real, diagonal matrix $D = \operatorname{diag}(\sqrt{\tilde{c}_2}, \dots, \sqrt{\tilde{c}_N})$ and S is the real, symmetric, tridiagonal matrix

$$S = \begin{pmatrix} -a_2 z - b_2 & \sqrt{z a_2 b_3} & & & \\ \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & \sqrt{z a_{r-1} b_r} & -a_r z - b_r & \sqrt{z a_r b_{r+1}} & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots \\ & & & \sqrt{z a_{N-1} b_N} & -b_N \end{pmatrix}. \tag{3.2}$$

The elements of D are the square roots of the equilibrium solution defined in (1.6). Kreer [**7**] also uses similarity transforms to obtain information about another version of the Becker–Döring equations.

The link between M and the symmetric matrix S allows us to extract a great deal of information about its eigensystem, and hence to explain how solutions of the ODE (1.4) behave. In particular, we show that the eigenvalues λ_j of M are real, distinct and satisfy

$$\lambda_{N-1} < \cdots < \lambda_1 < 0.$$

Hence M has a full basis of eigenvectors $\{\boldsymbol{q}_j : j = 1, \dots, N-1\}$ associated with the λ_j. The solution of the ODE system (1.4) is then

$$\boldsymbol{c}(t) = \tilde{\boldsymbol{c}} + \sum_{j=1}^{N-1} \alpha_j \mathrm{e}^{\lambda_j t} \boldsymbol{q}_j,$$

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

where $\tilde{\boldsymbol{c}}$ is the equilibrium solution defined element-wise in (1.6), and the α_j are determined by the initial condition $\tilde{\boldsymbol{c}} + \sum_{j=1}^{N-1} \alpha_j \boldsymbol{q}_j = \boldsymbol{c}(0)$. Of course, $\boldsymbol{c}(t) \to \tilde{\boldsymbol{c}}$ as $t \to \infty$ and the time-scale of the approach to equilibrium is $1/|\lambda_1|$.

Remark 3.1. The time-scale bound of Theorem 2.3 from the previous section can be regarded as giving the estimate

$$\lambda_1 \approx -\biggl(Q_N z^N \sum_{k=1}^{N-1} A_k(z)\biggr)^{-1}, \tag{3.3}$$

which can be very small (see Remark 2.5).

In this section we obtain more detailed estimates of eigenvalues λ_1, λ_2 and λ_{N-1} and we show that for the common choice of parameters (1.3) λ_1 can be very small relative to the rest of the eigenvalues, including λ_2. For large time t,

$$\boldsymbol{c}(t) \approx \tilde{\boldsymbol{c}} + \alpha_1 \mathrm{e}^{\lambda_1 t} \boldsymbol{q}_1,$$

and the solution will appear not to change for t between $O(1/|\lambda_2|)$ and $O(1/|\lambda_1|)$, when the slow decay towards the equilibrium solution shows up. The solution could appear to be in equilibrium when it is far from it—the characteristic problem of metastability.

To obtain estimates of the sizes of the eigenvalues we use the Gerschgorin Circle Theorem and results involving the Rayleigh quotient given in the next two lemmas.

Lemma 3.2 (Gerschgorin). *Let $A = (a_{ij})$ be an arbitrary complex $d \times d$ matrix. Then each of the eigenvalues of A lies in the union of the discs*

$$C_i = \biggl\{\zeta \in \mathbb{C} : |\zeta - a_{ii}| \leqslant \sum_{j \neq i} |a_{ij}|\biggr\}.$$

Proof. See, for example, [**10**, Chapter 2.13]. □

Lemma 3.3 (Rayleigh quotient). *The Rayleigh quotient $\rho(S, \boldsymbol{x})$ of the real, symmetric matrix S with any real vector $\boldsymbol{x} \neq \boldsymbol{0}$ is defined by $\rho(S, \boldsymbol{x}) \equiv (\boldsymbol{x}^{\mathrm{T}} S \boldsymbol{x})/(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x})$. It satisfies*

$$\min_j \lambda_j \leqslant \rho(S, \boldsymbol{x}) \leqslant \max_j \lambda_j,$$

where λ_j are the eigenvalues of S. Furthermore, if λ_i is the closest eigenvalue to ρ, then

$$|\lambda_i - \rho(S, \boldsymbol{x})| \leqslant \frac{\|\boldsymbol{r}\|_2^2}{\mathrm{gap}'},$$

where the residual $\boldsymbol{r} \equiv (S\boldsymbol{x} - \rho \boldsymbol{x})/\|\boldsymbol{x}\|_2$ and $\mathrm{gap}' \equiv \min_{j \neq i} |\lambda_j - \rho(S, \boldsymbol{x})|$.

Proof. See, for example, [**4**, Chapter 5.2 and Theorem 5.5]. □

The Gerschgorin Lemma and the similarity transform (3.1) linking M to the symmetric matrix S allow us to prove the next theorem about the eigenvalues of M.

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

Theorem 3.4. *The matrices M and S share the same eigenvalues, which are real, distinct and strictly negative. Hence, after ordering, the eigenvalues satisfy*

$$\lambda_{N-1} < \cdots < \lambda_2 < \lambda_1 < 0,$$

and we note in particular that M is not singular.

Proof. M and S are related by the similarity transformation (3.1) and hence have the same eigenvalues [**10**, Chapter 1.5]. S is real and symmetric and so the eigenvalues are real. The sub- and super-diagonal entries of the tridiagonal matrix S are non-zero (that is $\sqrt{za_{r-1}b_r} \neq 0$) and hence the eigenvalues are distinct [**10**, Chapter 5.37].

Using the Gerschgorin Lemma 3.2 applied to M^{T} (which has the same eigenvalues as M), the maximum value a real eigenvalue can have is at the right edge of the union of the Gerschgorin disks. Hence

$$\lambda_1 \leqslant \max_{r=3,\dots,N-1}(-a_2z - b_2 + |a_2z|, -a_rz - b_r + |a_rz| + |b_r|, -b_N + |b_N|) = 0,$$

since a_r, b_r, z are positive.

To show that $\lambda_1 \neq 0$, we note that if M is singular, then there exists $\boldsymbol{c} \neq \boldsymbol{0}$ such that $M\boldsymbol{c} = \boldsymbol{0}$. This system of equations can be written as

$$-b_2c_2 - J_2 = 0, \quad J_2 - J_3 = 0, \quad \dots, \quad J_{N-2} - J_{N-1} = 0, \quad J_{N-1} = 0,$$

where $J_r = za_rc_r - b_{r+1}c_{r+1}$. The solution is $c_2 = J_2 = \cdots = J_{N-1} = 0$, which can only hold when $\boldsymbol{c} = \boldsymbol{0}$. Hence M is not singular. □

Now we use the results above to derive bounds on the eigenvalues λ_{N-1} and λ_2.

Theorem 3.5. *The largest eigenvalue (in magnitude) λ_{N-1} of M and S is bounded below by*

$$\min_{r=3,\dots,N-1}(-2a_2z - b_2, -2a_rz - 2b_r, -2b_N) \leqslant \lambda_{N-1}, \tag{3.4}$$

and above by

$$\lambda_{N-1} \leqslant -\tfrac{1}{2}(a_2z + a_3z + b_2 + b_3 + 2\sqrt{za_2b_3}).$$

Proof. The first result is obtained by application of the Gerschgorin Lemma 3.2 to M^{T} and the second uses the Rayleigh quotient Lemma 3.3 with $\boldsymbol{x} = (1, -1, 0, \dots, 0)^{\mathrm{T}}$. □

Remark 3.6. If a_r, b_r are given by (1.3), then Theorem 3.5 implies that

$$-2z - \mathrm{e} \leqslant \lambda_{N-1} \leqslant -z - 1.35\sqrt{z} - 2.26,$$

so that $\lambda_{N-1} = O(1)$ if $z = O(1)$.

Theorem 3.7. *The second smallest eigenvalue*

$$\lambda_2 \leqslant \lambda_2^{(+)} = \min\{\lambda_2^{(+)}(S), \lambda_2^{(+)}(M)\}, \tag{3.5}$$

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

where

$$\lambda_2^{(+)}(S) = \max_{r=3,\dots,N-2}\left\{\begin{matrix} -a_2z - b_2 + \sqrt{za_2b_3}, \\ -a_rz - b_r + \sqrt{za_{r-1}b_r} + \sqrt{za_rb_{r+1}}, \\ -a_{N-1}z - b_{N-1} + \sqrt{za_{N-2}b_{N-1}} \end{matrix}\right\}$$

and

$$\lambda_2^{(+)}(M) = \max_{r=3,\dots,N-2}\left\{\begin{matrix} -a_2z - b_2 + b_3, \\ -a_rz - b_r + a_{r-1}z + b_{r+1}, \\ -a_{N-1}z - b_{N-1} + a_{N-2}z. \end{matrix}\right\}$$

Proof. The result $\lambda_2 \leqslant \lambda_2^{(+)}(S)$ comes from applying the Gerschgorin Lemma 3.2 to the leading principal submatrix S' of degree one less than S to get an upper bound on λ'_1 (the maximum eigenvalue of S'). All the eigenvalues of S' are real because it is symmetric, and hence all the λ'_j are less than or equal to the rightmost point of the Gerschgorin disks. Furthermore, because S is real and symmetric, the eigenvalues of S' are interleaved with those of S (see, for example, [**10**, Chapter 2.47]) giving

$$\lambda_{N-1} \leqslant \lambda'_{N-2} \leqslant \lambda_{N-2} \leqslant \lambda'_{N-3} \leqslant \cdots \leqslant \lambda'_2 \leqslant \lambda_2 \leqslant \lambda'_1 \leqslant \lambda_1. \tag{3.6}$$

The result that we need is $\lambda_2 \leqslant \lambda'_1$.

The result $\lambda_2 \leqslant \lambda_2^{(+)}(M)$ follows in the same way because $M' = D'S'D'^{(-1)}$, where the prime denotes the leading principal submatrix. □

Remark 3.8. With parameter choice (1.3), Theorem 3.7 gives

$$\lambda_2 \leqslant \lambda_2^{(+)}(M) = b_{N-1} - b_{N-2} < -\tfrac{2}{9}N^{-4/3},$$

which is algebraically rather than exponentially small as $N \to \infty$.

Finally, in this section we use the results above to derive bounds on the maximum eigenvalue λ_1 of M and S. We already know from Theorem 3.4 that $\lambda_1 < 0$, and we are now going to establish more definite bounds on its size. We find one lower bound and three upper bounds which apply in different parameter regions. Two of the upper bounds on λ_1 come by application of the Gerschgorin Lemma 3.2 to M and S, respectively. The lower bound and the third upper bound are found using the Rayleigh Quotient Lemma 3.3 applied with a very good guess at the eigenvector associated with λ_1.

Our guess at the eigenvector associated with eigenvalue λ_1 comes from results for the nonlinear Becker–Döring equation in [**6**]. They indicate that the solution eventually behaves like

$$c_r(t) \approx \tilde{c}_r\left(1 - \mathcal{J}(t)\sum_{k=1}^{r-1}\frac{1}{a_kQ_kz^{k+1}}\right) = M_r(z,\mathcal{J}(t)), \quad r = 2,\dots,N,$$

where $\mathcal{J}(t) \to 0$ as $t \to \infty$ and M_r was defined in (2.1) in the previous section. We use the analogy with the large time behaviour

$$\boldsymbol{c}(t) \approx \tilde{\boldsymbol{c}} + \alpha_1\boldsymbol{q}_1\mathrm{e}^{\lambda_1 t}$$

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

of our linear problem (1.4) to guess that the eigenvector $\boldsymbol{p}_1 = D^{-1}\boldsymbol{q}_1$ of S is approximated by $\hat{\boldsymbol{p}}$, where

$$\hat{p}_r = \sqrt{\tilde{c}_r}\left(\sum_{k=1}^{r-1} \frac{1}{a_k Q_k z^{k+1}}\right) = \sqrt{\tilde{c}_r}\left(\sum_{k=1}^{r-1} \frac{1}{z a_k \tilde{c}_k}\right), \tag{3.7}$$

for $r = 2, \dots, N$. Perhaps surprisingly,

$$S\hat{\boldsymbol{p}} = (0, \dots, 0, -1/\sqrt{\tilde{c}_N})^{\mathrm{T}}, \tag{3.8}$$

making the algebraic manipulations required below relatively simple.

Theorem 3.9. *The smallest eigenvalue* λ_1 *of* M *and* S *is bounded below by*

$$-\frac{a_1 z^2}{\tilde{c}_N} < \frac{-\hat{p}_N}{\|\hat{\boldsymbol{p}}\|_2^2 \sqrt{\tilde{c}_N}} = \rho(S, \hat{\boldsymbol{p}}) \leqslant \lambda_1, \tag{3.9}$$

where $\hat{\boldsymbol{p}}$ *is given in (3.7). It is bounded above by* $\lambda_1 < 0$,

$$\lambda_1 \leqslant \max_{r=3,\dots,N-1} \left\{ \begin{array}{c} -a_2 z - b_2 + b_3, \\ -a_r z - b_r + a_{r-1} z + b_{r+1}, \\ a_{N-1} z - b_N \end{array} \right\} \tag{3.10}$$

and

$$\lambda_1 \leqslant \max_{r=3,\dots,N-1} \left\{ \begin{array}{c} -a_2 z - b_2 + \sqrt{z a_2 b_3}, \\ -a_r z - b_r + \sqrt{z a_{r-1} b_r} + \sqrt{z a_r b_{r+1}}, \\ \sqrt{z a_{N-1} b_N} - b_N. \end{array} \right\} \tag{3.11}$$

Furthermore, when $\lambda_2^{(+)} < 2\rho(S, \hat{\boldsymbol{p}})$,

$$\lambda_1 \leqslant \rho(S, \hat{\boldsymbol{p}}) + \frac{(\|\hat{\boldsymbol{p}}\|_2^2 - \hat{p}_N^2)}{\tilde{c}_N \|\hat{\boldsymbol{p}}\|_2^4 (\rho(S, \hat{\boldsymbol{p}}) - \lambda_2^{(+)})}, \tag{3.12}$$

where $\lambda_2^{(+)}$ *is defined in Theorem 3.7.*

Proof. The sharper lower bound in (3.9) comes from the first part of Lemma 3.3 and the observation (3.8), while the weaker bound follows from

$$\frac{\hat{p}_N}{\|\hat{\boldsymbol{p}}\|_2^2 \sqrt{\tilde{c}_N}} < \frac{1}{\hat{p}_N \sqrt{\tilde{c}_N}} = \frac{z}{\tilde{c}_N \sum_{r=1}^{N-1} 1/(a_r \tilde{c}_r)} < \frac{a_1 z^2}{\tilde{c}_N}.$$

The upper bound $\lambda_1 < 0$ comes from Theorem 3.4, while (3.10) and (3.11) are obtained by application of the Gerschgorin Lemma 3.2 to M and S, respectively. The final upper bound (3.12) comes from the second part of Lemma 3.3 after some algebra and the observation that if $\lambda_2^{(+)}(z) < 2\rho(S, \hat{\boldsymbol{p}})$, then λ_1 is the closest eigenvalue to ρ and

$$\mathrm{gap}' \equiv \min_{j \neq 1} |\lambda_j - \rho(S, \hat{\boldsymbol{p}})| = \rho(S, \hat{\boldsymbol{p}}) - \lambda_2 \geqslant \rho(S, \hat{\boldsymbol{p}}) - \lambda_2^{(+)}(z) > -\rho(S, \hat{\boldsymbol{p}}) > 0.$$

□

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

Remark 3.10. With parameters (1.3) and monomer concentration $z > 1$, inequality (3.9) of Theorem 3.9 gives

$$-\exp((N-1)^{2/3} - (N-2)\log z) = \frac{-z^2}{\tilde{c}_N} < \lambda_1,$$

so that λ_1 is exponentially small as $N \to \infty$. When combined with Theorem 3.7, we see that

$$\left|\frac{\lambda_1}{\lambda_2}\right| \leqslant \tfrac{9}{2} N^{4/3} \exp((N-1)^{2/3} - (N-2)\log z),$$

which is also exponentially small as $N \to \infty$. However, when $z = 1$ the inequality (3.10) gives

$$\lambda_1 < b_N - b_{N-1} < -\tfrac{2}{9} N^{-4/3},$$

which is not exponentially small. This ties in with the observations in § 1 that metastability is only expected when $z > 1$.

The results of this and the previous section are linked through the observations in Remarks 2.5 and 3.10. The same exponentially small function appears in each.

4. Illustrative examples

In this section we illustrate how the eigenvalues λ_1, λ_2 and λ_{N-1} influence solutions of (1.1), then examine their behaviour and that of the estimates derived in the previous two sections over a range of values of z and N.

Figure 1 shows the result of solving (1.1) with the rate constants (1.3) and pure monomer initial data

$$c_1 = z > 0, \qquad c_r(0) = 0, \quad r = 2, \dots, N.$$

The plot is of the relative 'distance' from equilibrium $\tilde{c}_r$ of the concentration of r-particle clusters c_r,

$$|c_r(t) - \tilde{c}_r| / |c_r(0) - \tilde{c}_r|,$$

against time t. The larger clusters reach a very slowly varying metastable state between $t = |\lambda_2|^{-1}$ and $|\lambda_1|^{-1}$, and finally decay away to their true equilibrium as t increases. The smaller clusters get close to equilibrium much more quickly than the larger ones and the behaviour of the smaller clusters appears to be simpler, changing with time-scales between $t = |\lambda_{N-1}|^{-1}$ and $|\lambda_2|^{-1}$. However, on closer examination the concentrations of smaller clusters also settle into plateaux between $t = |\lambda_2|^{-1}$ and $|\lambda_1|^{-1}$, but their heights are too small to distinguish in Figure 1. If N and/or z are increased, then the ratio λ_2/λ_1 will increase dramatically (see Remark 3.10) and this plateau period will be much longer.

To get some idea of the time-scales possible, the eigenvalue estimates from the previous sections are plotted in Figures 2 and 3 and compared with the results of a standard numerical eigenvalue routine (bisection based on Sturm sequences [**4**, Chapter 5.3.4]). We note that using any numerical method there are great difficulties in resolving extremely

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

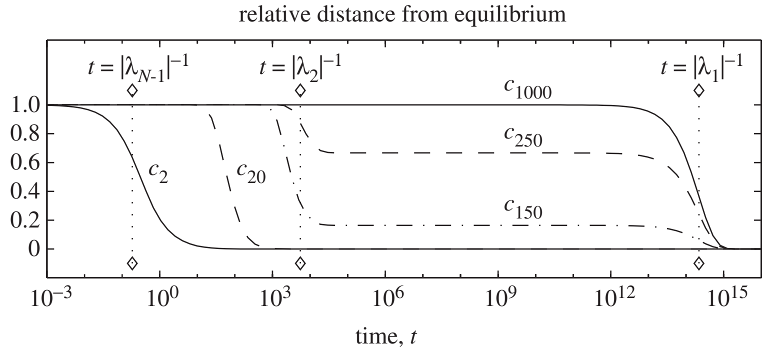

Figure 1. Solution of the Becker–Döring system with $N = 1000$ and $z = 1.12$. The relative distance from equilibrium is defined by $|c_r(t) - \tilde{c}_r|/|c_r(0) - \tilde{c}_r|$. The vertical lines with diamonds mark the time-scales of the slowest and fastest eigencomponents. The estimate from § 2 predicts a time-scale of 9.4×10^{12} in the approach to equilibrium.

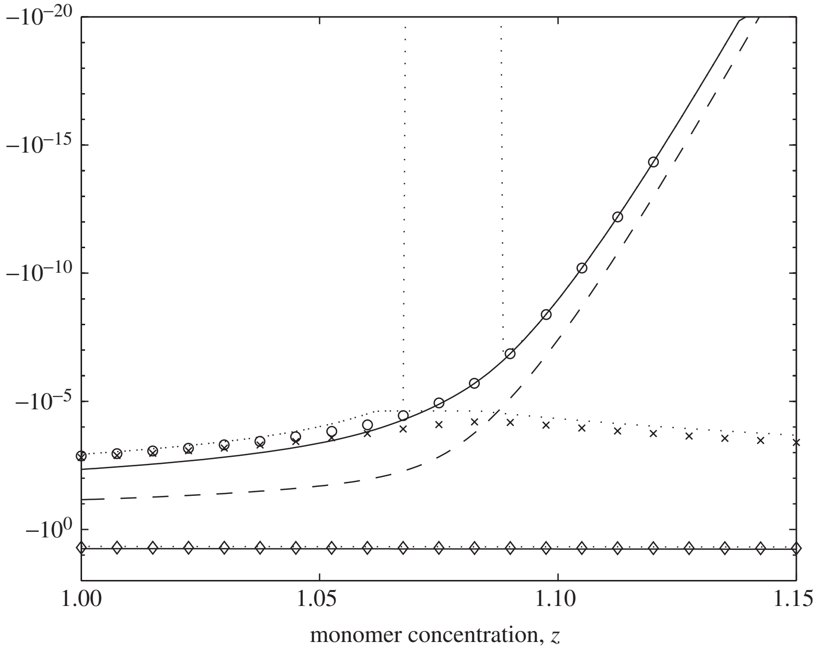

Figure 2. Eigenvalues and bounds on the eigenvalues versus monomer concentration z (system size $N = 1000$). The bounds derived in § 3 and the estimate from Theorem 2.3 (via (3.3)) are shown. The λ_1 and λ_2 upper bounds coincide at the left of the plot, and the upper and lower bounds on λ_1 coincide at the right of the plot. The 'exact' eigenvalues marked with symbols are determined numerically (○, λ_1; ×, λ_2; ◇, λ_{N-1}; ······, upper bounds; ——, lower bounds; – – –, Theorem 2.3 estimate).

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

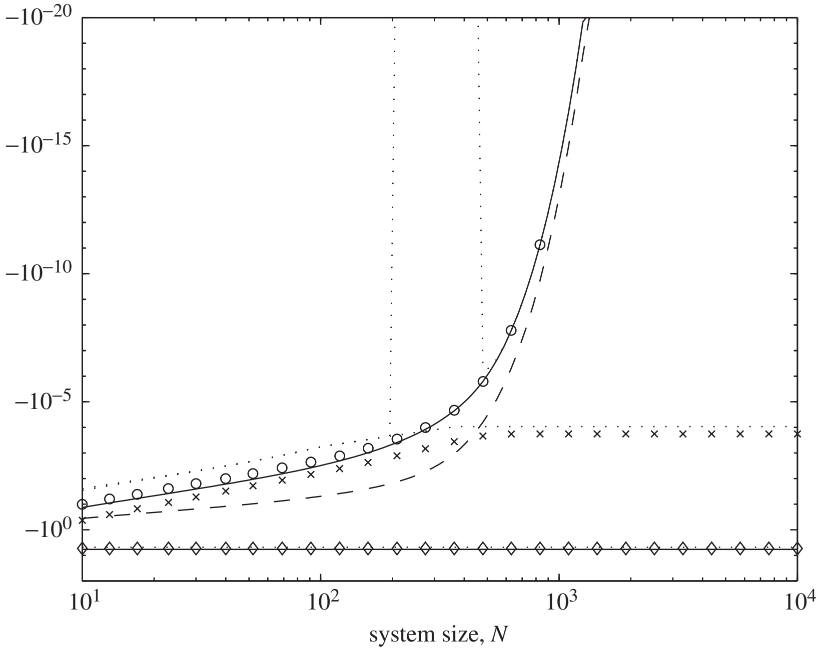

Figure 3. Similar to Figure 2, but with monomer concentration $z = 1.12$ fixed and system size varying (∘, λ_1; ×, λ_2; ◇, λ_{N-1}; ······, upper bounds; ——, lower bounds; – – –, Theorem 2.3 estimate).

small eigenvalues, and the calculations are very time consuming for large matrices. For these reasons we were not able to get direct numerical approximations of the smallest values of λ_1 in the figures, although the upper and lower bounds provide a very accurate value for λ_1 there.

Figures 2 and 3 show the estimate of λ_1 given in (3.3) and obtained from Theorem 2.3 in § 2. It overestimates the *size* of λ_1 (and hence underestimates the time-scale of the approach to equilibrium) over most of the range shown, but it does get better as z and/or N increase.

Figures 2 and 3 also show the sharper bounds on eigenvalue λ_1 given in Theorem 3.9. The theorem gives upper bounds which apply for different ranges of z. When $z < b_N$ the bounds (3.10) and (3.11) apply, and they are coincident with the upper bound on λ_2. At intermediate values of z, the best bound we have is $\lambda_1 < 0$, and this shows clearly between the near-vertical lines on the figures. When z and N are big enough, the lower bound (3.9) and the upper bound (3.12) are very tight indeed, because then $\hat{\boldsymbol{p}}$ given by (3.7) is a very good guess at the first eigenvector of S, and the quantity gap$'$ estimated by the difference between the lower bound on λ_1 and the upper bound on λ_2 is relatively large. At the larger values of z and N in the figures, the upper and lower bounds on λ_1 agree to within machine precision.

We plot the tightest bounds on λ_{N-1} and λ_2 from Theorems 3.5 and 3.7 in Figures 2 and 3 and see that they do not vary much in comparison with λ_1. In fact for large N

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

and fixed $z = 1.12$ in Figure 3, the upper bound on λ_2 reaches a constant value given by $\lambda_2^{(+)}(S)$ from Theorem 3.7 with $r = 239$ (the location of the maximum of the function $-a_r z - b_r + \sqrt{za_{r-1}b_r} + \sqrt{za_r b_{r+1}}$ given the rate constants (1.3)).

5. Concluding remarks

The behaviour of the Becker–Döring Equation (1.1) can be explained by looking at the extremal eigenvalues of its coefficient matrix M given by (1.5).

(1) The smallest eigenvalue λ_1 can become extremely small as the system size and/or monomer concentration increase, leading to very slow evolution of solutions (see, for example, Remarks 3.1 and 3.10).

(2) λ_1 can also become isolated from the rest of the spectrum (Remark 3.10), giving a long metastable period between time $|\lambda_2|^{-1}$ and $|\lambda_1|^{-1}$, where solutions can appear not to change. See the figures for illustrations of this.

The eigenvalues can be estimated in various ways with varying degrees of accuracy depending on the system parameters. However, for parameters (1.3) with large enough system size and/or monomer concentration, the upper and lower bounds on λ_1 in Theorem 3.9 agree to better than 15 significant digits.

Different versions of the Becker–Döring equations were described in § 1. One apparently minor variation on the system (1.1) is to replace the truncation $\dot{c}_N = J_{N-1}$ by $\dot{c}_N = J_{N-1} - za_N c_N$, and this system has been studied in [**7**, **8**] as an intermediate step to obtain estimates for the infinite Becker–Döring equations. It is interesting to note that the ODE system that results from this minor modification does not have an exponentially small eigenvalue (with parameters (1.3)) and so does not exhibit metastability for finite system size. The proof of this follows directly from that in Theorem 3.7 and Remark 3.8.

Finally, we note that our results cannot be used directly to estimate the duration of the metastable time-scale in the nonlinear (constant-density) version of the truncated equations. In the constant-density case the time-scale *decreases* as the metastable monomer concentration value increases (see [**3**, Figure 4.1]), while in the constant-monomer-concentration case the time-scale *increases* as monomer concentration increases. Suitable estimates for the constant-density finite-system-size case can be found in [**6**].

References

1. J. M. Ball, J. Carr and O. Penrose, The Becker–Döring equations: basic properties and asymptotic behaviour of solutions, *Commun. Math. Phys.* **104** (1986), 657–692.
2. R. Becker and W. Döring, Kinetische behandlung der keimbildung in übersättigten Dämpfern, *Annln Phys.* **24** (1935), 719–752.
3. J. Carr, D. B. Duncan and C. H. Walshaw, Numerical approximation of a metastable system, *IMA J. Numer. Analysis* **15** (1995), 505–521.
4. J. W. Demmel, *Applied numerical linear algebra* (SIAM, 1997).

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press

5. D. B. Duncan and A. R. Soheili, Approximating the Becker–Döring cluster equations, *Appl. Numer. Math.* **37** (2001), 1–29.
6. R. M. Dunwell, The Becker–Döring cluster equations, PhD Thesis, Department of Mathematics, Heriot-Watt University (1997).
7. M. Kreer, Classical Becker–Döring cluster equations: rigorous results on metastability and long-time behaviour, *Annln Phys.* **2** (1993), 398–417.
8. O. Penrose, Metastable states for the Becker–Döring cluster equations, *Commun. Math. Phys.* **124** (1989), 515–541.
9. O. Penrose and J. L. Lebowitz, Towards a rigorous molecular theory of metastability, in *Studies in statistical mechanics*, vol. VII, *Fluctuation phenomena* (ed. E. W. Montroll & J. L. Lebowitz), pp. 293–340 (North-Holland, Amsterdam, 1979, 1987).
10. J. H. Wilkinson, *The algebraic eigenvalue problem* (Oxford University Press, 1965).

https://doi.org/10.1017/S0013091500000882 Published online by Cambridge University Press