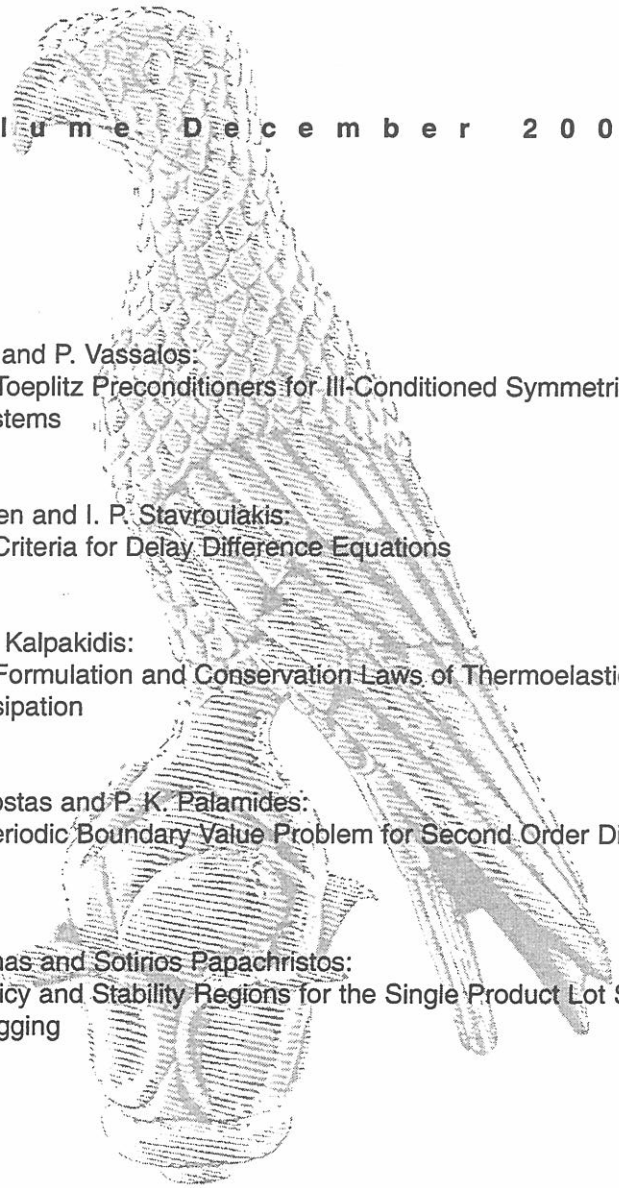
TECHNICAL REPORT

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

UNIVERSITY OF IOANNINA
DEPARTMENT OF MATHEMATICS



Number 4 Volume December 2000

NEW BAND TOEPLITZ PRECONDITIONERS FOR ILL-CONDITIONED SYMMETRIC POSITIVE DEFINITE TOEPLITZ SYSTEMS

D. NOUTSOS* AND P. VASSALOS†

Abstract. It is well known that Preconditioned Conjugate Gradient (PCG) methods are widely used to solve ill-conditioned Toeplitz linear systems $T_n(f)x = b$. In this paper we present a new preconditioning technique for the solution of symmetric Toeplitz systems generated by nonnegative functions f with zeros of even order. More specifically, f is divided by the appropriate trigonometric polynomial g of the smallest degree, with zeros the zeros of f, to eliminate its zeros. Using rational approximation we approximate $\sqrt{\frac{f}{g}}$ by $\frac{p}{q}$ and consider $\frac{p^2 g}{q^2}$ as a very satisfactory approximation of f. We propose the matrix $M_n = B_n^{-1}(p)B_n(p^2 g)B_n^{-1}(p)$ as a preconditioner whence a good clustering of the spectrum of its preconditioned matrix is obtained. We also show that the proposed technique can be very flexible, a fact that is confirmed by various numerical experiments so that in many cases it constitutes a much more efficient strategy than the existing ones.

Key words. low rank correction, Toeplitz matrix, conjugate gradient, rational interpolation and approximation, preconditioner

AMS subject classifications. Primary 65F10, 65F15

1. Introduction. In this paper we use and analyze band Toeplitz matrices as preconditioners for the solution of the $n \times n$ ill-conditioned symmetric and positive definite Toeplitz system

$$T_n(f)x = b \tag{1.1}$$

by the Preconditioned Conjugate Gradient (PCG) method, where the matrix $T_n(f) \in \mathbb{R}^{n\times n}$ is produced by a real-valued, even, 2π- periodic function defined in the fundamental interval $[-\pi, \pi]$. Then, the (j,k) element of $T_n(f)$ is given by the Fourier coefficient of f, i.e

$$T_n(f)_{j,k} = T_{j-k} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-\mathbf{i}(j-k)x}dx, \qquad 1 \le j,k < n,$$

where $\mathbf{i}$ is the imaginary unit.

Toeplitz matrices arise very often in a wide variety of applications, as e.g., in the numerical solution of differential equations using finite differences, in statistical problems (linear prediction), in Wiener-Hopf kernels, in Markov chains, in image and signal processing, e.t.c. (see [8], [3], [19]). The generating function f plays a significant role in the location and distribution of the eigenvalues of Toeplitz matrix [8], [4] and in many cases is a priori known. As it is known for the spectrum of $T_n(f)$ there holds $\sigma(T_n(f)) \subseteq [ess\inf f, ess\sup f]$.

Superfast direct methods can solve system (1.1) in $O(n\log^2 n)$ operations, but their stability properties for ill-conditioned Toeplitz matrices are still unclear; see, for instance, [3].

*Department of Mathematics, University of Ioannina, GR-451 10, Ioannina, Greece (dnoutsos@cc.uoi.gr).

†Department of Mathematics, University of Ioannina, GR-451 10, Ioannina, Greece (pvassal@pythagoras.math.uoi.gr). The research of this author was supported by Hellenic Foundation of Scholarships (HFS).

The classical iterative methods such as Jacobi, Gauss–Seidel and SOR are not effective since the associated spectral radius tends to 1 for large n. The method which is widely used for the solution of such systems is the PCG method. The factors that affect the convergence features of this method are the magnitude of the condition number $\kappa_2(T_n(f))$ and the distribution of the eigenvalues. So a good preconditioner must cluster the eigenvalues of the preconditioned system as much as possible and make the eigenvalues that might lie outside the cluster have magnitude independent of n.

If the generating function is continuous and positive then problem (1.1) will not be ill-conditioned and the condition number can not increase proportionally to n although it can be very large. In this case system (1.1) can be handled by using a preconditioner belonging to some Trigonometric matrix algebras (circulant ,τ, Hartley, [18], [17], [9]) or by band Toeplitz preconditioners with weakly increasing bandwidth defined by a polynomial operator $\mathcal{S}_n$ as was proposed in [16]. Theoretically, the latter class of preconditioners seems to perform better as $n \to \infty$ since the number of PCG iterations tends to 1 while in the former cases this number tends to a constant.

When f has any zeros, then system (1.1) is ill-conditioned and the condition number $\kappa_2(T_n(f))$ increases proportionally to n^α where α is the largest number of the multiplicities of the zeros of f [4], [14]. To best handle this case it is necessary to know the number of the zeros of f. If this number is not even then the most suitable technique for this situation [13], fails to make the condition number of the preconditioned matrix independent of its dimension n and the problem is still open. On the other hand things dramatically change when the number of zeros is even.

In this case, it was R. Chan [4] who first proposed as a preconditioner for system (1.1) the Toeplitz band matrix $B_n(g)$ whose generating function g is a trigonometric polynomial that has the same zeros with the same multiplicities as those of f. Next, in [5], g was not only considered as having the zeros of f but also its degree was increased so that it provided additional degrees of freedom to approximate f and to minimize the relative error $\|\frac{f-g}{g}\|_\infty$ over all trigonometric polynomials g of a fixed degree l. The generating function g is then computed by the Remez algorithm, which can be very expensive, from the computational point of view, especially when f has a large number of zeros.

Recently, Serra [15] has extended this method by proposing alternative techniques to minimize $\|\frac{f-g}{g}\|_\infty$. More specifically, he chose as g, $z_k g_{l-k}$ where z_k is the trigonometric polynomial of minimum degree k that has all the zeros of f with their multiplicities and g_{l-k} is the trigonometric polynomial of degree $l-k$ which is the best Chebyshev approximation of $\hat{f} = \frac{f}{z_k}$ from the space $\mathcal{P}_{l-k}$ of all trigonometric polynomials of degree at most $l-k$. In addition, in the same work [15], it was also proposed another way of constructing g_{l-k} by interpolating $\hat{f}$ at the $l-k+1$ zeros of the $(l-k+1)$-st degree Chebyshev polynomial of the first kind.

We remark that it has been proved [7], that preconditioners belonging to the aforementioned matrix algebra, when they can be defined, produce weak clustering, i.e., the eigenvalues of the preconditioned matrix are such that for every $\epsilon > 0$ there exists a positive β so that, except for rare exceptions, $O(n^\beta)$ of the eigenvalues lie in the interval $(0, \epsilon)$.

In this paper we extend the previous methods in order to achieve a better clustering for the eigenvalues of the preconditioned matrix and propose a way of constructing a class of preconditioners based on rational approximation or on interpolation to the positive and continuous function $\sqrt{\frac{f}{z_k}}$ with z_k defined previously.

The outline of the present work is as follows. In Section 2 we recall some useful issues about the rational approximation, while in Section 3 we introduce the technique of constructing the new class of preconditioners based on rational approximation to $\sqrt{\frac{f}{z_\rho}}$ and analyze the convergence of the PCG method. In Section 4 we study the flexibility and possible modifications of our method, analyze its cost per iteration and compare it with that of previous techniques. Finally, in Section 5, results of illustrative numerical experiments are exhibited and concluding remarks are made.

2. Preliminaries. In what follows we assume that the generating function f is defined in $[-\pi, \pi]$, is 2π- periodic, continuous, nonnegative and has zeros of even order.

We define by z_k the trigonometric polynomial of minimum degree k containing all the zeros of f with their multiplicities. Then we define $r_{lm} = \frac{p_l}{q_m}$ as the best rational approximation of $\hat{f} = \sqrt{\frac{f}{z_k}}$ in the uniform norm, i.e.,

$$\|\hat{f} - r_{lm}\|_\infty = \min_{r \in \mathcal{R}(l,m)} \|\hat{f} - r\|_\infty,$$

where $\mathcal{R}(l, m)$ denotes the set of rational functions r, with $p \in \mathcal{P}_l$, $q \in \mathcal{P}_m$ and r is irreducible, that is p and q have no zeros in common.

It is known that when f belongs to some special class of functions [10] then the order of magnitude of the maximum error of an approximation from the space $\mathcal{R}(l, m)$ is better than the corresponding error in the space $\mathcal{P}(l + m)$. In general, we hope that taking advantage of the flexible nature of rational functions this set will be a stronger tool than its competitor the polynomial one. For example, it is obvious that polynomials are not suitable for approximating functions having sharp peaks near the center of their ranges and are slowly varying when $|x|$ increases. Such kind of behavior can be obtained by continuous functions which are not differentiable at some points. However, it is easy to overcome this difficulty by using rational functions.

The next theorem establishes the fact that rational approximation of continuous functions in $[-\pi, \pi]$ is always possible and unique.

THEOREM 2.1. *Let f in $C[-\pi, \pi]$. Then there exists $r^* \in \mathcal{R}(l, m)$ such that*

$$\|f - r^*\| < \|f - r\|$$

for all $r \in \mathcal{R}(l, m)$, $r \neq r^*$.

Proof. See [12], pp. 121,125. □

3. Construction of the Preconditioner. Let f be a $2\pi-$periodic, nonnegative function belonging to $C[-\pi, \pi]$ with zeros $x_1, x_2, \cdots, x_s$ of multiplicities $2\mu_1, 2\mu_2, \cdots, 2\mu_s$, respectively, and $2\mu_1 + 2\mu_2 + \cdots + 2\mu_s = \rho$. First, we define

$$z_\rho = \prod_{i=1}^{s} (1 - \cos(x - x_i))^{\mu_i}$$

which is the trigonometric polynomial of minimum degree ρ having all the zeros of f. By dividing f by z_ρ, all its zeros are eliminated and the ratio $\frac{f}{z_\rho}$ becomes a real positive function.

Then, we define the function $\hat{f} = \sqrt{\frac{f}{z_\rho}}$ and approximate it with the rational trigonometric function $r_{l,m} = \frac{p_l}{q_m}$ where l, m are the degrees of the numerator and the

denominator, respectively. Since $\frac{p_l}{q_m}$ is the best rational approximation of $\sqrt{\frac{f}{z_\rho}}$ for certain l and m we are led to the conclusion that $\frac{p_l^2}{q_m^2}$ may be a good approximation of $\frac{f}{z_\rho}$. This means that there exists a small $\epsilon > 0$ such that

$$\left\| \frac{f}{z_\rho} - \frac{p_l^2}{q_m^2} \right\|_\infty < \epsilon$$

or, equivalently, that there exists a small $\delta > 0$ such that

$$\left\| \frac{q_m^2}{z_\rho p_l^2} f - 1 \right\|_\infty < \delta.$$

The last inequality means that the values of $\frac{q_m^2}{z_\rho p_l^2} f$ are clustered in a small region near the constant number 1. In matrix analog, this means that taking $T_n\left(\frac{z_\rho p_l^2}{q_m^2}\right)$ as a preconditioner matrix for the solution of (1.1), the eigenvalues of $T_n^{-1}\left(\frac{z_\rho p_l^2}{q_m^2}\right) T_n(f)$ are clustered in a small region near 1 and the PCG method will become very fast. Unfortunately, this matrix is a full Toeplitz matrix, is hard to construct, is costly to invert and so it is useless as a preconditioner. Instead, we are led to the idea of separating the numerator and the denominator of the ratio $\frac{z_\rho p_l^2}{q_m^2}$ and use as a preconditioner matrix the product of three band Toeplitz matrices. More specifically, the preconditioner we propose for the solution of system (1.1) is

$$M_n = B_{nm}^{-1}(q) B_{n\hat{l}}(p^2 z_\rho) B_{nm}^{-1}(q), \qquad \hat{l} = 2l + \rho, \tag{3.1}$$

where the second index in the matrices represents their halfbandwidth, while the first one their dimension. The following statements prove the basic assumptions a preconditioner must satisfy and also describe the spectrum of the preconditioned matrix $M_n^{-1} T_n$.

THEOREM 3.1. *The matrix M_n is symmetric and positive definite for every n.*

Proof. Its symmetry is implied directly from the definition (3.1). On the other hand, the eigenvalues of $B_{n\hat{l}}(p^2 z_\rho)$ belong to the interval $(\min p_l^2 z_\rho, \max p_l^2 z_\rho)$, where $0 = \min p_l^2 z_\rho < \max p_l^2 z_\rho \leq 2^\rho$. Therefore, $B_n(p_l^2 z_\rho)$ is symmetric and positive definite. Furthermore, q_m has no zeros in $[-\pi, \pi]$ because it results from the rational approximation to a function which is strictly positive in $[-\pi, \pi]$. So, $B_{nm}(q)$ is symmetric and invertible. Then, for every $x \in \mathbb{R}^n$, $x \neq 0$, we have

$$x^T M_n x = x^T B_{nm}^{-1}(q) B_{n\hat{l}}(p^2 z_\rho) B_{nm}^{-1}(q) x = y^T B_{n\hat{l}}(p^2 z_\rho) y > 0,$$

where $y = B_{nm}^{-1}(q) x$. Hence M_n is symmetric and positive definite. □

Theorem 3.1 suggests that the matrix M_n can be taken as a preconditioner matrix. It then remains to study the convergence rate of the PCG method or, equivalently, how the eigenvalues of the matrix $M_n^{-1} T_n$ are distributed. For this, we give without proof the following Lemma and then we state and prove our main result in Theorem 3.2.

LEMMA 3.1. *Suppose $A, B \in \mathbb{R}^{n \times n}$ are symmetric matrices such that*

$$A = B + \epsilon c c^T,$$

where $c \in \mathbb{R}^n$, $c^T c = 1$. *If* $\epsilon > 0$ *then*

$$\lambda_1(B) \leq \lambda_1(A) \leq \lambda_2(B) \leq \cdots \leq \lambda_n(B) \leq \lambda_n(A)$$

while if $\epsilon \leq 0$, *then*

$$\lambda_1(A) \leq \lambda_1(B) \leq \lambda_2(A) \leq \cdots \leq \lambda_n(A) \leq \lambda_n(B)$$

provided that the eigenvalues are labeled in nondecreasing order of magnitude. In either case

$$\lambda_k(A) = \lambda_k(B) + t_k\epsilon, \qquad k = 1, 2, \cdots, n,$$

where $t_k \geq 0, \quad k = 1, 2, \cdots, n,$ *and* $\sum_{k=1}^{n} t_k = 1$.

Proof. See Wilkinson [20], pp. 97-98. □

THEOREM 3.2. *Let* $\lambda_i(M_n^{-1}T_n)$, $i = 1(1)n$, *denote the eigenvalues of* $M_n^{-1}T_n$ *and* m *the degree of the denominator* q_m *of the rational approximation. Then, at least* $n - 4m$ *eigenvalues of the preconditioned matrix lie in* $(h_{\min}, h_{\max})$, *at most* $2m$ *are greater than* $h_{\max}$ *and at most* $2m$ *are in* $(0, h_{\min})$, *where* $h = \frac{fq^2}{p^2 z_\rho}$.

Proof. Obviously the matrix

$$M_n^{-1}T_n = B_{nm}(q)B_{n\hat{l}}^{-1}(p^2 z_\rho)B_{nm}(q)T_n(f)$$

is similar to the matrix

$$B_{n\hat{l}}^{-\frac{1}{2}}(p^2 z_\rho)B_{nm}(q)T_n(f)B_{nm}(q)B(p^2 z_\rho)_{n\hat{l}}^{-\frac{1}{2}}. \tag{3.2}$$

Then, since $B_{nm}(q)$ is a band matrix with halfbandwidth m, the matrix

$$B_{nm}(q)T_n(f)B_{nm}(q)$$

can be written as a sum of a Toeplitz matrix and a low rank correction matrix, i.e.,

$$B_{nm}(q)T_n(f)B_{nm}(q) = T_n(q^2 f) + \Delta, \tag{3.3}$$

where Δ is a symmetric 'border' matrix with nonzero elements only in the first and last m rows and columns. So $rank(\Delta) \leq 4m$ is independent of n. Then, from (3.2) and (3.3) we obtain that

$$\begin{aligned} \overbrace{B_{n\hat{l}}^{-\frac{1}{2}}(p^2 z_\rho)B_{nm}(q)T_n(f)B_{nm}(q)B_{n\hat{l}}^{-\frac{1}{2}}(p^2 z_\rho)}^{\mathrm{E}} &= \overbrace{B_{n\hat{l}}^{-\frac{1}{2}}(p^2 z_\rho)T_n(q^2 f)B_{n\hat{l}}^{-\frac{1}{2}}(p^2 z_\rho)}^{\widetilde{\mathrm{E}}} \\ &\quad + B_{n\hat{l}}^{-\frac{1}{2}}(p^2 z)\Delta B_{n\hat{l}}^{-\frac{1}{2}}(p^2 z). \end{aligned} \tag{3.4}$$

Since a matrix product does not have rank larger than that of each of the factors involved, there exist $\alpha_i > 0$, $c_i \in \mathbb{R}^n$, $i = 1(1)m_+$, and $\beta_i > 0$, $d_i \in \mathbb{R}^n, i = 1(1)m_-$, with $m_+ + m_- \leq 4m$, such that (3.4) can be written as

$$E - \widetilde{E} = \sum_{i=1}^{m_+} \alpha_i c_i c_i^T - \sum_{i=1}^{m_-} \beta_i d_i d_i^T.$$

So applying successively $m_+ + m_-$ times Lemma 3.1 gives

$$h_{\min} \leq \lambda_i(E) \leq h_{\max}, \qquad m_- < i \leq n - m_+,$$

and the theorem is proved. □

It is clear from the previous analysis and statements that contrary to what happens with other band Toeplitz preconditioners, the one we propose of the 'premultiplier' matrix $B_{nm}(q)$, may make some of the eigenvalues lie outside the approximation interval $[h_{\min}, h_{\max}]$. We will prove now that the spectral radius of the preconditioned matrix is bounded by a constant number independent of n. For this, first, we state and prove the following lemma.

LEMMA 3.2. *Let B_n be a $n \times n$ symmetric and positive definite band Toeplitz matrix with halfbandwidth s. Then the $k \times k$ principal and trailing submatrices of B_n^{-1} as well as the $k \times k$ submatrices consisting from the first k rows and the last k columns (right upper corner) or from the last k rows and the first k columns (left lower corner) of B_n^{-1}, are bounded for every fixed k independent of n .*

Proof. For principal and trailing submatrices, this property has been proved in [6] for $k = s$. We will prove the validity of this property for $k = s + 1$ and the proof of every fixed k can be completed by induction. From the foundamental relation

$$\sum_{l=1}^{s+1} b_{1l}(B_n^{-1})_{lj} = \delta_{1j},$$

where δ_{1j} is the Kroneker δ, we obtain successively that

$$(B_n^{-1})_{s+1,j} = \frac{1}{b_{1,s+1}} \left(\delta_{1j} - \sum_{l=1}^{s} b_{1l}(B_n^{-1})_{lj} \right), \; j = 1, 2, \ldots, s. \tag{3.5}$$

Since all the elements in the righthand side of (3.5) are bounded, so are the elements $(B_n^{-1})_{s+1,j}$, $j = 1, 2, \cdots, s$. From the symmetry of B_n^{-1} we obtain that the elements $(B_n^{-1})_{j,s+1}$, $j = 1, 2, \cdots, s$, are also bounded. One more application of (3.5) for $j = s + 1$, gives us that the element $(B_n^{-1})_{s+1,s+1}$ is bounded and the proof for the principal submatrices is complete. Since, B_n^{-1} is a persymmetric matrix the elements of the trailing matrix are the same as those of the principal one in reverse order. So the $k \times k$ trailing matrix is also bounded.

It remains to prove the validity of the property for the submatrices in the right upper corner and in the left lower corner of B_n^{-1}. These matrices are transposes of each other due to the symmetry of B_n^{-1}. From the positive definiteness of B_n^{-1} we have that

$$|(B_n^{-1})_{ij}| < \frac{(B_n^{-1})_{ii} + (B_n^{-1})_{jj}}{2}, \; i = 1, \ldots, k, \;\; j = n - k + 1, \ldots, n.$$

The elements in the righthand side are the diagonal elements of the $k \times k$ principal and trailing submatrices, respectively, which are bounded and the proof is complete. □

The following theorem proves that the eigenvalues of $M^{-1}T$ have an upper bound.

THEOREM 3.3. *Under the assumptions of Theorem 3.2 there exists a constant c, independent of n, such that $\rho\left(M_n^{-1}T_n(f)\right) \leq c$, for every n.*

Proof. We begin the proof by using some relations connecting the spectral radii and the Rayleigh quotients of symmetric matrices. The fact that all the matrices are

positive definite, is also used.

$$
\begin{aligned}
\rho\left(M_n^{-1}T_n(f)\right) &= \rho\left(B_{nm}(q)B_{n\hat{l}}^{-1}(p^2z_\rho)B_{nm}(q)T_n(f)\right)\\
&= \rho\left(B_{n\hat{l}}^{-\frac{1}{2}}(p^2z_\rho)B_{nm}(q)T_n(f)B_{nm}(q)B_{n\hat{l}}^{-\frac{1}{2}}(p^2z_\rho)\right)\\
&= \max_{x\neq 0}\frac{x^TB_{n\hat{l}}^{-\frac{1}{2}}(p^2z_\rho)B_{nm}(q)T_n(f)B_{nm}(q)B_{n\hat{l}}^{-\frac{1}{2}}(p^2z_\rho)x}{x^Tx}\\
&= \max_{x\neq 0}\left(\frac{x^TT_n(f)x}{x^TB_{nm}^{-1}(q)B_{n\hat{l}}(p^2z_\rho)B_{nm}^{-1}(q)x}\cdot\frac{x^TB_{n\hat{l}}(p^2z_\rho)x}{x^TB_{n\hat{l}}(p^2z_\rho)x}\right)\\
&= \max_{x\neq 0}\left(\frac{x^TT_n(f)x}{x^TB_{n\hat{l}}(p^2z_\rho)x}\cdot\frac{x^TB_{n\hat{l}}(p^2z_\rho)x}{x^TB_{nm}^{-1}(q)B_{n\hat{l}}(p^2z_\rho)B_{nm}^{-1}(q)x}\right) \qquad (3.6)\\
&\leq \max_{x\neq 0}\frac{x^TT_n(f)x}{x^TB_{n\hat{l}}(p^2z_\rho)x}\cdot\max_{x\neq 0}\frac{x^TB_{n\hat{l}}(p^2z_\rho)x}{x^TB_{nm}^{-1}(q)B_{n\hat{l}}(p^2z_\rho)B_{nm}^{-1}(q)x}\\
&= M_1\max_{x\neq 0}\frac{x^TB_{nm}(q)B_{n\hat{l}}(p^2z_\rho)B_{nm}(q)x}{x^TB_{n\hat{l}}(p^2z_\rho)x}\\
&= M_1\max_{x\neq 0}\frac{x^T\left(B_{n,\hat{l}+2m}(q^2p^2z_\rho)+\Delta\right)x}{x^TB_{n\hat{l}}(p^2z_\rho)x}\\
&\leq M_1\left(M_2+\max_{x\neq 0}\frac{x^T\Delta x}{x^TB_{n\hat{l}}(p^2z_\rho)x}\right)\\
&= M_1\left(M_2+\rho\left(B_{n\hat{l}}^{-1}(p^2z_\rho)\Delta\right)\right).
\end{aligned}
$$

In (3.6) we have taken $M_1=\max_{x\neq 0}\frac{x^TT_n(f)x}{x^TB_{n\hat{l}}(p^2z_\rho)x}=\rho\left(B_{n\hat{l}}^{-1}(p^2z_\rho)T_n(f)\right)$ and $M_2=\max_{x\neq 0}\frac{x^TB_{n,\hat{l}+2m}(q^2p^2z_\rho)x}{x^TB_{n\hat{l}}(p^2z_\rho)x}=\rho\left(B_{n\hat{l}}^{-1}(p^2z_\rho)B_{n,\hat{l}+2m}(q^2p^2z_\rho)\right)$ which are bounded, since the generating functions $\frac{f}{p^2z_\rho}$ and $\frac{q^2p^2z_\rho}{p^2z_\rho}=q^2$, respectively, are bounded functions in $[-\pi,\pi]$. In (3.6), the matrix product $B_{nm}(q)Bn\hat{l}(p^2z_\rho)B_{nm}(q)$ was written as the band Toeplitz matrix $B_{n,\hat{l}+2m}(q^2p^2z_\rho)$, generated by the function $q^2p^2z_\rho$, plus the low rank correction matrix Δ.

It is known [2] that the matrix Δ is given by

$$\Delta=B_{nm}(q)H(q)H(p^2z_\rho)+B_{nm}(q)H^R(q)H^R(p^2z_\rho)+H(q)H(qp^2z_\rho)+H^R(q)H^R(qp^2z_\rho),$$

where $H(q)$, $H(p^2z_\rho)$ and $H(qp^2z_\rho)$ are Hankel matrices produced by the trigonometric polynomials q, p^2z_ρ and qp^2z_ρ, respectively, while H^R denotes the matrix obtained from H by reversing the order of its rows and columns.

It is obvious that Δ is a low rank correction matrix that has nonzero elements

only in the upper left and lower right triangles as this is illustrated below

$$\Delta = \begin{pmatrix} * & \cdots & * & 0 & \ldots & 0 \\ \vdots & \ddots & 0 & \ddots & 0 & \vdots \\ * & 0 & \ddots & 0 & & 0 \\ 0 & \ddots & 0 & \ddots & 0 & * \\ \vdots & 0 & & 0 & \ddots & \vdots \\ 0 & \ldots & 0 & * & \ldots & * \end{pmatrix}.$$

It is clear that the elements of Δ are bounded and the size of the triangles depends only on the bandwidths m and $\hat{l}$ and are independent of n.

It remains to prove that $\rho\left(B_{n\hat{l}}^{-1}(p^2 z_\rho)\Delta\right)$ is bounded. For this, we write the matrices in the following block forms

$$B_{n\hat{l}}^{-1}(p^2 z_\rho) = \begin{pmatrix} B_1 & * & B_2 \\ * & * & * \\ B_2^T & * & B_1^R \end{pmatrix}, \qquad \Delta = \begin{pmatrix} D & & \\ & O & \\ & & D^R \end{pmatrix},$$

where B_1, B_2 are $k \times k$ matrices if D has k nonzero anti-diagonals.

Since the only nonzero columns of the matrix $B_{n\hat{l}}^{-1}(p^2 z_\rho)\Delta$ are its first k and last k ones, the nonidentically zero eigenvalues of $B_{n\hat{l}}^{-1}(p^2 z_\rho)\Delta$ will be the eigenvalues of the matrix

$$\begin{pmatrix} B_1 D & B_2 D^R \\ B_2^T D & B_1^R D^R \end{pmatrix}.$$

In view of Lemma 3.2 this matrix is bounded and so are its eigenvalues which proves the present statement. □

So, the eigenvalues that are greater than $h_{\max}$, have an upper bound. An open question remains regarding the eigenvalues that may lie in the interval $(0, h_{\min})$. However, strong numerical evidence suggests that in the spectrum of the preconditioned matrix obtained by our approach (see Figures 5.1, 5.2, 5.3), these eigenvalues have a lower bound independent of n. Moreover, as one can see from Figures (5.1(b)-(d), 5.2(b), 5.3(b)), the out of the main interval eigenvalues appear in pairs. In addition, the elements of each pair tend to each other as n tends to infinity. In view of this observation the convergence analysis of the PCG method in [1] assures us that our method will not be seriously affected and the convergence of it will remain superlinear which is the optimal cost for this method.

4. Computational analysis and modifications of the method. In this section we will try to compare, from the computational point of view, our preconditioner with the most recent band-Toeplitz preconditioner proposed in [15]. The latter has in general the best performance from all the previous ones, when the generating function f is nonnegative and has zeros of even order.

The main computational cost in every PCG iteration is due to the Toeplitz matrix-vector product $T_n(f)x$ and to the solution of a system with coefficient matrix the preconditioner itself. The first one is the same for both methods and can be computed by means of Fast Fourier Transform (FFT) in $30(n \log 2n)$ operations (ops) in

a sequential machine or in $O(\log 2n)$ steps in the parallel PRAM model of computation, when $O(n)$ processors are used. For the inversion of the preconditioners things slightly change. If we use band Toeplitz preconditioners then their halfbandwidth $\hat{l}_1$ represents the degree l_1 of the Chebyshev approximation plus the degree ρ of the trigonometric polynomial which eliminates the zeros of f. The inversion of such type of matrices can be achieved using the LDL^T factorization method in $n(\hat{l}_1^2 + 8\hat{l}_1 + 1)$ ops. We mention that this method is preferable from the band Cholesky procedure because the latter requires the computation of n square roots, which is quite expensive when n is large.

In the case of our preconditioner the inversion requires two band matrix vector products of total cost $n(4m + 2)$ ops, where m is the halfbandwidth and coincides with the degree of the denominator in the rational approximation. In addition, the inversion of $B_{n,\hat{l}_2}$, as in the previous case, can be performed in $n(\hat{l}_2^2 + 8\hat{l}_2 + 1)$ ops, where $\hat{l}_2 = \rho + 2l_2$ and l_2 represents the degree of the numerator of the rational approximation. So the total cost per iteration for this step of the algorithm of the PCG method is about

$$Cost_{it} = n(\hat{l}_2^2 + 8\hat{l}_2 + 4m + 3).$$

When n is large, the complexity of the method is strongly dominated by the first step which requires $O(n \log 2n)$ ops and the methods are essentially equivalent in complexity per iteration. Thus the costs of finding $B_{n,\hat{l}_1}^{-1}$ and $B_{n,m}B_{n,\hat{l}_2}^{-1}B_{n,m}$, where $l_1 = l_2 + m$, are comparable.

In case n is not large enough, taking $l_2 = \frac{l_1}{2} - 1$ and making some calculations, we can see that the two preconditioning strategies are approximately equivalent even when $m = \rho l_1$.

According to this observation, if we have two candidates of rational approximations of f with almost the same relative error and degrees (l_1, m_1), (l_2, m_2) with $l_1 + m_1 \approx l_2 + m_2$, it is preferable, from the computation point of view, to choose as the generating function for our preconditioner the one which has the larger m and smaller l.

Finally, we will focus on the calculation of rational approximation of degree (l, m) of a positive continuous function f. In the recent literature many different strategies that produce this kind of approximation [11] can be found. Each of them is most suitable for certain classes of functions but the one which is based on the Remez algorithm seems to be, in general, quite efficient for a large variety of functions. The starting point of this category of algorithms is to construct a rational approximation using rational interpolation and then this rational approximation is used to generate a better approximation until an alternative set of $m + l + 2$ points is achieved. This procedure consists of adjusting the choice of the interpolation points in such a way as to ensure that the relative error decreases. In practice this method can fail in some cases. Usually, problems are caused either from the fact that the extreme values of the relative error occur more than $m + l + 2$ times, or the starting rational interpolation has zeros in the interval in which this approximation is sought. The first difficulty is usually overcome by seeking a rational approximation of a different degree or by designing a more robust algorithm. A trick that often works in the latter case is, instead of asking again for a rational approximation of a different degree, to start with an approximation that is valid over a shorter interval and use it as a starting point for an approximation on a slightly larger interval. Iterative application of this procedure may enable us to obtain a final approximation in the desired interval.

TABLE 5.1
Number of iterations for $f_1(x)$

n	B_n^{*1}	$\hat{B}_n^1$	B_n^{*3}	$\hat{B}_n^3$	B_n^{*4}	$\hat{B}_n^4$	$M_n^{0,1}$	$R_n^{0,1}$	$M_n^{1,1}$	$R_n^{1,1}$	$M_n^{1,2}$	$R_n^{1,2}$
16	9	8	9	7	7	6	8	7	6	6	5	5
32	10	10	11	8	9	7	10	9	7	7	6	6
64	13	12	11	10	9	8	11	11	9	9	8	8
128	15	15	12	11	10	10	12	13	11	11	10	10
256	16	16	12	13	10	10	13	13	12	12	11	11
512	16	16	13	13	10	11	13	14	13	13	11	12

For the convergence rate of the approximation method we can not give a theoretical result, but the facts that its computational cost is independent of n and the computations are done only once for a given function make us believe that this problem does not play an important role in the whole procedure.

4.1. Modifications of the method. The idea of constructing a preconditioner from a rational approximation of a function can be used in exactly the same way in case of rational interpolation at the Chebyshev points. The advantage of this modification is the easiness of its calculation. Nevertheless, it is worth noticing that we can not assure that this interpolation would not have zeros in the interval of approximation. Despite this, whenever the preconditioning gives us poor results, this technique may give, at least for certain classes of f, results similar to the corresponding ones by the best Chebyshev approximation.

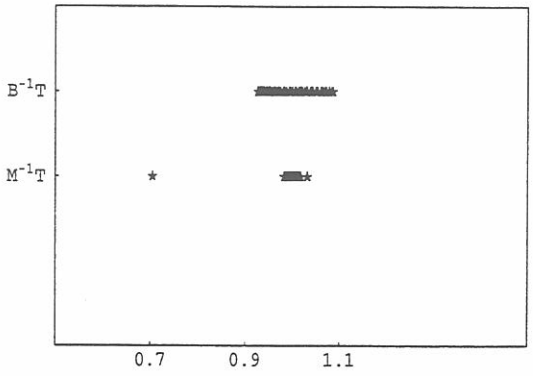
Another modification of this kind of preconditioning would be the following. First, we approximate the function $\frac{f}{z_\rho}$ by a rational approximation $\frac{p_l}{r_k}$, where k can be very large. Then we approximate the function $\sqrt{r_k}$ using a polynomial Chebyshev approximation q_m. Finally, the ratio $\frac{p_l}{q_m^2}$ is considered as an approximation of $\frac{f}{z_\rho}$. So, the preconditioner matrix $\tilde{M}$ for the solution of (1.1) would be

$$\tilde{M}_n = B_{nm}^{-1}(q) B_{n\hat{l}}(pz) B_{nm}^{-1}(q), \quad \hat{l} = l + \rho, \tag{4.1}$$

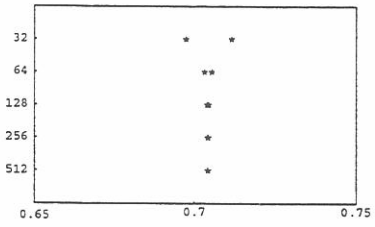
instead of M_n in (3.1). After this, all the previous theory developed holds the same.

The main point of this method is to approximate directly $\frac{f}{z_\rho}$ instead of $\sqrt{\frac{f}{z_\rho}}$ and possibly with a polynomial of higher degree in the denominator. Then considering that this can take care of every possible abnormalities of f, we approximate the denominator by a polynomial of lower degree by the Chebyshev technique. We remark here, that numerical experiments show that this matrix is not in general so good as a preconditioner compared with M_n or with the band-Toeplitz preconditioner obtained in [15]. This is because we make approximations in two levels. First, we take the rational approximation and then the Chebyshev approximation of the square root of the denominator of the first approximation. So, the overall approximation error seems to become much larger.

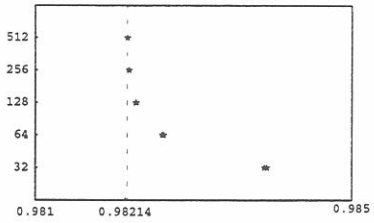
5. Numerical examples and concluding remarks. In this section, we present some numerical examples. The aim of these examples is twofold: i) to show, by numerical evidence, the correctness of our observations regarding the asymptotical spectral analysis of the preconditioned matrices and ii) to compare the convergence rate of our preconditioner with that of the band Toeplitz preconditioner proposed in [15]. We



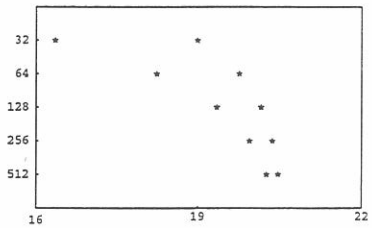
(a) The main mass of the eigenvalues of the preconditioned matrices



(b) The lower extreme pair



(c) The second upper pair



(d) The upper extreme pair

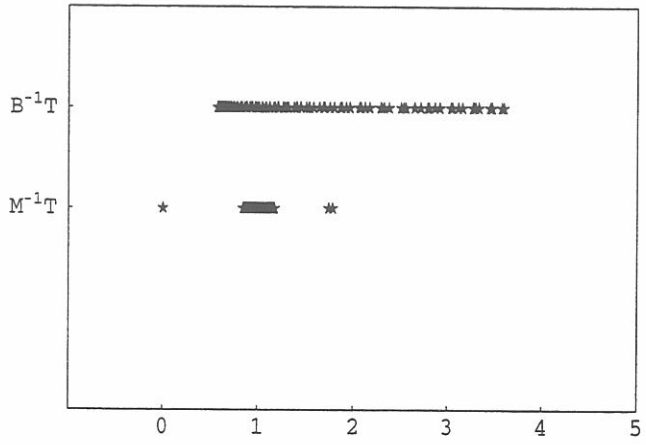
FIG. 5.1. *Spectra of* $(M_n^{2,2})^{-1}T_n(f_1)$ *and* $(B_n^{*5})^{-1}T_n(f_1)$ *for* $n = 128$ *and behavior of the pairs of eigenvalues that lie outside the interval* $[h_{\min}, h_{\max}]$ *with* $h_{\min} = 0.98214$

TABLE 5.2
Number of iterations for $f_2(x)$

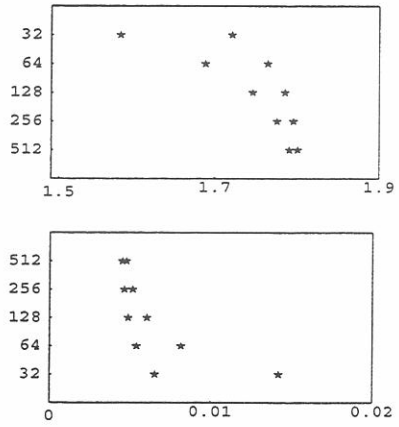
n	B_n^{*3}	B_n^{*4}	B_n^{*5}	B_n^{*6}	$M_n^{1,1}$	$R_n^{2,2}$
16	8	8	7	8	8	6
32	13	13	12	11	11	7
64	19	18	15	13	12	9
128	24	19	17	14	12	11
256	25	21	18	15	13	13
512	27	22	18	16	14	14

use the latter to compare it with ours because it is the most efficient technique for preconditioning Toeplitz matrices generating by functions with zeros of even order. Our test functions are the following

$$i)\ \ f_1(x) = x^4, \qquad ii)\ \ f_2(x) = \frac{2x^4}{1 + 25x^2}$$

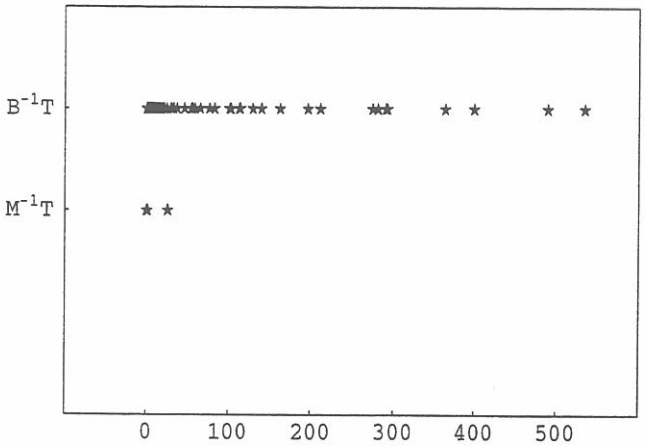


(a)

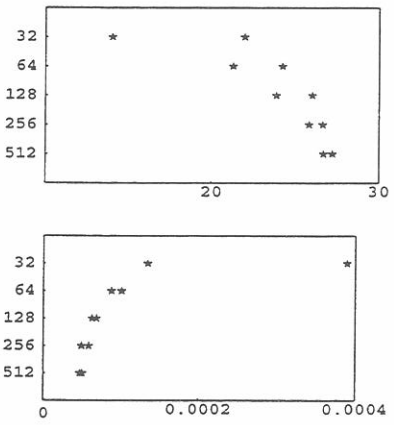


(b) The two pairs of extreme eigenvalues

FIG. 5.2. *Spectra of* $(M_n^{1,1})^{-1}T_n(f_2)$ *and* $(B_n^{*3})^{-1}T_n(f_2)$ *for* $n = 128$ *and behavior of the pairs of eigenvalues that lie outside the interval* $[h_{min}, h_{max}]$



(a)



(b) The two pairs of extreme eigenvalues

FIG. 5.3. *Spectra of* $(M_n^{1,2})^{-1}T_n(f_3)$ *and* $(B_n^{*3})^{-1}T_n(f_3)$ *for* $n = 256$ *and behavior of the pairs of eigenvalues that lie outside the interval* $[h_{min}, h_{max}]$

and

$$iii)\quad f_3(x) = \begin{cases} (x-3)^4(x-1)^2 & 0 \le x \le \pi, \\ (x+3)^4(x+1)^2 & -\pi \le x \le 0. \end{cases}$$

An effort was made to choose functions of different behaviors which produce ill-

TABLE 5.3
Number of iterations for $f_3(x)$.

n	B_n^{*3}	B_n^{*5}	B_n^{*7}	$M_n^{1,2}$	$R_n^{(1,2)}$
16	9	7	7	9	8
32	17	14	13	18	11
64	34	28	22	21	14
128	65	48	36	21	20
256	111	69	54	23	24
512	152	93	66	23	27

conditioned matrices T_n. The Toeplitz matrices produced have Euclidean condition numbers of order $O(n^4)$. In our experiments we solve the system $T_n(f)x = b$ where b is the vector having all its components equal to one. As a starting initial guess of solution the zero vector is used and as a stopping criterion the validity of $\frac{\|r_k\|_2}{\|r_0\|_2} \leq 10^{-7}$ is considered, where r_k is the residual vector after k iterations. The matrices and the rational approximations were performed using Mathematica in order to have more accurate results while all the other computations were performed using Matlab.

In the Tables we report the number of iterations needed until convergence is achieved in each case, B_n^{*l} denotes the optimal band Toeplitz preconditioner [15] which is generated by the trigonometric polynomial $z_\rho g_l$, with g_l being the best Chebyshev approximation of $\frac{f}{z_\rho}$ out of $\mathcal{P}_l$, $\hat{B}_n^l$ is the band Toeplitz preconditioner where $\hat{g}_l$ is the interpolation polynomial at the Chebyshev points, $M_n^{l,m}$ denotes our main proposed preconditioner obtained by the best rational approximation procedure of degree (l, m) and $R_n^{l,m}$ denotes the preconditioner that results after applying rational interpolation of degree (l, m).

In Figures 5.1(a), 5.2(a), 5.3(a), the spectra of the matrices $M_n^{-1}T_n(f_i)$, $i = 1, 2, 3$, are illustrated, while in 5.1(b)-(d), 5.2(b), 5.3(b) we focus on the behavior of the pairs of eigenvalues of the matrix lying outside the interval $[h_{\min}, h_{\max}]$ for different values of n. The boundness and the convergence in pairs is obvious in all figures. Especially, we stress the case of figures (5.1) and (5.3) where as we expected from the theory at most eight eigenvalues would lie outside the interval $[h_{\min}, h_{\max}]$ but in practice, for the first test function, only three pairs of eigenvalues lie outside this interval, one of which (the second lower pair) moves very close to the lower bound $h_{\min} = 0.98214$ while, for the third test function, only two pairs lie outside this interval. Finally, we remark that in the case of f_3 and for $n = 512$, the preconditioning by band Toeplitz B^{*3} "clusters" the eigenvalues of the preconditioned matrix in $[0.5, 584.3]$, B^{*5} in $[0.36, 104.7]$ while $M^{1,2}$ collects the main mass of them in $[0.67, 1.65]$ and $R^{1,2}$ collects it in $[0.95, 14.25]$.

REFERENCES

[1] O.AXELSSON AND G.LINDSKÖG, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 52 (1986), pp. 499-523.

[2] A. BÖTTCHER AND B. SILBERMANN, *Introduction to Large Truncated Toeplitz Matrices*, Springer Verlag, 1998.

[3] J. R. BUNCH, *Stability of methods for solving Toeplitz systems of equations,* SIAM J. Sci. Stat. Comput., 6 (1985), pp. 349-364.

[4] R. CHAN, *Toeplitz Preconditioners for Toeplitz Systems with Nonnegative Generating Functions,* SIAM J. of Numer. Anal., 11 (1991), pp. 333-345.

[5] R. CHAN AND P. TANG, *Fast Band-Toeplitz preconditioners for Hermitian Toeplitz systems,* SIAM J. Sci. Comp., 15 (1994), pp. 164-171.

[6] F. DI BENEDETTO, *Analysis of Preconditioning Techniques for ill-conditioned Toeplitz matrices,* SIAM J. Sci. Comput., 16 (1995), pp. 682-697.

[7] F. DI BENEDETTO AND S. SERRA, *A unifying approach to abstract matrix algebra preconditioning,* Numer. Math., 82 (1999), pp. 57-90.

[8] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications,* 2nd edition, Chelsea, New York, 1984.

[9] X.A. JIN, *Hartley Preconditioners for Toeplitz Systems generated by positive continuous functions,* BIT, 34 (1994), pp. 367-371.

[10] G. LORENTZ, *Approximation of Functions,* 2nd edition, Chelsea, New York, 1986.

[11] M.J.D. POWELL, *Approximation theory and methods,* Cambridge Univ. Press, 1982.

[12] T. RIVLIN, *Introduction to the Approximation of functions,* Dover Pubs, 1981.

[13] S. SERRA, *New PCG based algorithms for the solution of Hermitian Toeplitz systems,* Calcolo, 32 (1995), pp. 154-176.

[14] S. SERRA, *On the extreme spectral properties of Toeplitz matrices generated by L^1 functions with several minima (maxima),* BIT, 36 (1996), pp. 135-142.

[15] S. SERRA, *Optimal, Quasi-Optimal and Superlinear Band-Toeplitz preconditioners for asymptotically ill-conditioned positive definite Toeplitz Systems,* Math. Comp., 66 (1997), pp. 651-665.

[16] S. SERRA, *Toeplitz preconditioners constructed from linear approximation processes,* SIAM J. Matrix. Anal. Appl., 20 (1998), pp. 446-465

[17] S. SERRA, *A Korovkin-type Theory for finite Toeplitz operators via matrix algebras,* Nuner. Math., 82 (1999), pp. 117-142.

[18] G. STRANG, *A proposal for Toeplitz matrix calculation,* Stud. Appl. Math., 74 (1986), pp. 171-176.

[19] H. WIDOM, *Toeplitz Matrices* , In Studies in real and Complex analysis, I. Hirshman Jr. Ed., Math. Ass. Am., 1965.

[20] J.H. WILKINSON, *The Algebraic Eigenvalue Problem,* Oxford Press, Oxford, 1965.

Oscillation Criteria for Delay Difference Equations

JIANHUA SHEN
Department of Mathematics, Hunan Normal University
Changsha, Hunan 410081, China
E-mail: jhsh@public.cs.hn.cn

I.P. STAVROULAKIS
Department of Mathematics, University of Ioannina
451 10 Ioannina, Greece
E-mail: ipstav@cc.uoi.gr

Abstract. This paper is concerned with the oscillation of all solutions of the delay difference equation

$$x_{n+1} - x_n + p_n x_{n-k} = 0, \quad n = 0, 1, 2, ...$$

where $\{p_n\}$ is a sequence of nonnegative real numbers and k is a positive integer. Some new oscillation conditions are established. These conditions concern the case when none of the well-known oscillation conditions

$$\limsup_{n\to\infty} \sum_{i=0}^{k} p_{n-i} > 1 \quad \textit{and} \quad \liminf_{n\to\infty} \frac{1}{k} \sum_{i=1}^{k} p_{n-i} > \frac{k^k}{(k+1)^{k+1}}$$

is satisfied.

Key words: Oscillation, nonoscillation, delay difference equation.

AMS Subject Classification (1991): 39A 10.

1. INTRODUCTION

In the last few decades the oscillation theory of delay differential equations has been extensively developed. The oscillation theory of discrete analogues of delay differential equations has also attracted growing attention in the recent few years. The reader is referred to [1-5,9,10,15,16,18,20-23]. In particular, the problem of establishing sufficient conditions for the oscillation of all solutions of the delay difference equation

$$x_{n+1} - x_n + p_n x_{n-k} = 0, \quad n = 0, 1, 2, ... \tag{1.1}$$

where $\{p_n\}$ is a sequence of nonnegative real numbers and k is a positive integer, has been the subject of many recent investigations. See, for example, [2-7,9,15,16,18,20,21,23] and the references cited therein. Strong interest in Eq. (1.1) is motivated by the fact that it represents a discrete analogue of the delay differential equation

$$x'(t) + p(t)x(t-\tau) = 0, \quad p(t) \geq 0, \quad \tau > 0. \tag{1.2}$$

By a solution of (1.1) we mean a sequence $\{x_n\}$ which is defined for $n \geq -k$ and which satisfies (1.1) for $n \geq 0$. A solution $\{x_n\}$ of (1.1) is said to be *oscillatory* if the terms x_n of the solution are not eventually positive or eventually negative. Otherwise the solution is called *nonoscillatory*.

In 1989, Erbe and Zhang [9] and Ladas, Philos and Sficas [16] studied the oscillation of Eq. (1.1) and proved that all solutions oscillate if

$$\limsup_{n\to\infty} \sum_{i=0}^{k} p_{n-i} > 1, \tag{1.3}$$

or

$$\liminf_{n\to\infty} p_n > \frac{k^k}{(k+1)^{k+1}}, \tag{1.4}$$

or

$$\liminf_{n\to\infty} \frac{1}{k} \sum_{i=1}^{k} p_{n-i} > \frac{k^k}{(k+1)^{k+1}}. \tag{1.5}$$

Observe that (1.5) improves (1.4).

It is interesting to establish sufficient conditions for the oscillation of all solutions of (1.1) when (1.3) and (1.5) are not satisfied. (For Eq. (1.2), this question has been investigated by many authors, see, for example, [8,11-14,19] and the references cited therein). In 1993, Yu, Zhang and Qian [23] and Lalli and Zhang [18] derived some results in this direction. Unfortunately, the main results in [23,18] are not correct. This is because these results are based on a false discrete version of Koplatadze-Chanturia Lemma (a counterexample is given in [5]).

In 1998 Domshlak [4], studied the oscillation of all solutions and the existence of nonoscillatory solution of Eq. (1.1) with r -periodic positive coefficients $\{p_n\}, p_{n+r} = p_n$. It is very important that in the following cases where $\{r = k\}, \{r = k+1\}, \{r = 2\}, \{k = 1, r = 3\}$ and $\{k = 1, r = 4\}$ the results obtained are stated in terms of necessary and sufficient conditions, and their checking is very easy.

Following this historical (and chronological) review we also mention that in the case where

$$\frac{1}{k} \sum_{i=1}^{K} p_{n-i} \geq \frac{k^k}{(k+1)^{k+1}} \quad \textit{and} \quad \lim_{n\to\infty} \frac{1}{k} \sum_{i=1}^{k} p_{n-i} = \frac{k^k}{(k+1)^{k+1}},$$

the oscillation of (1.1) has been studied in 1994 by Domshlak [3] and in 1998 by Tang [21] (see also Tang and Yu [22]). In a case when p_n is asymptotically close to one of the periodic critical states, unimprovable results about oscillation preperties of the equation

$$x_{n+1} - x_n + p_n x_{n-1} = 0$$

were obtained by Domshlak in 1999 [6] and in 2000 [7].

The aim of this paper is to use some new techniques and improve the methods previously used to obtain new oscillation conditions for (1.1). Our results are based on two new lemmas established in section 2.

For convenience, we will assume that inequalities about values of sequences are satisfied eventually for all large n.

2. SOME NEW LEMMAS

Lemma 2.1. *Let the number $h \geq 0$ be such that*

$$\frac{1}{k}\sum_{i=1}^{k} p_{n-i} \geq h \quad \textit{for large } n. \tag{2.1}$$

Assume that (1.1) has an eventually positive solution $\{x_n\}$. Then $h \leq k^k/(k+1)^{k+1}$ and

$$\limsup_{n\to\infty} \frac{x_n}{x_{n-k}} \leq [d(h)]^k, \tag{2.2}$$

where $d(h)$ is the greater real root of the algebraic equation

$$d^{k+1} - d^k + h = 0, \quad \textit{on the interval } [0,1]. \tag{2.3}$$

Proof. Since (1.5) implies that all solutions of (1.1) oscillate, but (1.1) has an eventually positive solution, from (2.1), it follows that $h \leq k^k/(k+1)^{k+1}$ must hold. We now prove (2.2). To this end, we let

$$w_n = \frac{1}{k}\sum_{i=1}^{k} \frac{x_{n-i}}{x_{n-i-1}}. \tag{2.4}$$

and first prove that $\limsup_{n\to\infty} w_n \leq d(h)$. From (1.1), it follows that $\{x_n\}$ is eventually decreasing and so for large n, we have $x_{n-i-1} \geq x_{n-i}$ for $i = 1, 2, ..., k$. This implies that

$$w_n = \frac{1}{k}\sum_{i=1}^{k} \frac{x_{n-i}}{x_{n-i-1}} \leq 1 := d_1. \tag{2.5}$$

Thus, $\limsup_{n\to\infty} w_n \leq d(h)$ holds for $h = 0$ because of $d(0) = 1$. We now consider the case when $0 < h \leq k^k/(k+1)^{k+1}$. From (1.1), we have

$$x_{n-i-1} = x_{n-i} + p_{n-i-1}x_{n-i-k-1}, \quad i = 1, 2, ..., k. \tag{2.6}$$

Using the Arithmetic-Geometric Mean Inequality in (2.5), we have

$$\left(\frac{x_{n-1}}{x_{n-k-1}}\right)^{1/k} \leq d_1,$$

and so

$$\frac{x_{n-i-k-1}}{x_{n-i-1}} \geq d_1^{-k}, \quad i = 1, 2, ..., k.$$

Dividing both sides of (2.6) by x_{n-i-1} and using the last inequality, we have

$$1 = \frac{x_{n-i}}{x_{n-i-1}} + p_{n-i-1}\frac{x_{n-i-k-1}}{x_{n-i-1}} \geq \frac{x_{n-i}}{x_{n-i-1}} + d_1^{-k}p_{n-i-1}.$$

Summing both sides of the last inequality from $i = 1$ to $i = k$, we obtain

$$\sum_{i=1}^{k} \frac{x_{n-i}}{x_{n-i-1}} \leq k - d_1^{-k}\sum_{i=1}^{k} p_{n-i-1}.$$

This, in view of (2.1), leads to

$$w_n \leq 1 - d_1^{-k}\frac{1}{k}\sum_{i=1}^{k} p_{n-i-1} \leq 1 - \frac{h}{d_1^k} := d_2.$$

Using the last inequality and repeating the above arguments, we have

$$w_n \leq 1 - \frac{h}{d_2^k} := d_3.$$

Following this iterative procedure, by induction, we have

$$w_n \leq 1 - \frac{h}{d_m^k} := d_{m+1}, \quad m = 1, 2, ... \tag{2.7}$$

It is easy to see that $1 = d_1 > d_2 > \cdots > d_m > d_{m+1} > 0, m = 1, 2, ...$ Therefore, the limit $\lim_{m\to\infty} d_m = d$ exists and satisfies (2.3). Since (2.7) holds for all $m = 1, 2, ..., \{d_m\}$ is decreasing and $d(h)$ is the greater real root of the equation (2.3), it follows that $\limsup_{n\to\infty} w_n \leq d(h)$ holds. Finally, using the Arithmetic-Geometric Mean Inequality , we have

$$\limsup_{n\to\infty}\left(\frac{x_{n-1}}{x_{n-k-1}}\right)^{1/k} \leq \limsup_{n\to\infty}\frac{1}{k}\sum_{i=1}^{k}\frac{x_{n-i}}{x_{n-i-1}} \leq d(h).$$

This implies (2.2). The proof is complete.

We describe by the following proposition and remark the number $d(h)$.

Proposition 2.1. *For Eq. (2.3), the following statements hold true:*

(i) if $h = 0$, then (2.3) has exactly two different real roots $d_1 = 0$ and $d_2 = 1$.

(ii) if $0 < h < k^k/(k+1)^{k+1}$, then (2.3) has exactly two different real roots d_1 and d_2 such that

$$d_1 \in (0, k/(k+1)), \quad d_2 \in (k/(k+1), 1).$$

(iii) if $h = k^k/(k+1)^{k+1}$, then (2.3) has a unique real root $d = k/(k+1)$.

The proof of the above Proposition is easy and is omitted.

Remark 2.1. From Proposition 2.1, we see that the number $d(h)$ in Lemma 2.1 satisfies

$$d(h) \text{ is } \begin{cases} = 1, & h = 0 \\ \in (k/(k+1), 1), & 0 < h < k^k/(k+1)^{k+1} \\ = k/(k+1), & h = k^k/(k+1)^{k+1}. \end{cases}$$

Lemma 2.2. *Let the number $M \geq 0$ be such that*

$$\sum_{i=1}^{k} p_{n-i} \geq M \quad \text{for large } n. \tag{2.8}$$

Assume that Eq. (1.1) has an eventually positive solution $\{x_n\}$. Then $M \leq k^{k+1}/(k+1)^{k+1}$ and

$$\limsup_{n\to\infty} \frac{x_{n-k}}{x_n} \prod_{i=1}^{k} \sum_{j=1}^{k} p_{n-i+j} \leq [\overline{d}(M)]^k, \tag{2.9}$$

where $\overline{d}(M)$ is the greater real root of the algebraic equation

$$d^{k+1} - d^k + M^k = 0, \quad \text{on } [0,1]. \tag{2.10}$$

Proof. As in the proof of Lemma 2.1, we have that $M \leq k^{k+1}/(k+1)^{k+1}$ must hold. We now prove (2.9). To this end, we let

$$\overline{w}_n = \frac{1}{k} \sum_{i=1}^{k} \frac{x_{n-i}}{x_{n-i+1}} \left(\sum_{j=1}^{k} p_{n-i+j} \right). \tag{2.11}$$

and first prove that

$$\limsup_{n\to\infty} \overline{w}_n \leq \overline{d}(M). \tag{2.12}$$

From (1.1), we have

$$x_{n+j+1} - x_{n+j} + p_{n+j}x_{n+j-k} = 0, \quad j = 0, 1, ..., k-1.$$

Summing the above equality from $j = 0$ to $j = k-1$, we have

$$x_n = x_{n+k} + \sum_{j=0}^{k-1} p_{n+j}x_{n+j-k}. \tag{2.13}$$

Since $\{x_n\}$ is eventually decreasing, it follows that

$$x_n > \sum_{j=0}^{k-1} p_{n+j}x_{n+j-k} \geq \left(\sum_{j=0}^{k-1} p_{n+j}\right) x_{n-1},$$

and so for $i = 1, 2, ..., k$, we have

$$\frac{x_{n-i}}{x_{n-i+1}} \left(\sum_{j=1}^{k} p_{n-i+j}\right) < 1.$$

Summing the last inequality from $i = 1$ to $i = k$, we obtain

$$\overline{w}_n = \frac{1}{k}\sum_{i=1}^{k} \frac{x_{n-i}}{x_{n-i+1}} \left(\sum_{j=1}^{k} p_{n-i+j}\right) < 1 := d_1. \tag{2.14}$$

Thus (2.12) holds for $M = 0$ because of $\overline{d}(0) = 1$. We now consider the case when $0 < M \leq k^{k+1}/(k+1)^{k+1}$. Using (2.8) and the Arithmetic-Geometric Mean Inequality in (2.14), we have

$$M\left(\frac{x_{n-k}}{x_n}\right)^{1/k} < d_1 \quad or \quad \frac{x_{n-k}}{x_n} < \frac{d_1^k}{M^k}. \tag{2.15}$$

Since $\{x_n\}$ is eventually decreasing, from (2.13), for $i = 1, 2, ..., k$, we have

$$\begin{aligned} x_{n-i+1} &= x_{n+k-i+1} + \sum_{j=0}^{k-1} p_{n-i+j+1}x_{n-i+j-k+1} \\ &\geq x_{n+k-i+1} + \sum_{j=1}^{k} p_{n-i+j}x_{n-i}, \end{aligned}$$

and so

$$1 \geq \frac{x_{n+k-i+1}}{x_{n-i+1}} + \sum_{j=1}^{k} p_{n-i+j}\frac{x_{n-i}}{x_{n-i+1}}. \tag{2.16}$$

The last inequality, in view of (2.15), yields

$$1 > \frac{M^k}{d_1^k} + \sum_{j=1}^{k} p_{n-i+j}\frac{x_{n-i}}{x_{n-i+1}}.$$

Summing the last inequality from $i = 1$ to $i = k$, we obtain

$$k > \frac{kM^k}{d_1^k} + \sum_{i=1}^{k} \frac{x_{n-i}}{x_{n-i+1}} \left(\sum_{j=1}^{k} p_{n-i+j} \right).$$

Thus

$$\overline{w}_n = \frac{1}{k} \sum_{i=1}^{k} \frac{x_{n-i}}{x_{n-i+1}} \left(\sum_{j=1}^{k} p_{n-i+j} \right) < 1 - \frac{M^k}{d_1^k} := d_2. \tag{2.17}$$

Using the inequality (2.17) and repeating the above arguments, we have

$$\overline{w}_n < 1 - \frac{M^k}{d_2^k} := d_3.$$

Following this iterative procedure, by induction, we have

$$\overline{w}_n < 1 - \frac{M^k}{d_m^k} := d_{m+1}, \quad m = 1, 2, ... \tag{2.18}$$

Now (2.12) follows from similar proof as in Lemma 2.1. Next, using the Arithmetic-Geometric Mean Inequality in (2.12) we have

$$\limsup_{n\to\infty} \left(\frac{x_{n-k}}{x_n} \prod_{i=1}^{k} \sum_{j=1}^{k} p_{n-i+j} \right)^{1/k} \leq \limsup_{n\to\infty} \frac{1}{k} \sum_{i=1}^{k} \frac{x_{n-i}}{x_{n-i+1}} \left(\sum_{j=1}^{k} p_{n-i+j} \right) \leq \overline{d}(M),$$

which leads to (2.9). The proof is complete.

Observe that the number M in Lemma 2.2 satisfies

$$0 \leq M^k \leq \left(\frac{k^{k+1}}{(k+1)^{k+1}} \right)^k \leq \frac{k^k}{(k+1)^{k+1}},$$

and the last equality holds if and only if $k = 1$. Thus, from Proposition 2.1, we have the following conclusion about the equation (2.10).

Proposition 2.2. *For Eq. (2.10), the following statements hold true:*

(i) if $M = 0$, then (2.10) has exactly two different real roots $d_1 = 0$ and $d_2 = 1$.

(ii) if $k \neq 1$ and $0 < M \leq k^{k+1}/(k+1)^{k+1}$, then (2.10) has exactly two different real roots d_1 and d_2 which satisfy

$$d_1 \in (0, k/(k+1)), \quad d_2 \in (k/(k+1), 1).$$

(iii) if $k = 1$, then (2.10) has two real roots of the form

$$d_1 = \frac{1 - \sqrt{1 - 4M}}{2} \quad \text{and} \quad d_2 = \frac{1 + \sqrt{1 - 4M}}{2}.$$

Remark 2.2. The number $\overline{d}(M)$ in Lemma 2.2 satisfies

$$\overline{d}(M) \quad is \quad \begin{cases} = 1, & M = 0 \\ \in (k/(k+1), 1), & k \neq 1, 0 < M \leq k^{k+1}/(k+1)^{k+1} \\ = (1 + \sqrt{1 - 4M})/2, & k = 1. \end{cases}$$

This implies that $\overline{d}(M) \leq 1$ and the equality holds if and only if $M = 0$. Observe that (2.8) implies

$$\prod_{i=1}^{k} \sum_{j=1}^{k} p_{n-i+j} \geq M^k.$$

Thus, from (2.9), we have

$$\liminf_{n\to\infty} \frac{x_n}{x_{n-k}} \geq [\overline{d}(M)]^{-k} M^k.$$

3. OSCILLATION CRITERIA FOR EQ. (1.1)

In this section, by using the results in section 2, we establish new oscillation criteria for (1.1). From section 1, we see that all solutions of (1.1) oscillate if (1.3), or (1.4) or (1.5) is satisfied. Therefore, we establish oscillation conditions for (1.1) in the case when none of these conditions is satisfied. Let

$$\mu = \liminf_{n\to\infty} \frac{1}{k} \sum_{i=1}^{k} p_{n-i}. \tag{3.1}$$

Theorem 3.1. *Assume that* $0 \leq \mu \leq k^k/(k+1)^{k+1}$ *and that there exists an integer* $l \geq 1$ *such that*

$$\limsup_{n\to\infty} \left\{ \sum_{i=1}^{k} p_{n-i} + [\overline{d}(k\mu)]^{-k} \prod_{i=1}^{k} \sum_{j=1}^{k} p_{n-i+j} \right.$$
$$\left. + \sum_{m=0}^{l-1} [d(\mu)]^{-(m+1)k} \sum_{i=1}^{k} \prod_{j=0}^{m+1} p_{n-jk-i} \right\} > 1, \tag{3.2}$$

where $\overline{d}(k\mu)$ *and* $d(\mu)$ *are the greater real roots of the equations*

$$d^{k+1} - d^k + (k\mu)^k = 0 \tag{3.3}$$

and

$$d^{k+1} - d^k + \mu = 0, \tag{3.4}$$

respectively. Then all solutions of (1.1) oscillate.

Proof. Assume, for the sake of contradiction, that (1.1) has an eventually positive solution $\{x_n\}$. We consider the two possible cases:

CASE 1. $\mu = 0$. In this case we have $\overline{d}(k\mu) = d(\mu) = 1$. From (1.1), we have

$$x_{n-i} = x_{n-i+1} + p_{n-i}x_{n-k-i}, \quad i = 1, 2, ..., k.$$

Summing both sides of the above equality from $i = 1$ to $i = k$ leads to

$$x_{n-k} = x_n + \sum_{i=1}^{k} p_{n-i}x_{n-k-i}. \tag{3.5}$$

From (1.1), for any positive integer j, we have

$$x_{n-k-j} = x_{n-k-j+1} + p_{n-k-j}x_{n-k-j-k}. \tag{3.6}$$

Substituting (3.6) for $j = i$ into (3.5), we have

$$x_{n-k} = x_n + \sum_{i=1}^{k} p_{n-i}x_{n-k-i+1} + \sum_{i=1}^{k} p_{n-i}p_{n-k-i}x_{n-i-2k}.$$

Substituting (3.6) for $j = i + k$ into the last equality, we have

$$\begin{aligned} x_{n-k} &= x_n + \sum_{i=1}^{k} p_{n-i}x_{n-k-i+1} + \sum_{i=1}^{k} p_{n-i}p_{n-k-i}x_{n-2k-i+1} \\ &\quad + \sum_{i=1}^{k} p_{n-i}p_{n-k-i}p_{n-2k-i}x_{n-i-3k}. \end{aligned}$$

By induction, it is easy to prove that

$$\begin{aligned} x_{n-k} &= x_n + \sum_{i=1}^{k} p_{n-i}x_{n-k-i+1} + \sum_{i=1}^{k} p_{n-i}p_{n-k-i}x_{n-2k-i+1} \\ &\quad + \sum_{i=1}^{k} p_{n-i}p_{n-k-i}p_{n-2k-i}x_{n-3k-i+1} + \cdots + \\ &\quad + \sum_{i=1}^{k} p_{n-i}p_{n-k-i}\cdots p_{n-lk-i}x_{n-(l+1)k-i+1} \\ &\quad + \sum_{i=1}^{k} p_{n-i}p_{n-k-i}\cdots p_{n-(l+1)k-i}x_{n-i-(l+2)k}. \end{aligned}$$

Removing the last term of the last equality, we have

$$x_{n-k} \geq x_n + \sum_{i=1}^{k} p_{n-i}x_{n-k-i+1} + \sum_{m=0}^{l-1}\sum_{i=1}^{k} x_{n-(m+2)k-i+1}\prod_{j=0}^{m+1} p_{n-jk-i}. \tag{3.7}$$

In the proof of Lemma 2.2, we have (2.14) holds. Using the Arithmetric-Geometric Mean Inequality in (2.14), we have

$$\left(\frac{x_{n-k}}{x_n}\prod_{i=1}^{k}\sum_{j=1}^{k}p_{n-i+j}\right)^{1/k}<1,$$

and so

$$x_n>\left(\prod_{i=1}^{k}\sum_{j=1}^{k}p_{n-i+j}\right)x_{n-k}. \tag{3.8}$$

Substituting (3.8) into (3.7) and using the fact that $\{x_n\}$ is eventually decreasing, we have

$$x_{n-k}>\left(\sum_{i=1}^{k}p_{n-i}+\prod_{i=1}^{k}\sum_{j=1}^{k}p_{n-i+j}+\sum_{m=0}^{l-1}\sum_{i=1}^{k}\prod_{j=0}^{m+1}p_{n-jk-i}\right)x_{n-k}.$$

Dividing both sides of the last inequality by x_{n-k}, and taking the limit superior as $n\to\infty$, we have

$$1\geq\limsup_{n\to\infty}\left\{\sum_{i=1}^{k}p_{n-i}+\prod_{i=1}^{k}\sum_{j=1}^{k}p_{n-i+j}+\sum_{m=0}^{l-1}\sum_{i=1}^{k}\prod_{j=0}^{m+1}p_{n-jk-i}\right\}.$$

This contradicts (3.2).

CASE 2. $0<\mu\leq k^k/(k+1)^{k+1}$. In this case, for any $\eta\in(0,\mu)$, we have

$$\frac{1}{k}\sum_{i=1}^{k}p_{n-i}\geq\mu-\eta. \tag{3.9}$$

From (3.7), we have

$$x_{n-k}\geq x_n+\sum_{i=1}^{k}p_{n-i}x_{n-k}+\sum_{m=0}^{l-1}x_{n-(m+2)k}\sum_{i=1}^{k}\prod_{j=0}^{m+1}p_{n-jk-i}. \tag{3.10}$$

By Lemma 2.2, we have

$$x_n\geq\{[\overline{d}(k(\mu-\eta))]^{-k}-\eta\}\prod_{i=1}^{k}\sum_{j=1}^{k}p_{n-i+j}x_{n-k}, \tag{3.11}$$

where $\overline{d}(k(\mu-\eta))$ is the greater real root of the equation

$$d^{k+1}-d^k+k^k(\mu-\eta)^k=0. \tag{3.12}$$

By Lemma 2.1, we have

$$x_{n-(m+2)k}\geq\{[d(\mu-\eta)]^{-(m+1)k}-\eta\}x_{n-k}, \tag{3.13}$$

where $d(\mu-\eta)$ is the greater real root of the equation

$$d^{k+1}-d^k+(\mu-\eta)=0. \tag{3.14}$$

Now substituting (3.11) and (3.13) into (3.10), we obtain

$$\begin{aligned} x_{n-k} \geq & \sum_{i=1}^{k} p_{n-i}x_{n-k}+\{[\overline{d}(k(\mu-\eta))]^{-k}-\eta\}\prod_{i=1}^{k}\sum_{j=1}^{k}p_{n-i+j}x_{n-k} \\ & +\sum_{m=0}^{l-1}\{[d(\mu-\eta)]^{-(m+1)k}-\eta\}\sum_{i=1}^{k}\prod_{j=0}^{m+1}p_{n-jk-i}x_{n-k}. \end{aligned}$$

Dividing both sides of the last inequality by x_{n-k} then taking the limit superior as $n\to\infty$, we have

$$\begin{aligned} 1\geq \limsup_{n\to\infty}\Bigg\{ & \sum_{i=1}^{k}p_{n-i}+\{[\overline{d}(k(\mu-\eta))]^{-k}-\eta\}\prod_{i=1}^{k}\sum_{j=1}^{k}p_{n-i+j} \\ & +\sum_{m=0}^{l-1}\{[d(\mu-\eta)]^{-(m+1)k}-\eta\}\sum_{i=1}^{k}\prod_{j=0}^{m+1}p_{n-jk-i}\Bigg\}. \end{aligned}$$

Letting $\eta\to 0$, we have $\overline{d}(k(\mu-\eta))\to\overline{d}(k\mu)$ and $d(\mu-\eta)\to d(\mu)$, so that the last inequality contradicts (3.2). The proof is now complete.

Notice that when $k=1$, from Remark 2.1 and Remark 2.2, we have $d(\mu)=\overline{d}(\mu)=(1+\sqrt{1-4\mu})/2$, so condition (3.2) reduces to

$$\limsup_{n\to\infty}\left\{Cp_n+p_{n-1}+\sum_{m=0}^{l-1}C^{m+1}\prod_{j=0}^{m+1}p_{n-j-1}\right\}>1, \tag{3.15}$$

where $C=2/(1+\sqrt{1-4\mu})$, $\mu=\liminf_{n\to\infty}p_n$. Therefore, from Theorem 3.1, we have the following corollary.

Corollary 3.1. *Assume that $0\leq\mu\leq 1/4$ and that (3.15) holds. Then all solutions of the equation*

$$x_{n+1}-x_n+p_nx_{n-1}=0 \tag{3.16}$$

oscillate.

A condition obtained from (3.15) and whose checking is more easy is given in next corollary.

Corollary 3.2. *Assume that $0\leq\mu\leq 1/4$ and that*

$$\limsup_{n\to\infty}p_n>\left(\frac{1+\sqrt{1-4\mu}}{2}\right)^2. \tag{3.17}$$

Then all solutions of (3.16) oscillate.

Proof. When $\mu = 0$, by condition (1.3), all solutions of (3.17) oscillate. For the case when $0 < \mu \leq 1/4$, by Theorem 3.1, it suffices to prove that (3.17) implies (3.15). Notice

$$\frac{1+\sqrt{1-4\mu}}{2} = 1 - \frac{\mu}{1-C\mu},$$

by (3.17) and $\mu = \liminf_{n\to\infty} p_n$, there exists $\varepsilon \in (0, \mu)$ such that $p_n \geq \mu - \varepsilon$ and

$$C \limsup_{n\to\infty} p_n > 1 - \frac{\mu-\varepsilon}{1-C(\mu-\varepsilon)}.$$

The last inequality, in view of the fact that $[C(\mu-\varepsilon)]^m \to 0$ as $m \to \infty$, implies that for some sufficiently large integer $l > 1$

$$\begin{aligned} C \limsup_{n\to\infty} p_n &> 1 - \frac{(\mu-\varepsilon)\{1-[C(\mu-\varepsilon)]^{l+1}\}}{1-C(\mu-\varepsilon)} \\ &= 1 - (\mu-\varepsilon) - C(\mu-\varepsilon)^2 - \cdots - C^l(\mu-\varepsilon)^{l+1}, \end{aligned}$$

which leads to (3.15), because

$$p_{n-1} + \sum_{m=0}^{l-1} C^{m+1} \prod_{j=0}^{m+1} p_{n-j-1} \geq (\mu-\varepsilon) + C(\mu-\varepsilon)^2 + \cdots + C^l(\mu-\varepsilon)^{l+1}.$$

The proof is complete.

Observe that when $\mu = 1/4$, condition (3.17) reduces to $\limsup_{n\to\infty} p_n > 1/4$, which can not be improved in the sense that the lower bound $1/4$ can not be replaced by a smaller number. Indeed, by Theorem 2.3 in [9], we see that Eq. (3.16) has a nonoscillatory solution if $p_n \leq 1/4$ for large n. Note, however, that even in the critical state $\lim_{n\to\infty} p_n = 1/4$ Eq. (3.16) can be either oscillatory or nonoscillatory. For example, if $p_n = \frac{1}{4} + \frac{c}{n^2}$ then Eq. (3.16) will be oscillatory in case $c > 1/4$ and nonoscillatory in case $c < 1/4$ (the Kneser-like theorem, [3]).

Example. Consider the equation

$$x_{n+1} - x_n + \left(\frac{1}{4} + a\sin^4\frac{n\pi}{8}\right) x_{n-1} = 0,$$

where $a > 0$ is a constant. It is easy to see that

$$\liminf_{n\to\infty} p_n = \liminf_{n\to\infty} \left(\frac{1}{4} + a\sin^4\frac{n\pi}{8}\right) = \frac{1}{4},$$

$$\limsup_{n\to\infty} p_n = \limsup_{n\to\infty}\left(\frac{1}{4} + a\sin^4\frac{n\pi}{8}\right) = \frac{1}{4} + a.$$

Therefore, by Corollary 3.2, all solutions of the equation oscillate. However, none of the conditions (1.3)-(1.5) and those appear in [4,20,23] is satisfied.

The following corollary concerns the case when $k > 1$.

Corollary 3.3. *Assume that* $0 \le \mu \le k^k/(k+1)^{k+1}$ *and that*

$$\limsup_{n\to\infty}\sum_{i=1}^{k} p_{n-i} > 1 - [\overline{d}(k\mu)]^{-k}(k\mu)^k - \frac{k[d(\mu)]^{-k}\mu_*^2}{1-[d(\mu)]^{-k}\mu_*}, \tag{3.18}$$

where $\mu_* = \liminf_{n\to\infty} p_n$, *and* $\overline{d}(k\mu), d(\mu)$ *are as in Theorem 3.1. Then all solutions of (1.1) oscillate.*

Proof. If $\mu = 0$ (then $\mu_* = 0$), then, by (1.3), all solutions of (1.1) oscillate. If $\mu_* = 0, \mu > 0$, then (3.18) reduces to

$$\limsup_{n\to\infty}\sum_{i=1}^{k} p_{n-i} > 1 - [\overline{d}(k\mu)]^{-k}(k\mu)^k. \tag{3.19}$$

From (3.1) and (3.19), for some sufficiently small $\eta \in (0,\mu)$ we have

$$\frac{1}{k}\sum_{i=1}^{k} p_{n-i} \ge \mu - \eta, \quad \limsup_{n\to\infty}\sum_{i=1}^{k} p_{n-i} > 1 - [\overline{d}(k\mu)]^{-k}(k(\mu-\eta))^k. \tag{3.20}$$

Thus, we obtain

$$[\overline{d}(k\mu)]^{-k}\prod_{i=1}^{k}\sum_{j=1}^{k} p_{n-i+j} \ge [\overline{d}(k\mu)]^{-k}(k(\mu-\eta))^k.$$

From this and the second inequality of (3.20), we see that (3.2) holds. By Theorem 3.1, all solutions of (1.1) oscillate. We now consider the case when $0 < \mu_* \le k^k/(k+1)^{k+1}$. By Theorem 3.1, it suffices to prove that condition (3.18) implies condition (3.2). From (3.18), it follows that, for some sufficiently small $\eta \in (0,\mu_*)$ we have

$$\limsup_{n\to\infty}\sum_{i=1}^{k} p_{n-i} > 1 - [\overline{d}(k\mu)]^{-k}(k(\mu-\eta))^k - \frac{k[d(\mu)]^{-k}(\mu_*-\eta)^2}{1-[d(\mu)]^{-k}(\mu_*-\eta)}.$$

This, in view of the fact that $[[d(\mu)]^{-k}(\mu_*-\eta)]^m \to 0$ as $m \to \infty$, implies that for some sufficiently large integer $l > 1$

$$\limsup_{n\to\infty}\sum_{i=1}^{k} p_{n-i} \;>\; 1 - [\overline{d}(k\mu)]^{-k}(k(\mu-\eta))^k$$

$$
\begin{aligned}
& - \frac{k(\mu_* - \eta)^2 [d(\mu)]^{-k} \{1 - [[d(\mu)]^{-k}(\mu_* - \eta)]^l\}}{1 - [d(\mu)]^{-k}(\mu_* - \eta)} \\
= \; & 1 - [\overline{d}(k\mu)]^{-k}(k(\mu - \eta))^k - k(\mu_* - \eta)^2 [d(\mu)]^{-k} \\
& \times \{1 + [d(\mu)]^{-k}(\mu_* - \eta) + [d(\mu)]^{-2k}(\mu_* - \eta)^2 \\
& + \cdots + [d(\mu)]^{-(l-1)k}(\mu_* - \eta)^{l-1}\}.
\end{aligned}
$$

This leads to (3.2) because

$$
\begin{aligned}
& [\overline{d}(k\mu)]^{-k} \prod_{i=1}^{k} \sum_{j=1}^{k} p_{n-i+j} + \sum_{m=0}^{l-1} [d(\mu)]^{-(m+1)k} \sum_{i=1}^{k} \prod_{j=0}^{m+1} p_{n-jk-i} \\
\geq \; & [\overline{d}(k\mu)]^{-k}(k(\mu - \eta))^k + k(\mu_* - \eta)^2 [d(\mu)]^{-k} + k(\mu_* - \eta)^3 [d(\mu)]^{-2k} \\
& + \cdots + k(\mu_* - \eta)^{l+1} [d(\mu)]^{-lk}.
\end{aligned}
$$

The proof is complete.

REFERENCES

1. R.P. Agarwal and P.J.Y. Wong, Advanced Topics in Difference Equations, Kluwer Academic Publishers, 1997.
2. Sui Sun Cheng and Guang Zhang, "Virus" in several discrete oscillation theorems, *Applied Math. Letters*, **13** (2000), 9-13.
3. Y. Domshlak, Sturmian comparison method in oscillation study for discrete difference equations, I, *Differential and Integral Equations*, **7** (1994), 571-582.
4. Y. Domshlak, Delay-difference equations with periodic coefficients: sharp results in oscillation theory, *Math. Inequal. Appl.*, **1** (1998), 403-422.
5. Y. Domshlak, What should be a discrete version of the Chanturia-Koplatadze Lemma? *Functional Differential Equations* (to appear).
6. Y. Domshlak, Riccati Difference Equations with almost periodic coefficients in the critical state, *Dynamic Systems Appl.*, **8** (1999), 389-399.
7. Y. Domshlak, The Riccati Difference Equations near "extremal" critical states, *J. Difference Equations Appl.*, **6** (2000), 387-416.
8. Á. Elbert and I.P. Stavroulakis, Oscillations of first order differential equations with deviating arguments, Univ. of Ioannina, T. R. N^0 172 1990, *Recent trends in differential equations* 163-178, *World Sci. Ser. Appl. Anal.*, World Sci. Publishing Co. (1992).
9. L. Erbe and B.G. Zhang, Oscillation of discrete analogues of delay equations, *Differential and Integral Equations*, **2** (1989), 300-309.

10. I. Györi and G. Ladas, Oscillation Theory of Delay Differential Equations with Applications, Clarendon Press, Oxford, 1991.
11. J. Jaroš and I.P. Stavroulakis, Oscillation tests for delay equations, *Rocky Mountain J. Math.*, **29** (1999), 197-207.
12. M. Kon, Y.G. Sficas and I.P. Stavroulakis, Oscillation criteria for delay equations, *Proc. Amer. Math. Soc.*, **128** (2000), 2989-2997.
13. R. Koplatadze and T. Chanturia, On oscillatory and monotonic solutions of first order delay differential equations with deviating arguments, *Differential'nye Uravnenija*, **18** (1982), 1463-1465 (Russian).
14. M.K. Kwong, Oscillation of first-order delay equations, *J. Math. Anal. Appl.*, **156** (1991), 274-286.
15. G. Ladas, Recent developments in the oscillation of delay difference equations, In *International Conference on Differential Equations, Stability and Control*, Dekker, New York, 1990.
16. G. Ladas, C. Philos and Y. Sficas, Sharp conditions for the oscillation of delay difference equations, *J. Appl. Math. Simulation*, **2** (1989), 101-112.
17. G.S. Ladde, V. Lakshmikantham and B.G. Zhang, Oscillation Theory of Differential Equations with Deviating Arguments, Marcel Dekker, New York, 1987.
18. B. Lalli and B.G. Zhang, Oscillation of difference equations, *Colloquium Math.*, **65** (1993), 25-32.
19. Ch.G. Philos and Y.G. Sficas, An oscillation criterion for first order linear delay differential equations, *Canad. Math. Bull.*, **41** (1998), 207-213.
20. I.P. Stavroulakis, Oscillation of delay difference equations, *Computers Math. Applic.*, **29** (1995), 83-88.
21. X.H. Tang, Oscillations of delay difference equations with variable coefficients, (Chinese), *J. Central South Univ. of Technology*, **29** (1998), 287-288.
22. X.H. Tang and J.S. Yu, Oscillation of delay difference equations, *Computers Math. Applic.*, **37** (1999), 11-20.
23. J.S. Yu, B.G. Zhang and X.Z. Qian, Oscillations of delay difference equations with oscillating coefficients, *J. Math. Anal. Appl.*, **177** (1993), 432-444.

VARIATIONAL FORMULATION AND CONSERVATION LAWS OF THERMOELASTICITY WITHOUT DISSIPATION

Vassilios K. Kalpakides
Department of Mathematics, University of Ioannina, Ioannina, GR-45110, E-mail: vkalpak@cc.uoi.gr

January 26, 2001

Abstract

The concern of this work is the derivation of conservation laws for the Green–Naghdi theory of non–linear thermoelasticity without dissipation. The lack of dissipation allows for a variational formulation which is used for the application of Noether's theorem. Also, the balance laws in material manifold and the exact conditions under which they hold are rigorously studied.

Keywords: Conservation laws, material momentum, material forces, configurational mechanics

1 Introduction

This work is devoted to the Lagrangian formulation and the derivation of conservation laws of non-linear thermoelasticity provided by Green and Naghdi hereafter called G-N theory. Green and Naghdi [1] formulated an alternative formulation of what is called hyperbolic thermoelasticity in which disturbances propagate with finite wave speeds. The main feature of this theory is that it does not admit energy dissipation. This by turn allows us to ask if it is possible to consider the G-N theory as a field theory as in the case of hyperelasticity, thus to look for a variational formulation of Hamilton type. Though short time has elapsed since the G-N theory appeared, many researchers have already presented various results concerning it. Especially, we mention the work of Dascalu and Maugin [2] who provided the material momentum equation for G–N theory. For a detailed discussion we refer to the review papers of Chandrasekharaiah [3] and Hetnarski and Ignaczak [4].

Our main concern lies in conservation laws related to this theory, especially in the framework of the so called material or configurational mechanics as refered by Maugin [5]. By this term is meant an approach to continuum mechanics focused on the material configuration providing new insights and a unified view of disparate topics like fracture, phase transitions, dislocations etc are obtained. The interested reader can find a nice and extended exposition of this view in the aforementioned book of Maugin and in the recently published books of Gurtin [6] and of Kienzler and Herrmann [7]. To obtain conservation laws we use the celebrated theorem of Noether exploiting by this manner the obtained variational formulation. It is well known and widely used by many researchers in continuum mechanics that according to Noether's theorem it is possible to obtain a conservation low for every given variational symmetry. But it seems that what Lovelock and Rund [8] call "invariance identity", that is a necessary and sufficient condition for the action integral to be invariant under a given infinitesimal group of transformations is not often used. We used this condition rather systematically not only to explore necessary conditions for invariance of the lagrangian but rather to obtain the non-homogeneous terms of the material balance laws, i.e., the so called material forces or some kind of moment of such forces. We use throughout the paper alternatively the vectorial as well as the index notation to represent Cartesian vectors and tensors, thus rectangular coordinate systems are adopted in all cases. The motion of a thermoelastic body is described by the smooth mapping

$$x_\alpha = x_\alpha(X_A),$$

where $A, \alpha = 1, 2, 3, 4, \quad X_4 = t, \quad x_4 = \Theta, \quad \Theta = \Theta(\mathbf{X}, t) = \Theta(X_A)$ is the temperature scalar field. Also, we use the notation $\mathbf{X}$ to denote the material space variable, and $\mathbf{x}$ for the spatial position of the particle $\mathbf{X}$ at time t. In a coordinate system these variables will be written as $X_L, L = 1, 2, 3$ and $x_i, i = 1, 2, 3$, respectively. Thus, the motion is alternatively written

$$x_i = x_i(X_L, t), \quad \Theta = \Theta(X_L, t)$$

Generally, if it is not otherwise denoted, Greek indices will range from 1 to 4, while the lower-case Latin ones will range from 1 to 3. Also, the capital letters K, L, M,...will range from 1 to 3 and A, B,... from 1 to 4. We use two distinct differential operators $\frac{\partial}{\partial X_A}$ and $\frac{D}{DX_A}$. The former is the usual partial derivative operator while the latter denotes the partial derivative which accounts for the underlying function composition. For instance,

$$\frac{D}{DX_A} F(X_B, x_\gamma(X_B)) = \frac{\partial F}{\partial X_A} + \frac{\partial F}{\partial x_\gamma} \frac{\partial x_\gamma}{\partial X_A}.$$

Also, the usual notations $GradF = \nabla_R F = \frac{DF}{DX_L}$, $Div\mathbf{F} = \frac{DF_L}{DX_L}$ and $\dot{F} = \frac{DF}{Dt}$ for gradient, divergence and material time derivative, respectively are used.

2 Preliminaries on Green–Naghdi Theory of Thermoelasticity and Noether's Theorem

According to G–N theory the field equations of thermoelasticity of type II [1] i.e., the momentum and energy equations are given respectively as follows

$$\frac{\partial \mathbf{p}}{\partial t} - Div\mathbf{T} = \mathbf{0}, \tag{1}$$

$$-(\frac{D\Psi}{Dt} + \frac{D\Theta}{Dt}\eta) + tr(\mathbf{T}\dot{\mathbf{F}}) - \mathbf{S}\cdot\nabla_R\Theta = 0 \tag{2}$$

where $\mathbf{p} = \rho_R\mathbf{v}$ is the physical momentum, $\mathbf{v} = \frac{\partial \mathbf{x}}{\partial t}$ is the velocity field, Ψ is the free energy function per unit volume, $\mathbf{T}$ is the first Piola–Kirchhoff stress tensor, $\mathbf{F}$ is the deformation gradient tensor, $\mathbf{S}$ is the entropy flux vector and η is the entropy density per unit volume.
Also, the constitutive equations are given in the form

$$\mathbf{T} = \frac{\partial \Psi}{\partial \mathbf{F}}, \quad \mathbf{S} = -\frac{\partial \Psi}{\partial \beta}, \quad \eta = -\frac{\partial \Psi}{\partial \dot{\alpha}}. \tag{3}$$

In a rectangular coordinate system the first two of them can be written

$$T_{Li} = \frac{\partial \Psi}{\partial x_{i,L}}, \quad S_L = -\frac{\partial \Psi}{\partial \beta_L}, \tag{4}$$

where $\beta = \nabla_R\alpha$, $\alpha = \alpha(\mathbf{X}, t)$ is the so called by Green and Naghdi *thermal displacement field*, a scalar field earlier introduced by other reasearchers as well. According to G–N theory the thermal displacement is a primitive concept and the temperature field is defined to be the time derivative of α, thus

$$\Theta(\mathbf{X}, t) := \frac{\partial}{\partial t}\alpha(\mathbf{X}, t)$$

In the sequence, for the needs of the present work we give some fundamental elements related to variational symmetries and Noether's theorem. Let a C^2 function

$$L = L(X_A, x_\alpha, x_{\alpha,A}), \quad A = 1, 2, ...n, \alpha = 1, 2, ...m,$$

where $X_A \in G$, G is a smooth domain of R^n and $x_\alpha(X_A)$ is a sufficiently smooth function. Consider the functional $I : C^2(G) \to R$ given as follows

$$I(x_\alpha) = \int_G L(X_A, x_\alpha, x_{\alpha,A})dV. \tag{5}$$

Hereafter, we shall refer to functional (5) as *action integralal.* The necessary condition for the functional I to attain an extremum is given by the well known Euler–Lagrange equations

$$\frac{\partial L}{\partial x_\gamma} - \frac{D}{DX_A}(\frac{\partial L}{\partial x_{\gamma,A}}) = 0, \quad \forall X_A \in G \tag{6}$$

where the summation convention is used when repeated indices are appeared.

Consider now the $n+m$–dimensional Euclidean space E^{m+n} made up by the depedent and indepedent variables and the continuous group (actually it is a Lie group) of point transformations in this space

$$\begin{aligned} \tilde{X}_A &= \tilde{X}_A(X_B, x_\beta; \epsilon_w), \\ \tilde{x}_\alpha &= \tilde{x}_\alpha(X_B, x_\beta; \epsilon_w), \quad w = 1, 2, ..., \mu \end{aligned} \tag{7}$$

with

$$\tilde{X}_A(X_B, x_\beta; \mathbf{0}) = X_A, \quad \tilde{x}_\alpha(X_B, x_\beta; \mathbf{0}) = x_\alpha,$$

where $\tilde{X}_A$ and $\tilde{x}_\alpha$ are C^∞ with respect to X_B, x_β and analytic with respect to e_w in the domain of their definition. ϵ_w denotes the μ–dimensional parameter of the group.

The corresponding group of infintesimal transformations will be given by the relations

$$\tilde{X}_A = \tilde{X}_A + \epsilon_w Z_A^w, \tag{8}$$

$$\tilde{x}_\alpha = \tilde{x}_\alpha + \epsilon_w \zeta_\alpha^w, \tag{9}$$

where

$$Z_A^w(X_B, x_\beta) = \frac{\partial \tilde{X}_A}{\partial \epsilon_w}(\epsilon_w = \mathbf{0}), \quad \zeta_\alpha^w(X_B, x_\beta) = \frac{\partial \tilde{x}_\alpha}{\partial \epsilon_w}(\epsilon_w = \mathbf{0}). \tag{10}$$

The vector field over the space E^{m+n} defined by the relation

$$\mathbf{V}_w = Z_A^w \frac{\partial}{\partial X_A} + \zeta_\alpha^w \frac{\partial}{\partial x_\alpha} \tag{11}$$

will be called the *infinitesimal generator* of the group (7).

We will say that the vector field (11) or equivalently the group (7) is a variational symmetry [9, 10] of the action integral (5) if the latter is invariant under any member of the the transformation group (7), that is

$$\int_G L(X_A, x_\alpha, x_{\alpha,A}) dV = \int_{\tilde{G}} L(\tilde{X}_A, \tilde{x}_\alpha, \tilde{x_{\alpha,A}}) d\tilde{V}, \tag{12}$$

where a tilde over a quantity denotes the transformation of this quantity under eqs. (7).

Next theorem [10] will provide the so–called infinitesimal critirion in order a functional to be invariant under a continuous group of point transformations.

Theorem *The group of transformations* (7) *is a variational symmetry of the functional* (5) *if an only if*

$$\mathbf{V}_w^{(1)} L + L \frac{D Z_A^w}{D X_A} = 0, \quad w = 1, 2, ..., \mu, \tag{13}$$

where $\mathbf{V}_w^{(1)}$ *denotes the first prolongation* [10] *of the infinitesimal generator* (11).

After that we can easily prove that eq. (13) is equivalent to equation:

$$\frac{\partial L}{\partial X_A} Z_A^w + \frac{\partial L}{\partial x_\alpha} \zeta_\alpha^w + \frac{\partial L}{\partial x_{\alpha,A}} \left(\frac{D\zeta_\alpha^w}{DX_A} - x_{\alpha,B} \frac{DZ_B^w}{DX_A} \right) + L \frac{DZ_A^w}{DX_A} = 0, \tag{14}$$

refered by [8] as *invariance identity.*
We proceed now to the Noether's theorem of which we give a version [8] convenient to our objective

Theorem of Noether *If the functional* (5) *is invariant under the* μ*–parameter group of transformations given by eqs.* (9–10), *there exist* μ *conservation laws of Euler–Lagrange equations* (6) *given by*

$$-\frac{D\theta_A^w}{DX_A} = \left[\frac{\partial L}{\partial x_\gamma} - \frac{D}{DX_A} \left(\frac{\partial L}{\partial x_{\gamma,A}} \right) \right] (\zeta_\gamma^w - x_{\gamma,B} Z_B^w) = 0, \tag{15}$$

where

$$\theta_A^w = -(L Z_A^w - \frac{\partial L}{\partial x_{\alpha,A}} x_{\alpha,B} Z_B^w + \frac{\partial L}{\partial x_{\alpha,A}} \zeta_\alpha^w). \tag{16}$$

3 The Variational Principle

Definition 3.1 *The Lagrangian function of a thermoelastic body without dissipation is defined to be of the form*

$$L(X_L, \dot{x}_i, \dot{\alpha}, \frac{\partial x_i}{\partial X_L}, \frac{\partial \alpha}{\partial X_L}) = \frac{1}{2} \rho_R(X_L) \dot{x}_i \dot{x}_i - \Psi(X_L, \frac{\partial x_i}{\partial X_L}, \dot{\alpha}, \frac{\partial \alpha}{\partial X_L}). \tag{17}$$

The above definition indirectly provides the independent constitutive variables which should be

$$\frac{\partial x_i}{\partial X_L}, \quad \Theta = \dot{\alpha}, \quad \beta = \frac{\partial \alpha}{\partial X_L}$$

in complete accordance with the corresponding ones that Green and Naghdi introduced in what they call thermoelasticity theory of type II [1]. After that, the functional I for the case under discussion will take the form

$$I(x_i, \alpha) = \int_{t_1}^{t_2} \int_\Omega L(X_L, \dot{x}_i, \dot{\alpha}, x_{i,L}, \alpha_{,L}) dV dt, \tag{18}$$

where Ω is a smooth domain of R^3 and $[t_1, t_2]$ an interval of R. Notice that L is not an explicit function of x_i by virtue of Galilean invariance (translations in physical space of placements). Neither is it an explicit function of α itself , this implying a s sort of *gauge invariance* very similar to that of electrostatic for the electric potential. To

proceed to the variational principle we have to add initial and boundary conditions. Let us suppose that the functions $x_\alpha = (x_i, \alpha)$ satisfy the following restrictions

$$\begin{aligned} &x_\alpha(X_L, t) = g_\alpha(X_L, t), \quad X_L \in \partial\Omega, \quad t \in [t_1, t_2] \\ &x_\alpha(X_L, t_1) = h_\alpha(X_L), \quad X_L \in \Omega, \\ &x_\alpha(X_L, t_2) = m_\alpha(X_L), \quad X_L \in \Omega, \end{aligned} \tag{19}$$

where $g_\alpha, h_\alpha, m_\alpha$ are C^2 functions on the domain of their definition and moreover fulfil the compatibiliy relations

$$g_\alpha(X_L, t_1) = h_\alpha(X_L), \quad g_\alpha(X_L, t_2) = m_\alpha(X_L).$$

The variational statement for the G–N theory can be writen as follows:

Proposition 3.1 *Let the constitutive relations* (3) *hold. Among all admissible functions of motion and thermal displacement for a thermoelastic body without dissipation which satisfy the initial–boundary conditions* (19) *those ones affording an extreme value to the action integral defined by* (17–18), *will satisfy the field equations* (1–2).

PROOF We can write in a more elegant and compact form the argument of the Lagrangian as

$$L = L\left(X_A, \frac{\partial x_\alpha}{\partial X_A}\right), \tag{20}$$

where now $x_4 := \alpha$.
Hence, the corresponding Euler–Lagrange equations, i.e. eqs (6), will take the form

$$\frac{D}{DX_A}\left(\frac{\partial L}{\partial x_{\alpha,A}}\right) = 0, \quad \forall X_A \in \Omega \times [t_1, t_2]. \tag{21}$$

What remains is to analyse equations (21). For our problem, they can be written as

$$\begin{aligned} &\frac{D}{DX_A}\left(\frac{\partial L}{\partial x_{\alpha,A}}\right) = \frac{D}{DX_L}\left(\frac{\partial L}{\partial x_{\alpha,L}}\right) + \frac{D}{Dt}\left(\frac{\partial L}{\partial \dot{x}_\alpha}\right) = \\ &\left(\frac{D}{DX_L}\left(\frac{\partial L}{\partial x_{i,L}}\right) + \frac{D}{Dt}\left(\frac{\partial L}{\partial \dot{x}_i}\right), \quad \frac{D}{DX_L}\left(\frac{\partial L}{\partial x_{4,L}}\right) + \frac{D}{Dt}\left(\frac{\partial L}{\partial \dot{x}_4}\right)\right) = \mathbf{0}. \end{aligned} \tag{22}$$

Taking into account the form of Lagrangian (17), equations (22) can be written as

$$\begin{aligned} &\frac{D}{DX_L}\left(\frac{\partial \Psi}{\partial x_{i,L}}\right) - \frac{D}{Dt}\left(\rho_R \dot{x}_i\right) = 0, \quad X_L \in \Omega, \quad t \in [t_1, t_2], \\ &\frac{D}{DX_L}\left(\frac{\partial \Psi}{\partial \alpha_L}\right) + \frac{D}{Dt}\left(\frac{\partial \Psi}{\partial \dot{\alpha}}\right) = 0, \quad X_L \in \Omega, \quad t \in [t_1, t_2]. \end{aligned}$$

Inserting now constitutive relations (3) into the above equations they transform to

$$\frac{DT_{Li}}{DX_L} - \frac{D}{Dt}(\rho_R \dot{x}_i) = 0, \tag{23}$$

$$\frac{DS_L}{DX_L} + \frac{D\eta}{Dt} = 0. \tag{24}$$

Thus the variational statemenet provides two equations. The first of them, i.e., eq. (23) is the equation of motion and coincides with the corresponding one of G–N theory, that is eq. (1). The second one, i.e. eq. (24), is an equation for the balance of entropy which is also included in the treatmenet of Green and Nahdhi [1]. Thus, the Proposition 3.1 is partly proved. □

Remark 3.1 As far as the variational principle is considered, the field equations for thermoelasticity without dissipation are eqs. (23)–(24) instead of (1)–(2) of G–N theory. The other required equation, i.e. eq. (2), is an energy equation and can not directly rise from a variational principle. What can be expected is to appear as a consequence of Noether's theorem considering invariance in time translations.

4 Variational Symmetries and Conservation Laws

Having obtained the variational principle, we can proceed to explore particular cases of variational symmetries related to it.

4.1 Invariance under Translations

First, we shall consider invariance in material space and time translations.

Lemma 4.1 *If the action integral of a thermoelastic body without dissipation is invariant under the group of space and time translations, then it is a homogeneous body.*

PROOF The group of translations in material space and time is given by the following relations

$$\begin{aligned} &\tilde{X}_A = X_A + \epsilon\delta_{wA}, \quad w = 1,2,3,4, \\ &\tilde{x}_\alpha = x_\alpha, \end{aligned} \tag{25}$$

which means that $Z_A^w = \delta_{wA}$ and $\zeta_\alpha^w = 0$. After that the proof is a simple consequence of the invariance identity (14), which for the group (25) results in

$$\frac{\partial L}{\partial X_L} = 0, \quad \frac{\partial L}{\partial t} = 0. \tag{26}$$

The second of (26) is satisfied by the definition of the Lagrangian and the first one is what we want to prove. □

Next, we give the main result of this subsection.

Proposition 4.1 *Let the motion and the temperature functions x_i and Θ satisfy the field equations* (23–24) *for a homogeneous thermoelastic body without dissipation through constitutive relations* (3), *on the domain $\Omega \times [t_1, t_2]$. Then the following conservation laws also hold on $\Omega \times [t_1, t_2]$.*

$$\frac{D}{DX_L}(L\delta_{KL} + T_{Li}x_{i,K} - S_L\beta_K) - \frac{D}{Dt}(\rho_R\dot{x}_i x_{i,K} + \eta\beta_K) = 0, \tag{27}$$

$$\frac{D}{DX_L}(T_{Li}\dot{x}_i - S_L\Theta) + \frac{D}{Dt}(L - \rho_R\dot{x}_i\dot{x}_i - \eta\Theta) = 0. \tag{28}$$

PROOF The assumptions of the proposition imply that the symmetry (25) holds, thus we can use the general form of the conservation laws (15). With the aid of eq. (16), the quantity θ_A^w for the case under discussion will take the form

$$\theta_A^w = -L\delta_{wA} + \frac{\partial L}{\partial x_{\alpha,A}}x_{\alpha,w} \quad w = 1, 2, 3, 4 \tag{29}$$

and the corresponding conservation laws will be

$$-\frac{D\theta_A^w}{DX_A} = \frac{D}{DX_A}(L\delta_{wA} - \frac{\partial L}{\partial x_{\alpha,A}}x_{\alpha,w}) = 0, \quad w = 1, 2, 3, 4 \tag{30}$$

Equation (30) can be analyzed in

$$\frac{D}{DX_A}(L\delta_{LA} - \frac{\partial L}{\partial x_{\alpha,A}}x_{\alpha,L}) = 0, \quad L = 1, 2, 3$$

and

$$\frac{D}{DX_A}(L\delta_{4A} - \frac{\partial L}{\partial x_{\alpha,A}}x_{\alpha,4}) = 0.$$

Furthermore, developing the last equations we obtain

$$\begin{aligned} &\frac{D}{DX_K}(L\delta_{LK} - \frac{\partial L}{\partial x_{i,K}}x_{i,L} - \frac{\partial L}{\partial \alpha_{,K}}\alpha_{,L}) - \frac{D}{Dt}(\frac{\partial L}{\partial \dot{x}_i}x_{i,L} + \frac{\partial L}{\partial \dot{\alpha}}\alpha_{,L}) = 0, \\ &-\frac{D}{DX_K}(\frac{\partial L}{\partial x_{i,K}}\dot{x}_i + \frac{\partial L}{\partial \alpha_{,K}}\dot{\alpha}) + \frac{D}{Dt}(L - \frac{\partial L}{\partial \dot{x}_i}\dot{x}_i - \frac{\partial L}{\partial \dot{\alpha}}\dot{\alpha}) = 0. \end{aligned} \tag{31}$$

Finally, inserting the constitutive relations in (31) we obtain the required relations (27–28), hence Proposition 4.1 has been proved. □

The second of the just obtained conservation laws, eq. (28), corresponds to time translations, thus it is related to the conservation of energy. In the sequence, we shall prove that it could provide the second of the equations of G–N theory, i.e., eq. (2). Indeed, inserting the lagrangian (17) into (28) we obtain

$$\frac{D}{DX_L}(T_{Li}\dot{v}_i - S_L\Theta) - \frac{D}{Dt}(\frac{1}{2}\rho_R v_i v_i + \Psi + \eta\Theta) = 0. \tag{32}$$

After some simple calculation and taking into account equation of motion (23), we can write eq. (32) in the form

$$-(\dot{\Psi} + \dot{\Theta}\eta) + T_{Li}v_{i,L} - S_L\Theta_{,L} = 0. \tag{33}$$

Equation (33) coincides with eq. (2), hence Proposition 3.1 has been completely prooved.

Remark 4.1 It is noted that to obtain eq. (33), the homogeneity of the Lagrangian with respect to the material space variables is not required. What is really necessary is the homogeneity with respect to time which is assumed by the Definition 3.1 of the Lagrangian . That's why there is no requirement of homogeneity in the statement of Proposition 3.1.

Remark 4.2 It is easy one to confirm that the invariance of the action integral (18) under the group of translations in physical space will provide the Euler–Lagrange equations, i.e, eqs. (27) and (28).

4.2 Invariance under the Scaling Group

In this case we will use the following one–parameter group of scalings in material and physical space

$$\begin{aligned} \tilde{X}_A &= X_A + \epsilon X_A, \\ \tilde{x}_\alpha &= x_\alpha - \epsilon x_\alpha, \end{aligned} \tag{34}$$

Thus invoking the relations (10) we obtain

$$Z_A = X_A, \quad \zeta_\alpha = -x_\alpha, \tag{35}$$

which by turn are substituted into the invariance identity (14) to obtain

Lemma 4.2 *The action integral of a thermoelastic body without dissipation is invariant under the transformation group* (34) *if and only if its Lagrangian fulfils the identity*

$$\frac{\partial L}{\partial X_A}X_A - 2(\frac{\partial L}{\partial x_{\alpha,A}})x_{\alpha,A} + 4L = 0. \tag{36}$$

The most interestig result for the scaling group concerns the linear case for which the following identity holds

Lemma 4.3 *Let us assume that a thermoelastic body without dissipation admits linear constitutive relations risen from* (3)*, then it's Lagrangian satisfies the following identity*

$$\frac{\partial L}{\partial x_{\alpha,A}} x_{\alpha,A} = 2L. \tag{37}$$

PROOF Let us assume that the free energy function is a quadratic function of the indepedent constitutive variables

$$\begin{aligned}
&\Psi(X_L, \frac{\partial x_i}{\partial X_L}, \dot{\alpha}, \frac{\partial \alpha}{\partial X_L}) = \\
&\frac{1}{2}[c_{ijKL} x_{i,K} x_{j,L} + e_{KL}\beta_K\beta_L + d\Theta^2] + \\
&\qquad c_{iKL} x_{i,K}\beta_L + c_{iK} x_{i,K}\Theta + d_K\Theta\beta_K \\
&\qquad\qquad i,j,K,L = 1,2,3,
\end{aligned} \tag{38}$$

so as to obtain the linear constitutive relations

$$\begin{aligned}
T_{Ki} &= \frac{\partial \Psi}{\partial x_{i,K}} = \frac{1}{2}(c_{ijKL} + c_{jiLK}) x_{j,L} + c_{iKL}\beta_L + c_{iK}\Theta, \\
-S_K &= \frac{\partial \Psi}{\partial \beta_K} = \frac{1}{2}(e_{KL} + e_{LK})\beta_L + c_{iKL} x_{i,L} + d_K\Theta, \\
-\eta &= \frac{\partial \Psi}{\partial \Theta} = d\Theta + c_{iK} x_{i,K} + d_K\beta_K,
\end{aligned} \tag{39}$$

where $c_{ijKL}, e_{KL}, d, d_K, c_{iKL}, c_{iK}$ are material functions depending on material space variables X_L. In the case of homogemeous body they reduce to material constants. After eq. (38), it is a matter of a straight calculation to obtain relationship (37). □

Proposition 4.2 *Let the motion and the temperature functions x_i and Θ satisfy the field equations* (23–24) *for a homogeneous thermoelastic body without dissipation through linear constitutive relations* (39)*, on the domain $\Omega \times [t_1, t_2]$. Then the following conservation law also holds on $\Omega \times [t_1, t_2]$.*

$$\begin{aligned}
&\frac{D}{DX_L}[(L\delta_{KL} + T_{Li} x_{i,K} - S_L\beta_K) X_K + (T_{Li}\dot{x}_i - S_L\Theta)t + T_{Li} x_i - S_L\alpha] + \\
&\frac{D}{Dt}[-(\frac{1}{2}\rho_R\dot{x}_i\dot{x}_i + e)t - (\rho_R\dot{x}_i x_{i,K} + \eta\beta_K) X_K - \rho_R\dot{x}_i x_i - \eta\alpha] = 0,
\end{aligned} \tag{40}$$

where $e = e(\mathbf{X}, t)$ is the internal density function per unit volume.

PROOF The linearity of the thermoelastic body implies that the Lagrangian meets the relation (37). Furthermore, in virtue of the homogeneity of the body, the invariance identity (36) also holds. Thus, for the case under discussion, the action integral is invariant under the group (34). That means we can invoke the general form for the conservation law given by eq. (15). Due to the fact that the group (34) has only one parameter the quantity (16) will take the form

$$\theta_A = -(LX_A - \frac{\partial L}{\partial x_{\alpha,A}} x_{\alpha,B} X_B - \frac{\partial L}{\partial x_{\alpha,A}} x_\alpha) \tag{41}$$

and the conservation law corresponding to the symmetry (34) will be

$$\begin{aligned} &-\frac{D\theta_A}{DX_A} = 0 \Rightarrow \\ &\frac{D}{DX_A}\left(LX_A - \frac{\partial L}{\partial x_{\alpha,A}} x_{\alpha,B} X_B - \frac{\partial L}{\partial x_{\alpha,A}} x_\alpha\right) = 0. \end{aligned} \tag{42}$$

The equation (42) can be analyzed as follows

$$\begin{aligned} &\frac{D}{DX_L}\left(LX_L - \frac{\partial L}{\partial x_{\alpha,L}}(x_{\alpha,M} X_M + \dot{x}_\alpha t) - \frac{\partial L}{\partial x_{\alpha,L}} x_\alpha\right) + \\ &\qquad \frac{D}{Dt}\left(Lt - \frac{\partial L}{\partial \dot{x}_\alpha}(x_{\alpha,M} X_M + \dot{x}_\alpha t) - \frac{\partial L}{\partial \dot{x}_\alpha} x_\alpha\right) = 0 \Rightarrow \end{aligned}$$

$$\begin{aligned} &\frac{D}{DX_L}\left(LX_L - \frac{\partial L}{\partial x_{i,L}}(x_{i,M} X_M + \dot{x}_i t) - \frac{\partial L}{\partial \alpha_{,L}}(\alpha_M X_M + \dot{\alpha} t) - \frac{\partial L}{\partial x_{i,L}} x_i - \frac{\partial L}{\partial \alpha_{,L}} \alpha\right) + \\ &\frac{D}{Dt}\left(Lt - \frac{\partial L}{\partial \dot{x}_i}(x_{i,M} X_M + \dot{x}_i t) - \frac{\partial L}{\partial \dot{\alpha}}(\alpha_M X_M + \dot{\alpha} t) - \frac{\partial L}{\partial \dot{x}_i} x_i - \frac{\partial L}{\partial \dot{\alpha}} \alpha\right) = 0, \end{aligned} \tag{43}$$

Taking now into account that the internal energy is related with the free energy function by

$$e = \Psi - \Theta\eta$$

and invoking the relations (3) and (17) we can easily obtain from (43) the required conservation law (40). □

4.3 Invariance under the Group of Rotations

In this section we will examine the invariance of action integral (18) under rotations of spatial variables. That is the group $SO(3)$ in the physical space which is given by the following equations

$$\begin{aligned} &\tilde{X_A} = X_A, \quad A = 1,2,3,4, \\ &\tilde{x_i} = Q_{ij} x_j, \quad i,j = 1,2,3, \\ &\tilde{x_4} = x_4, \end{aligned} \tag{44}$$

where $\mathbf{Q}$ is an orthogonal matrix with $det\mathbf{Q} = 1$. The corresponding infinitesimal group is given by

$$\begin{aligned} &\tilde{X_K} = X_K, \\ &\tilde{X_4} = X_4, \\ &\tilde{x_i} = x_i + e_{iwj}\epsilon_w x_j, \\ &\tilde{x_4} = x_4, \quad i,j,K,w = 1,2,3. \end{aligned} \tag{45}$$

So, we take for the quntities given by eqs. (10)

$$Z_K^w = Z_4^w = 0, \quad \zeta_i^w = e_{iwj}x_j, \quad \zeta_4^w = 0. \tag{46}$$

Inserting eqs. (46) into the invariance identity (14), we obtain the following result

Lemma 4.4 *The action integral of a thermoelastic body without dissipation is invariant under the transformation group* (44)*, if and only if its Lagrangian fulfils the identity*

$$e_{iwj}\frac{\partial L}{\partial x_{i,K}}x_{j,K} = 0. \tag{47}$$

Invoking the constitutive relations (4) we easily obtain

$$e_{wij}T_{Ki}x_{j,K} = 0. \tag{48}$$

After that, one can easily prove the following proposition

Proposition 4.2 *Let the motion and the temperature functions x_i and Θ satisfy the field equations* (23–24) *for a thermoelastic body without dissipation through constitutive relations* (3)*, on the domain $\Omega \times [t_1, t_2]$. Moreover, let the lagrangian fulfils the identity* (47)*. Then the following conservation law also holds on $\Omega \times [t_1, t_2]$.*

$$\frac{D}{DX_L}\left(e_{iwj}x_jT_{Ki}\right) - \frac{D}{Dt}\left(\rho_R e_{iwj}\dot{x}_i x_j\right) = 0 \tag{49}$$

Remark 4.3 Equation (49) is the balance of *angular momentum* for the G–N theory. Certainly, this equation jointly with equation of momentum (eq. 23) can provide us, as usually, the symmetry of the tensor $T_{Li}x_{j,L}$. But this is erroneous because it is an assumption for the Proposition 4.2 (see eq. 48) and not a consequence of it.

5 Material Balance Laws

So far, we have presented conservation laws of the G–N equations of thermoelasticity. From the point of view of material mechanics, it is interesting to focus on

what can be called *material balance laws*, that is the corresponding to conservation laws non–homogeneous equations. To obtain such equations we must allow for the presence of sources in the already derived equations. This by turn, can be done by relaxing the assumptions we have posed in order to obtain them. By this manner, for every conservation law, having found the conserved quantity, we can obtain a balance (non-conservation) law. Depending on the particular equation, we expect these non–homogeneous terms to be the so–called *material forces* or *moment* of such forces.We will apply this procedure to conservation laws (27) and (40).

• The Canonical Momentum Balance Law

We proceed now to the above mentioned procedure for the eq. (27) by removing the homogeneity of the Lagrangian from the assumptions of the Proposition 4.1. Let assume that the Langragian (17) does depend explicitly on material variables, that is

$$\frac{\partial L}{\partial X_L} = 0, \quad L = 1, 2, 3.$$

In this case no symmetry with respect to space translations can be secured. In spite of this, the Euler–Lagrange equations still hold and it is possible to produce a balance law by calculating the expression

$$\frac{D\theta_A^w}{DX_A}, \quad w = 1, 2, 3 \quad A = 1, 2, 3, 4$$

which certainly does not vanish identically any more. Indeed, calculating this term we obtain

$$\begin{aligned}
\frac{D\theta_A^w}{DX_A} &= -\frac{D}{DX_A}\left(L\delta_{wA} - \frac{\partial L}{\partial x_{\alpha,A}} x_{\alpha,w}\right) \\
&= -\frac{\partial L}{\partial X_w} - \frac{\partial L}{\partial x_{\alpha,A}} x_{\alpha,Aw} + \frac{D}{DX_A}\left(\frac{\partial L}{\partial x_{\alpha,A}}\right) x_{\alpha,w} + \frac{\partial L}{\partial x_{\alpha,A}} \frac{D}{DX_A}\left(x_{\alpha,w}\right) \\
&= -\frac{\partial L}{\partial X_w} + \frac{D}{DX_A}\left(\frac{\partial L}{\partial x_{\alpha,A}}\right) x_{\alpha,w}. \qquad (50)
\end{aligned}$$

The last term on the right hand side of (50) vanishes due to the Euler–Lagrange equations, thus we conclude

$$\frac{D\theta_A^w}{DX_A} = -\frac{\partial L}{\partial X_w}. \qquad (51)$$

But the l.h.s. of eq. (51) has already been calculated in Proposition 4.1, hence equating the r.h.t. of eqs. (30) and (51) and taking into account eq. (31a), we can write

$$\frac{D}{DX_L}(L\delta_{KL} + T_{Li}x_{i,K} - S_L\beta_K) - \frac{D}{Dt}(\rho_R \dot{x}_i x_{i,K} + \eta\beta_K) = \frac{\partial L}{\partial X_K}, \quad K, L = 1, 2, 3, \quad (52)$$

which is the expected material balance law refered by some authors [5] as *canonical momentum equation* or *pseudomomentum equation.*

• **The Scalar Moment of Canonical Momentum Balance Law**

In the same way, we can elaborate the conservation law (40) related to scaling symmetry. Hence, removing the assumptions of constitutive relations linearity and the homogeneity of the Lagrangian with respect to material variables we calculate directly the quantity

$$
\begin{aligned}
&-\frac{D\theta_A}{DX_A} = \\
&\frac{D}{DX_A}\left(LX_A - \frac{\partial L}{\partial x_{\alpha,A}}x_{\alpha,B}X_B - \frac{\partial L}{\partial x_{\alpha,A}}x_\alpha\right) = \\
&\frac{\partial L}{\partial X_L}X_L - 2(\frac{\partial L}{\partial x_{\alpha,A}})x_{\alpha,A} + 4L, \qquad (53)\\
&\qquad\qquad \alpha, A = 1,2,3,4, \quad L = 1,2,3.
\end{aligned}
$$

Inserting now into eq. (53) the already estimated left hand side from Proposition 4.2, we obtain the following balance law

$$
\begin{aligned}
&\frac{D}{DX_L}[(L\delta_{KL} + T_{Li}x_{i,K} - S_L\beta_K)X_K + (T_{Li}\dot{x}_i - S_L\Theta)t + T_{Li}x_i - S_L\alpha] + \\
&\qquad \frac{D}{Dt}[-(\frac{1}{2}\rho_R\dot{x}_i\dot{x}_i + e)t - (\rho_R\dot{x}_ix_{i,K} + \eta\beta_K)X_K - \rho_R\dot{x}_ix_i - \eta\alpha] = \\
&\qquad\qquad \frac{\partial L}{\partial X_L}X_L - 2(\frac{\partial L}{\partial x_{\alpha,A}})x_{\alpha,A} + 4L. \qquad (54)
\end{aligned}
$$

We remark that equation (54) is valid for *non-linear, non-homogeneous thermoelasticity* in the framework of G–N theory. Hence, for *linear thermoelasticity*, invoking Lemma 4.3, the balance law originated by scaling symmetry will become

$$
\begin{aligned}
&\frac{D}{DX_L}[(L\delta_{KL} + T_{Li}x_{i,K} - S_L\beta_K)X_K + (T_{Li}\dot{x}_i - S_L\Theta)t + T_{Li}x_i - S_L\alpha] + \\
&\frac{D}{Dt}[-(\frac{1}{2}\rho_R\dot{x}_i\dot{x}_i + e)t - (\rho_R\dot{x}_ix_{i,K} + \eta\beta_K)X_K - \rho_R\dot{x}_ix_i - \eta\alpha] = \frac{\partial L}{\partial X_L}X_L. \qquad (55)
\end{aligned}
$$

Eq. (55) holds for linear, non–homogeneous thermoelasticity and represents a balance law for scalar moment of pseudomomentum or canonical momentum. The coresponding balance equation in physical space is not often used because it does not play any role in the description of the equilibrium or the motion of a body as does, for instance, the momemtum or angular momentum equation. In the case of physical space, the factors that balance the rate of scalar momet of momentum are refered as *scalar moments* or *virials*. So the right hand side term of eq. (55) is a sort of *material scalar moment* or *material virial*.

6 Comparisons and Conclusions

The above–obtained results can be compared to previously appeared work of other researchers. We must especially refer to the work of Dascalu and Maugin [2] for G–N thermoelasticity and Maugin [5] and Fletcher [11] for elasticity. Let us return to eq. (27) which represents the canonical momentum equation. It can be witten in the form

$$\frac{D}{DX_L}((\frac{1}{2}\rho_R\dot{x}_i\dot{x}_i - \Psi)\delta_{KL} + T_{Li}x_{i,K} - S_L\beta_K) - \frac{D}{Dt}(\rho_R\dot{x}_i x_{i,K} + \eta\beta_K) = 0,$$

or

$$Div((\rho_R\frac{\mathbf{v}^2}{2} - \Psi)\mathbf{I} + \mathbf{TF} - \beta\otimes\mathbf{S}) - \frac{D}{Dt}(\rho_R\mathbf{F}^T\mathbf{v} + \eta\beta) = 0. \tag{56}$$

Equation (43) coincides with the corresponding one deriving through a vectorial approach in [2]. Conservation laws (27), (28) and (40) restricted to the case of elasticity are in full agreement with the corresponding ones given by [5] and [11]. To obtain a more clear collation we introduce the following definitions [5, 2]

$$\begin{aligned} b_{LK} &= -(L\delta_{LK} + T_{Li}x_{i,K} - S_L\beta_K),\\ Q_L &= T_{Li}\dot{x}_i - S_L\Theta,\\ \mathcal{H} &= \frac{1}{2}\rho_R\dot{x}_i\dot{x}_i + e,\\ \mathcal{P}_L &= -(\rho_R\dot{x}_i x_{i,L} + \eta\beta_L), \end{aligned} \tag{57}$$

for *Eshelby stress tensor, material flux energy, Hamiltonian,* and *pseudomomentum vector* respectively.
After definitions (57) , the balance laws (52) and (55) can be written in vectorial form as follows

$$-Div\mathbf{b} + \frac{D\mathcal{P}}{Dt} = \mathbf{f}^{inh}. \tag{58}$$

and

$$\begin{aligned} Div(-\mathbf{bX} \;+\; &\mathbf{Q}t + \mathbf{Tx} - \mathbf{S}\alpha) +\\ &\frac{D}{Dt}(-\mathcal{H}t + \mathcal{P}\cdot\mathbf{X} - \mathbf{p}\cdot\mathbf{x} - \eta\alpha) = \mathbf{f}^{inh}\cdot\mathbf{X}, \end{aligned} \tag{59}$$

where following Maugin [5] we have defined

$$f_L^{inh} = \frac{\partial L}{\partial X_L}$$

for *material force.* We recall that the last equation holds for non-homogeneous but *linear* thermoelasticity of G–N. Under this restriction and in the framework of elasticity it can be compared with eq. (4.89) of [5].

References

[1] A. E. GREEN and P.M. NAGHDI, *J. Elasticity* **31**, 189, (1993)

[2] C. DASCALU and G. A. MAUGIN , *J. Elasticity* **39**, 201 (1995)

[3] D. S. CHANDRASEKHARAIAH , *Appl. Mech. Rev.* **51**, 705 (1998)

[4] R. B. HETNARSKI and J. I. IGNACZAK , *Int. J. Solids Struct.* **37**, 215 (2000)

[5] G. A. MAUGIN, *Material Inhomogeneites in Elasticity*, In Applied Mathematics and Mathematical Computation, 3, Chapman and Hall, London (1993)

[6] M. E. GURTIN, *Configurational Forces as Basic Concepts of Continuum Physics*, In Applied Mathematical Sciences, 137, Sprimger, New York (2000)

[7] R. KIENZLER and G. HERRMANN, *Mechanics in Material Space*, Sprimger, Berlin (2000)

[8] D. LOVELOCK and H. RUND, *Tensors, Differential Forms and Variational Principles* John Wiley and sons, London (1975)

[9] G. W. BLUMAN, S. KUMEI *Symmetries and Differential Equations.* In Applied Mathematical Sciences, 81, Springer, New York (1989)

[10] P. J. OLVER, *Applications of Lie Groups to Differential Equations.* In Graduate Texts in Mathematics, 107, Springer, New York (1993)

[11] D. C. FLETCHER , *Arch. Rat. Mech. Anal.* **60**, 329 (1976)

Address: Assist. Professor Dr. V. K. KALPAKIDES
Division of Applied Mathematics and Mechanics,
Department of Mathematics,
University of Ioannina, GR-45110, Ioannina, GREECE
E-mail address: vkalpak@cc.uoi.gr

A NEARLY-PERIODIC BOUNDARY VALUE PROBLEM FOR SECOND ORDER DIFFERENTIAL EQUATIONS

G. L. KARAKOSTAS AND P. K. PALAMIDES

ABSTRACT. By utilizing a combination of properties of the consequent mapping with the Brouwer's fixed point theorem we obtain existence results for the nearly-periodic boundary value problem

$$x'' = f(t, x, x'), \ \ t \in [0, 1]$$

$$x(1) = Q_0^{-1}x(0), x'(1) = Q_1^{-1}x'(0),$$

where Q_0, Q_1 are complex valued nonsingular matrices.

1. INTRODUCTION

Let $\mathbb{C}^n$ denote the $n-$dimensional complex Euclidean linear space and let I be the interval $I := [0, 1]$. Let Ω be a convex, open subset of the product space $\mathbb{C}^n \times \mathbb{C}^n$ and let $f : I \times \Omega \to \mathbb{C}^n$ be a continuous function. In this paper we provide sufficient conditions for the existence of a (complex valued) solution x of the vector differential equation

$$x'' = f(t, x, x'), \ \ t \in I \tag{1.1}$$

satisfying the conditions

$$x(1) = Q_0^{-1}x(0), \ \ x'(1) = Q_1^{-1}x'(0), \tag{1.2}$$

where Q_0, Q_1 are nonsingular $n \times n$ complex valued matrices. The problem under investigation is inspired by the periodic problem (in the real case) concerning (1.1), for which the literature is volunimous, as well as by those problems presented in [2, 4]. In [2] the existence of a Sturm-Liouville boundary value problem is investigated, by transforming it into the equivalent form $Lx = Gx$ and then applying the Leray-Schauder's continuation theorem. Also we would like to refer to [5, p. 338], where by using the Wazewski's method it was shown that, if in (1.1) the function f satisfies the well known Hartman's condition for all $t \geq 0$, x and $y \neq 0$, then there is a time $t_0 > 0$ such that $x(t).x(t)$ is nonincreasing on $t \geq t_0$, where x is the solution of (the real version of) equation (1.1). For a two-point boundary value problem concerning a more general differential equation in a Hilbert space discussed by the authors in [7] the Schauder' s fixed point theorem is used. Notice that in [4], the existence

2000 *Mathematics Subject Classification.* Primary 34B15; Secondary 34C25.
Key words and phrases. Boundary value problems, nearly-periodic solutions, egress points.

of a solution x of the problem is investigated, where the nonsingular $n \times n$-square matrices Q_0 and Q_1 satisfy the inequalities

$$x.Q_0Q_1^{-1}y \leq 0 \text{ and } x.(Q_0 + Q_1^{-1})y \leq 0, \tag{1.3}$$

for all vectors $x, y \in \mathbb{R}^n$ with $x.y \leq 0$ and the matrix Q_0 is orthogonal. (The dot denotes the inner product in the real Euclidean space.)

The literature shows a great number of papers refered to both the scalar and the vector case for the problem (1.1), (1.2), see, e.g., [1, 9, 10, 11] and the references therein. In [3] Erbe by using a technique, which involves a direct application of properties of Leray-Schauder degree, instead of (1.3), he used the following condition:

$$\text{There is a } \mu > 0 \text{ such that } Q_1 = \mu Q_0.$$

A more general situation of the problem is discussed by the authors in [8]. In this paper we do use of the Hartman's condition and give information on the existence of solutions by combining properties of the consequent mapping with the Brouwer's fixed point theorem. Motivated from Erbe's technique, instead of the orthogonality condition on Q_0, we assume that the matrices Q_0, Q_1 satisfy the relation

$$\|Q_1\| \leq \|Q_0\| = \|Q_0^{-1}\| = 1, \tag{1.5}$$

where $\|.\|$ stands for the norm in the $n \times n$ complex matrix space congruent to the euclidean norm of the complex n-dimensional space $\mathbb{C}^n$, the norm which equals to the greatest absolute value of its eigenvalues.

2. Preliminaries

Let J be a fixed interval of the real line such that $I \subset J$. Consider equation (1.1) associated with the initial conditions

$$(\tau, x(\tau), x'(\tau)) =: (\tau, \xi, \eta) =: P \in I \times \Omega, \tag{2.1}$$

where the function $f : J \times \Omega \to \mathbb{C}^n$ is continuous. Let $X(P)$ be the family of all solutions of (1.1), (2.1). If x is such a solution, we shall write I_x to denote the connected set of all existence times of x lying in I and such that $0 \in I_x$. We let $D := J \times \Omega$ and consider this set as a subset of the euclidean space $\mathbb{R} \times \mathbb{C}^n$. Take a subset W of D such that both the sets $int(W)$ and $D - cl(W)$ are nonempty. (Here $int(W)$ denotes the interior and $cl(W)$ the closure of the set W.) Later on the set W will be completely definite.

Next we recall some classical definitions. Given a $\tau \in (0, 1]$, a point $P := (\tau, \xi, \eta)$ of the boundary of W (if such exists) is *a point of egress*, if, given any $x \in X(P)$, there is an $\epsilon > 0$ such that

$$\{(t, x(t), x'(t)) : t \in (\tau - \epsilon, \tau)\} \subset int(W).$$

Also, if $\tau < 1$, then P is *a strict egress point*, if, given any $x \in X(P)$, there is an $\epsilon > 0$ such that

$$\{(t, x(t), x'(t)) : t \in (\tau, \tau + \epsilon)\} \subset D - cl(W).$$

(See, e.g., [6].) We denote by W^e and W^{se}, respectively, the sets of egress and strict egress points of W.

A point P of the boundary of W is *a consequent point* of $P_0 := (\tau_0, \xi_0, \eta_0)$, if there is a solution passing from both these points and such that

$$\{(t, x(t), x'(t)) : t \in (\tau_0, \tau)\} \subset int(W).$$

The set of all consequent points of P_0 will be denoted by $C(P_0)$, while the so defined (set-valued) mapping

$$\mathrm{C} : N_c(W) \to W^e$$

is *the consequent mapping.* Here the symbol $N_c(W)$ stands for the set of all points of W whose the sets of the consequent points are nonempty.

Given a time $\tau \in (0, 1]$ we say that a point $P := (\tau, \xi, \eta)$ of the boundary of W is *a point of ingress* of W, if given any solution $x \in X(P)$ there is an $\epsilon > 0$ such that

$$\{(t, x(t), x'(t)) : t \in (\tau - \epsilon, \tau)\} \subset D - cl(W).$$

Also, in case $\tau < 1$, the point P is *a strict ingress point*, if given any $x \in X(P)$, there is an $\epsilon > 0$ such that

$$\{(t, x(t), x'(t)) : t \in (\tau, \tau + \epsilon)\} \subset int(W).$$

We denote by W^i and W^{si}, respectively, the sets of ingress and strict ingress points of W.

It is clear that, if uniqueness of the solutions holds, then the consequent mapping is a single valued function.

Now assume that X, Y are topological spaces and let F be an abstract set-valued mapping which maps the points of X to nonempty compact subsets of Y. Then F is upper-semicontinuous (usc) at a point x_0 of X, if for any open subset A of $F(x_0)$ there exists a neibghborhood U of x_0 such that the set $F(x)$ is a subset of A for all points x of U.

The following lemmas give sufficient conditions for the upper semi-continuity of the consequent mapping and some useful properties for a class of usc mappings. Notice that the consequent mapping C is included in this class (see, e.g., [6]).

Lemma 2.1. *If for any point P of $S_c(W)$ all functions in $X(P)$ egress strongly from W, then the consequent mapping C is usc at any point P and the image $C(P)$ is a continuum subset of the boundary of W.*

Lemma 2.2. *Let X, Y be metric spaces and let $F : X \to 2^Y$ be a usc set-valued mapping. If A is a continuum subset of X such that for every $x \in A$ the image $F(x)$ is a continuum, then the image $F(A) := \cup\{F(x) : x \in A\}$ is also a continuum subset of Y.*

3. The main results

This section is devoted to the main results of the paper. We shall denote by $\bar{z}$ the conjugate of the complex number z and by $Re[z]$ its real part. Also the "typical" inner product in the n-dimensional space will be denoted by $< ., . >$.

Assume that the open set Ω has the property that there is a real number $R > 0$ such that

$$V := \cup\{V(t) : t \in I\} \subset I \times \Omega,$$

where for each $t \in I$ we have set

$$V(t) := \{(t, x, y) : |x| \leq R, \ y \in \mathbb{C}^n\}.$$

In the sequel a bar over a matrix will denote the matrix with elements the complex conjugates of the elements of the original matrix.

Theorem. *Consider equation* (1.1) *where the continuous function* $f : J \times \Omega \to \mathbb{C}^n$ *satisfies the following conditions:*

(F_1) *For any* $t \in I$ *and* (t, x, y) *in the boundary of* $V(t)$ *the implication*

$$\text{if } Re[< \bar{x}, y >] = 0, \ \text{ then } \ Re[< \bar{x}, f(t, x, y) > + |y|^2] \neq 0$$

holds.

(F_2) *There is a positive real number* M *such that any solution* $x \in X(V(0))$ *with* $|x'(0)| \leq M$, *satisfies the inequality*

$$|x'(t)| < M,$$

for all $t \in I_x$ *such that* $t > 0$ *and* $(t, x(t), x'(t)) \in V$.

Also assume that the nonsingular $n \times n$ *complex matrices* Q_0, Q_1 *are such that*

(M_1) *condition (1.5) is satisfied, and*

(M_2) *for all* $x, y \in \mathbb{C}^n$ *with*

$$Re[< \bar{x}, y >] \geq 0,$$

it holds

$$Re[< \bar{Q}_0^{-1}\bar{x}, Q_1^{-1}y >] > 0,$$

Then the problem (1.1), (1.2) *admits a solution* $x(t), t \in I$ *such that*

$$|x(t)| \leq R,$$

for all $t \in I$.

Proof. First of all we would like to notice some remarks:

(a) If we restrict the function f on a compact subset Z of $J \times \Omega$ containing the set V in its interior, we can approximate it uniformly on the set Z by a sequence of functions $f_k(t, x, y)$, which are at least C^1 on Z. For such functions we have uniqueness of solutions passing through points at least of the interior of Z. So, if we show the existence of a sequence of solutions (x_k) of the corresponding problems, with initial conditions in $V(0)$, then these solutions are uniformly bounded by R, their first derivatives by M and their second derivatives by the real number $sup\{|f(t, x, y)| : (t, x, y) \in Z\}$. Hence, by the Arzela-Ascoli's theorem a limiting point of this sequence exists which (according to continuous dependence arguments) will be a solution of the original problem.

(b) Let K be a compact subset of $\mathbb{C}^n \times \mathbb{C}^n$ containing the set

$$E := \{(x, y) \in \mathbb{C}^n \times \mathbb{C}^n : |x| \leq R, \ |y| \leq M\}.$$

Define the continuous real valued function

$$S : (\lambda, x, y) \to Re[< e^{-i\lambda} \bar{Q}_0^{-1} \bar{x}, Q_1^{-1} y >]$$

and observe that, because of (M_2), there is a $\delta > 0$ such that

$$S(\lambda, x, y) > 0,$$

for all $(\lambda, x, y) \in [0, \delta] \times K$. Also, multiplying the matrix Q_0 by the complex factor $e^{i\lambda}$, for some real $\lambda \in (0, \delta)$, we can assume that the unit is not an eigenvalue of the matrix Q_0. Indeed, let us suppose that for each such λ, for which the matrix Q_0 does not have the unit as its eigenvalue, a solution x_λ exists for the corresponding problem. (Notice that each matrix of the form $e^{i\lambda} Q_0$ satisfies, also, condition (M_1).) Then, as in case (a) above, we can get an accumulation point (as the real parameter λ tends to zero), which by continuity, finally, will be a solution of the original problem.

Now consider the set W of all points (t, x, y) of V with $|y| \leq M$ and let W_0 be its cross section at $t = 0$, i.e. the set $W_0 := \{0\} \times E$. Let P_0 be a point in W_0 and let x be the unique solution passing from P_0. Then we distinguish two possibilities:

(i) The set I_x is a subset of I and it holds

$$|x(t)| < R,$$

for all $t \in [0, 1)$. Then we let $s := 1$. It is obvious that there is a point $P \in V(1)$ such that $(1, x(1), x'(1)) = P$.

(ii) Either

$$|x(0)| = R,$$

or there is a time $s \in I$ such that

$$|x(s)| = R \text{ and } |x(t)| < R,$$

for all $t \in [0, s)$.

In both these cases, from (F_2) we have

$$|x'(t)| < M,$$

for all $t \in (0, s)$.

We claim that the point $P := (s, x(s), x'(s))$ is a point of strict egress of the set W. Indeed, in case (i) this fact is obvious. So, consider case (ii), where we can also assume that $s < 1$.

Define the real valued function

$$\phi(t) := |x(t)|^2 - R^2, \ t \in I_x,$$

and observe first that

$$\phi(s) = 0.$$

If

$$\phi'(s) = 2Re[< \bar{x}(s), x'(s) >] > 0,$$

then, clearly, $P \in W^{se}$, in case $s > 0$. If

$$\phi'(s) > 0, \text{ and } s = 0,$$

then

$$P = P_0 \in W^{se}.$$

Notice that in this case we have

$$|x(s)| = |x(0)| = |x_0| = R.$$

Then we can set

$$C(P) = P_0,$$

because the point P might be considered as the consequent point of itself.

If

$$\phi'(s) < 0,$$

then P is a point of strict ingress of W. Clearly, this fact cannot be true in case $s > 0$. If

$$s = 0 \text{ and } \phi'(0) < 0,$$

then the solution x must satisfy either (i), or (ii) above (for a certain new time $s > 0$).

These arguments lead us to discuss only the case

$$s > 0 \text{ and } \phi'(s) = 0.$$

The later means that

$$Re[< \bar{x}(s), x'(s) >] = 0$$

and, so, from (F_1)

$$\phi''(s) = 2Re[< \bar{x}(s), f(s, x(s), x'(s)) > + |x'(s)|^2] \neq 0.$$

The case

$$\phi''(s) < 0$$

is impossible. If

$$\phi''(s) > 0,$$

then we have

$$\phi(t) > 0, \text{ for all } t \in (s, s + \epsilon),$$

for some $\epsilon > 0$. Thus $P \in W^{se}$. Therefore our claim is true.

So far we have proved that

$$C(P_0) = P = (s, x(s), x'(s)).$$

From (1.5) we get

$$|Q_0 x(s)| \leq \|Q_0\| |x(s)| = 1.R = R \tag{3.1}$$

and

$$|Q_1 x'(s)| \leq \|Q_1\| |x'(s)| \leq 1.M = M. \tag{3.2}$$

Next, consider the set E as above and define the mappings

$$H : (x, y) \to (0, x, y) : E \to V(0) \text{ and } h : (t, x, y) \to (x, y) : V \to E,$$

as well as the matrix

$$Q := diag[Q_0, Q_1].$$

Then, from Lemmas 2.1, 2.2, our remark (a) (in the beginning of the proof) and relations (3.1), (3.2), we conclude that the function

$$T(x, y) := Qh(C(H(x, y))) : E \to E$$

maps continuously the closed, convex, bounded set E into itself. Hence, by the Brouwer's fixed point theorem it follows that there is a point $(x_0, y_0) \in E$ such that

$$T(x_0, y_0) = (x_0, y_0).$$

This means that there is a solution x such that $(x(0), x'(0)) = (x_0, y_0) \in E$ and

$$Q_0 x(s) = x_0 \text{ and } Q_1 x'(s) = y_0, \tag{3.3}$$

for some $s \in [0, 1]$.

To finish the proof, it is enough to show that (3.3) is true only for $s = 1$. Indeed, to prove it, we assume, on the contrary, that $s \in [0, 1)$. If $s = 0$, then, as we noticed above, $C(P) = P_0$ and so, it holds $x(s) = x_0$ and $x'(s) = y_0$. Hence from (3.3) we get $Q_0 x_0 = x_0$, where, notice that $|x_0| = R > 0$. This is impossible, because, from our remark (b) above, the unit is not a eigenvalue of the matrix Q_0.

Let us assume that $s \in (0, 1)$. We distinguish two cases:

Case A. Suppose that

$$|x_0| = |x(s)| = R.$$

Then, by the definition of the consequent mapping, the initial point P_0 must be an ingress point of W, so

$$\phi'(0) = 2Re[< \bar{x}(0), x'(0) >] = 2Re[< \bar{x}_0, y_0 >] \leq 0. \tag{3.4}$$

For the same reason the consequent point P is a point of egress of W, hence

$$\phi'(s) = 2Re[< \bar{x}(s), x'(s) >] \geq 0.$$

Then from (3.3) we derive

$$Re[< \bar{Q}_0^{-1}\bar{x}_0), (Q_1^{-1}y_0) >] = Re[< \bar{x}(s), x'(s) >] \geq 0.$$

This fact together with (3.4) contradict to (F_2).

Case B. Suppose that

$$|x_0| < R = |x(s)|.$$

Then we get

$$R = |x(s)| = |Q_0^{-1}x_0| \leq \|Q_0^{-1}\||x_0| < 1.R = R,$$

a contradiction. This completes the proof of the theorem. □

Remark. Hypothesis (F_2) holds, if, for instance, we impose a Nagumo type condition to the function $f(t, x, y)$, namely, if we assume that $f(t, x, y)$ has at most a quadratic growth rate in the argument y.

REFERENCES

[1] J. W. Bebernes and K. Schmitt, *Periodic boundary value problems for systems of second order differential equations*, J. Differential Equations **13** (1973), 32-47.

[2] Dong Yujun, *On solvability of second-order Sturm-Liouville boundary value problems at resonance*, Proc. of AMS **126** (1998), 145-152.

[3] L. H. Erbe, *Boundary value problems for second order differential equations*, Lecture Notes Pure Appl. Math. (N.Y.) **105** (1987).

[4] L. H. Erbe and P. K. Palamides, *Boundary value problems for second order differential equations*, J. Math. Anal. Appl. **117** (1987), 80-92.

[5] J. K. Hale, *Ordinary Differential Equations*, Krieger Publ. Co., Malabar, Florida, 1980.

[6] L. Jackson and P. K. Palamides, *An existence theorem for a nonlinear two-point boundary value problem*, J. Differential Equations **53** (1984), 48-66.

[7] G. L. Karakostas and P. K. Palamides, *A boundary value problem for operator equations in Hilbert spaces*, (to appear).

[8] G. L. Karakostas and P. K. Palamides, *Boundary value problems with compatible boundary conditions*, (to appear).

[9] M. A. Krasnosel'skii, *Translation along trajectories of differential equations*, AMS, No. 19, Providence, 1968.

[10] N. G. Loyd, *Degree Theory*, Cambridge University Press, 1978.

[11] J. Mawhin, *Topological Degree Methods in Nonlinear Boundary Value Problems*, CBMS Regional Conf. Series No. 40, AMS, Providence, 1979.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF IOANNINA, 451 10 IOANNINA, GREECE
E-mail address: gkarako@cc.uoi.gr

OPTIMAL POLICY AND STABILITY REGIONS FOR THE SINGLE PRODUCT LOT SIZING PROBLEM WITH BACKLOGGING

Ioannis GANAS* and Sotirios PAPACHRISTOS
University of Ioannina
Dept. of Mathematics
Probability, Statistics & O.R. Section
451 10 Ioannina
Greece

ABSTRACT

We consider the single product lot sizing problem over a finite planning horizon. In each time period the product is subject to a constant demand, while unsatisfied demand is backlogged. Carrying and backlogging costs are known and remain constant over the planning horizon. The objective is to determine the product lot sizes so as to minimize the total relevant inventory and backlogging costs incurred over the planning horizon. The main contributions of the paper are: the derivation of an analytical expression for the cost of the optimal policy as a function of all parameters in the model, the development of an algorithm, which determines the optimal number of cycles, and the construction of the corresponding stability regions for any set of problem parameters values. The proposed algorithm requires very little computational effort and therefore is suitable for practical use.

Key words: Inventory, lot sizing, backlogging, stability

* Corresponding Author: iganas@cc.uoi.gr

1. Introduction

This paper considers the single product lot sizing problem over a finite planning horizon. In each time period there is a constant demand for the product in question. Demand not satisfied immediately from stock on hand, is completely backlogged and it is satisfied at the beginning of the next period. Holding, backlogging and ordering costs are linear, known and stationary over the planning horizon. The objective is to determine the product lot sizes which minimize the total costs incurred over the planning horizon. The optimal solution for this problem may be obtained using the dynamic programming algorithm proposed by Zangwill [7]. Blackburn et al [1] and Morton [4] have also developed algorithms providing the optimal solution for the problem under consideration. The main disadvantage of the above mentioned algorithms is that they are computationally unattractive. Recently, Federgruen et al [3] developed a simple optimal algorithm solving the problem in linear time. Chand et al [2], allowing for varying demand and holding cost, presented a procedure to determine setup cost stability regions, i.e. sets of setup cost values for which an optimal solution remains valid.

In a previous paper, Papachristos et al [5] studied the same problem, in the non backlogging case. First we considered the set of policies with say n setups. As it is obvious, there are more than one policies with n setups, while among them there is at least one, which is optimal. Working in this set of policies, we obtained the optimal policy. We then proposed a partition of the set of all admissible policies to a class of subsets. Each subset of this partition contains policies with numbers of setups belonging to a set of integers. This partition enabled us to express the *optimal total* cost, as an analytical function of the cost and demand parameters and the number n of setups taken by the policy. Using this analytical expression, we determined the optimal policy within the policies of each subset of this partition. For this optimal policy we constructed its stability region for the cost and demand parameters. This cost function was proved to be convex w.r.t. n, the number of setups considered in the policy. Using this convexity property and the analytical cost function, we presented an algorithm, which solves for the overall optimal policy and constructs its stability region for the cost and demand parameters.

In this paper, we extend the above mentioned results of [5] to the case where backlogging is allowed. The paper is organized as follows: the next section contains the notation, and the mathematical formulation of the problem. In the third section, we present results concerning the structure of the optimal policy and we obtain the optimal one in the set of policies with n setups. In the fourth section, the total cost of the optimal policy for any given number of setups is analytically determined as a function of problem's parameters and the number of setups under consideration. Using this expression we find the optimal policy within

set of policies which constitute a partition of the set of all admissible policies. An example is given explaining the ideas presented. In this section, we also prove the convexity of the total cost function with respect to n, the number of setups considered. This convexity property guarantees the existence of the overall optimal policy. In the fifth section we present an algorithm which determines the overall optimal policy and constructs its corresponding stability region, for any set of cost and demand parameters values. The use of the algorithm is explained with an example. The sixth section contains concluding remarks and directions for further research. The paper ends with an appendix where we give proofs for some of the proposed theorems.

2. Problem Formulation

We consider the single product lot sizing problem with the following characteristics:

1. Demand D for every period is known, constant, and satisfied at the beginning of each period.
2. Ordering (setup) cost S, is constant in every period, it is independent of the quantity ordered, and it is paid every time an order is placed.
3. Holding cost is h per unit of product per period charged to the end-of-period stock.
4. The planning horizon is composed of T discrete-time periods of equal length.
5. Shortages are allowed. Excess demand during a period is completely backlogged and satisfied from ordering at the beginning of the subsequent period.
6. Backlogging cost is b per unit of unsatisfied demand per period, charged to the end-of-period stock.
7. Lot-splitting is prohibited.
8. Lead time is equal to zero.
9. Starting and ending inventory are both set equal to zero.

The following notation will be used subsequently:

y mod (x) = 0 : x and y are positive integers and x is an integer divisor of y.

y mod (x) ≠ 0 : x and y are positive integers and x is not an integer divisor of y.

$\lceil x \rceil$: the smallest integer greater than or equal to x.

$\lfloor x \rfloor$: the largest integer less than or equal to x.

x_i : the lot size ordered at the beginning of period i.

I_i : stock available at the end of period i, with $I_0 = I_T = 0$.

$N_T = \{1, 2, ..., T\}$

$$f(x_i) = \begin{cases} 1, & \text{if } x_i > 0 \\ 0, & \text{if } x_i = 0 \end{cases}$$

The cost for any period i is $Sf(x_i) + h\max(I_i,0) + b\max(-I_i,0)$. Summing up this cost over T we obtain the total cost. The objective is to find those x_i which minimize the total cost over the planning horizon T. The mathematical formulation of the problem is the following:

$$\underset{x_i}{\text{Min}} \sum_{i=1}^{T} Sf(x_i) + h\max(I_i,0) + b\max(-I_i,0) \tag{1}$$

subject to:

$$I_i = \sum_{j=1}^{i} (x_j - D)$$

$$x_i \geq 0,\ I_0 = I_T = 0,\ i = 1, 2, ..., T \tag{2}$$

Any vector $x = (x_1, x_2, \ldots, x_T)$ satisfying (2) will be called a "*policy*". The optimal policy x^*, is the vector giving the minimum in (1), and may not be unique. Define a *cycle* as the time between two consecutive periods with zero end of period stock. In this paper, we consider only those policies, which are such that for any cycle consisting of k_i periods, shortages are allowed for the first ν_i periods, followed by λ_i periods with positive inventory. The holding and backlogging cost for the cycle is then given by $g_{k_i}(\lambda_i, \nu_i) = \frac{\nu_i(\nu_i+1)}{2} bD + \frac{\lambda_i(\lambda_i-1)}{2} hD$.

For the problem under consideration, it is known [7] that in searching for the optimal vector x^*, we must restrict our attention to those x_i values which are such that:

$$x_i = I_{i-1} + (j-i)D,\ i < j \leq T,\ \text{and}\ x_{i+1} = x_{i+2} = \ldots = x_j = 0,\ \text{if}\ I_{i-1} \leq 0$$

$$x_i = 0,\ \text{if}\ I_{i-1} > 0\ ,\ i = 1, 2, ..., T \tag{3}$$

This means that in searching for the optimal policy we must restrict ourselves to the set of policies, which order at period i only if the ending stock of period i-1 is negative, and they do not order if this is positive. The lot size must cover exactly the backlogged demand up to period i and the demand for an integer number of periods following period i. This is called the "*exact requirements criterion for the* x_i *values*". Any period i is called an order period if $x_i > 0$. We shall denote by $P(\bar{z})$ the set of policies satisfying condition (3). Obviously, x_i's satisfy the relation $\sum_{i=1}^{t} x_i = TD,\ 1 \leq t \leq T$, where t indicates the number of none zero x_i's at any policy from $P(\bar{z})$. The realization of the stock level for any such policy is graphically illustrated in Figure 1.

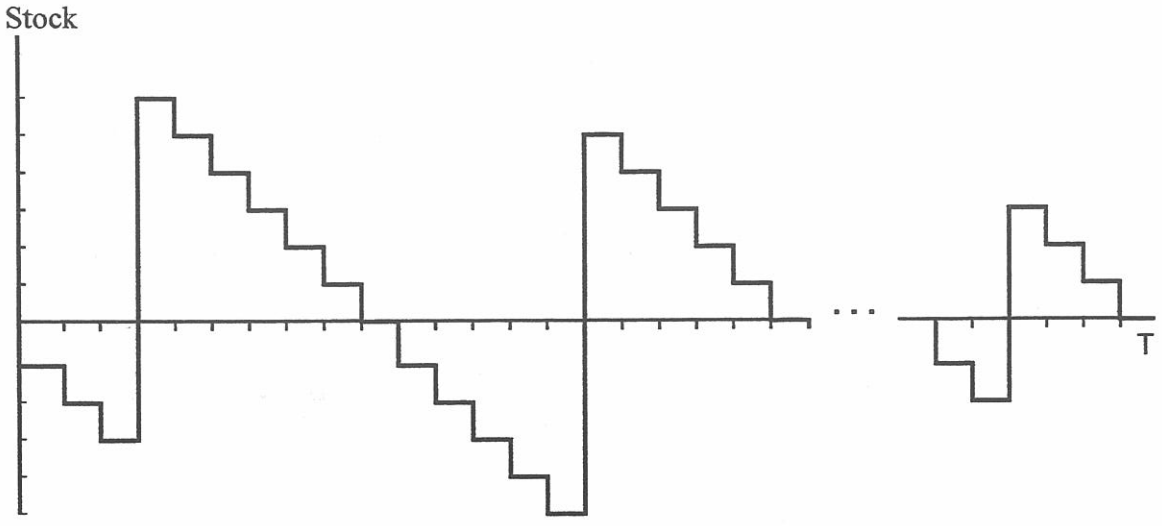


Figure 1. Stock level variation over planning horizon T

We now restrict our attention to those policies, which have exactly n cycles i.e. n setups, which in general have different lengths. Call this set of policies P(n). Obviously, P(n) $\subset$ $P(\bar{z})$. Suppose that any policy from P(n) has n_i cycles of length k_i, i = 1, 2, ..., t, and $k_i \neq k_j$, i≠j. Then we must have, $\sum_{i=1}^{t} n_i = n$ and $\sum_{i=1}^{t} n_i k_i = T$, $1 \le t \le T$. The cost for any such policy is

$$nS + \sum_{i=1}^{t} n_i g_{k_i}(\lambda_i, \nu_i).$$

Let $C_T(n,h,b)$ be the optimal cost for the above problem resulting from the application of the optimal policy with n cycles. Then

$$C_T(n,h,b) = nS + \min_{n_i,\lambda_i,\nu_i} \sum_{i=1}^{t} n_i g_{k_i}(\lambda_i, \nu_i)$$

$$\text{subject to } \sum_{i=1}^{t} n_i(\lambda_i + \nu_i) = T \tag{4}$$

$$\sum_{i=1}^{t} n_i = n \tag{5}$$

$$\lambda_i + \nu_i = \kappa_i$$

$$n_i,\ \lambda_i,\ \nu_i \in Z^+$$

The optimal policy at $P(\bar{z})$ will then be obtained by $C_T(h,b) = \min_{1 \le n \le T} C_T(n,h,b)$.

3. Structure of the Optimal Policy

For any $n \in N_T$ we set $\alpha = \left\lceil \frac{T}{n} \right\rceil$, $\beta = \left\lfloor \frac{T}{n} \right\rfloor = \alpha - 1$, in case Tmod(n) $\neq$ 0 and $\alpha = \frac{T}{n}$, $\beta = \alpha - 1$, in case Tmod(n) = 0. Consider the system

$$\begin{aligned} \alpha x + \beta y &= T \\ x + y &= n \end{aligned} \tag{6}$$

This has the unique integer solution $x = T - n\beta,\ y = \alpha n - T$. Note that in case Tmod(n) = 0, $y = 0$. We shall now prove the following:

Theorem 1. $C_T(n, \text{h}, \text{b}) = nS + x\bar{g}_\alpha + y\bar{g}_\beta$

where $\bar{g}_k = \min_{\lambda+\nu=\kappa} g_k(\lambda, \nu)$, and the λ_k^* value minimizing $g_k(\lambda, \nu)$, is the smallest integer satisfying the inequality:

$$\frac{b}{h+b}k \le \lambda_k^* \le 1 + \frac{b}{h+b}k$$

Proof. The proof is given in the Appendix •

Any policy, optimal or not, will have a number of cycles ranging from 1 up to T, i.e. a number belonging to the set N_T, where $N_T = \{1, 2, \ldots, T-1, T\}$. We shall now construct a partition of the set N_T in the following way. Let us consider the sets $B_i = \{n: \frac{T}{i+1} \le n < \frac{T}{i},\ n \in N\}$, where $i \in I = \{1, 2, \ldots, \left\lceil \frac{T}{2} \right\rceil, T-1\}$. Especially for B_1 we set $B_1 = \left\{n: \frac{T}{2} \le n \le \frac{T}{1},\ n \in N\right\}$, so as to include the value n = T. Obviously some B_i may be empty. These sets constitute a partition of N_T, which means that for any $n \in N_T$ there exists one and only one $i \in I$ such that $n \in B_i$. For any $n \in B_i$ we have $i < \frac{T}{n} \le i+1$. This inequality indicates that for any $i \in I$, the set B_i, if not empty, contains at most one n dividing T. If for any $n \in B_i$, $Tmod(n) \neq 0$, then obviously $\left\lceil \frac{T}{n} \right\rceil = i+1 = \alpha$, and $\left\lfloor \frac{T}{n} \right\rfloor = i = \beta$. If there exists one $n \in B_i$ such that Tmod(n) = 0, then for this n, $\frac{T}{n} = i+1 = \alpha$ and we set $\beta = i$. Therefore for any $n \in B_i$, $\alpha = i+1$ and $\beta = i$. The solution of (6) becomes:

$$x = T - ni,\ y = (i+1)n - T, \text{ for any } i \in I, \text{ and } n \in B_i \tag{7}$$

Due to the result of Theorem 1 and the above discussion, we have the following:

Theorem 2. The optimal cost for any policy in the set of policies P(n), $n \in B_i$, is given by

$$C_T(n, i, h, b) = nS + (T - ni)\bar{g}_{i+1} + [(i+1)n - T]\bar{g}_i, \ \forall\, n \in B_i \qquad (8)$$

where $\bar{g}_k = \min_{\lambda+\nu=K} g_k(\lambda, \nu)$, and λ_k^* is the minimizing value of $g_k(\lambda, \nu)$ •

This result is a first step towards having an analytical expression for the total cost function, with respect to the problem parameters, S, D, h and b. We have included i in the expression of the total cost function, in order to indicate its strong dependence from it.

Let us now examine the form of the optimal policy in the set of policies with *n* cycles. Due to the stationarity of the cost parameters we can take initially the *x* cycles of type *i+1*, and then the *y* cycles of length *i*. Consequently, the following theorem may be easily proved:

Theorem 3. For any $n \in B_i$, the optimal policy with *n* cycles is determined considering the following two cases:

Case 1: *Tmod(n) = 0*

The optimal policy consists of *n* cycles of length *i+1*, and is determined if

we order in any cycle *(i+1)*D

Case 2: *Tmod(n)* $\neq 0$

The optimal policy consists of x cycles of length i+1, y cycles of length i, and is determined if we

order in any cycle from the *x 's (i+1)*D, and

order in any cycle from the *y 's i*D •

This is a very useful and easily applicable result. If for any reason we are restricted to apply policies with only *n* cycles, the optimal one may be easily determined, by just finding $i \in I$, such that $n \in B_i$ and then substituting these *i* and *n* into (6) to determine *x*, the number of cycles of length *i+1*, and *y* the number of cycles of length *i*.

4. Total Cost Function Properties

An analytical expression for $C_T(n, i, h, b)$ with respect to the problem parameters S, D, h and b, requires similar expressions for the functions $\bar{g}_i$ and $\bar{g}_{i+1}$. The function $\bar{g}_i$ attains its minimum at $\lambda_i^* = \left\lceil \frac{b}{h+b} i \right\rceil$. This form of λ_i^* makes impossible the determination of an analytical expression for $\bar{g}_i$ as a function of h, b and i. This difficulty is overcomed by partitioning the (h,b) plane, into the sets $S_i(z) = \left\{ (h,b): z-1 \le \frac{b}{h+b} i \le z,\ z = 1, 2, \ldots, i \right\}$. An equivalent expression for these sets, servicing better the needs of this paper is:

$$S_i(z)=\{(h,b):\ (z-1)h\le(i-(z-1))b\ \wedge\ (i-z)b\le zh,\ z=1,\ 2,\ ...,\ i\}\ i\in I$$

It is easy to verify that these sets have the following properties:

a. they constitute a partition of the (h, b) plane w.r.t z

b. there is no inclusion relation between the sets $S_i(z)$ and $S_{i-1}(z)$

c. $S_i(z)\cap S_{i-1}(z)\neq\varnothing$ for $j=z-1,\ z+1$, and empty for all other j values

d. $S_i(z)\subset S_{i-1}(z)\cup S_{i-1}(z-1)$

In order to obtain a better understanding of the structure of the sets $S_i(z)$, a graphical illustration of the sets $S_5(z)$ for $z = 1,\ ...,\ 5$ is given in Figure 2.

It is obvious that for any (h, b) $\in S_i(z)$, $\lambda_i^*=z$. Therefore $\bar{g}_i$ becomes:

$$\bar{g}_i(z)=\frac{D}{2}\{z(z-1)h+(i-z)(i-z+1)b)\} \tag{9}$$

Based on the properties of the sets $S_i(z)$ and the expression in (8), we may obtain an analytical expression for $C_T(n, i, h, b)$. Obviously this requires finding which of the sets $S_{i+1}(z)\cap S_i(j)$ $j,\ z=1,\ ...,\ i$ are non empty. It can be easily proved that the only non empty sets are:

$$B_i(z,\ 1)=S_{i+1}(z)\cap S_i(z)=\{(h,b):\ (z-1)h\le(i+1-z)b\le zh,\ z=1,\ ...,\ i\},\ \text{and}$$

$$H_i(z,\ 1)=S_{i+1}(z)\cap S_i(z-1)=\{(h,b):\ (i+1-z)b\le(z-1)h\le(i+2-z)b,\ z=2,\ ...,\ i+1\}.$$

Taking $\bar{g}_i(z)$ from (9), and substituting into (8) we obtain the following analytic expression for the function $C_T(n, i, h, b)$:

$$C_T(n,\ i,\ h,\ b)=\begin{cases} nS+\dfrac{D}{2}nz(z-1)h+\dfrac{D}{2}\{2T-n(i+z)\}(i-z+1)b & (h,b)\in B_i(z,\ 1),\ z=1,...,i \\ nS+\dfrac{D}{2}(z-1)\{2T-n(2i-z+2)\}h+\dfrac{D}{2}n(i-z+1)(i-z+2)b & (h,b)\in H_i(z,\ 1),\ z=2,...,i+1 \end{cases} \tag{10}$$

To the best of our knowledge, we are not aware of any such analytical expression for the total cost function of this problem.

Now let us consider any set B_i containing at least two elements. Based on the expression (10), the optimal policy in the class of policies *P(n)*, for all $n\in B_i$, may be easily determined. In order to do this, we observe that $C_T(n, i, h, b)$ is a linear function with respect to *n*. Therefore, we only need to study the difference function $\Delta C_T = C_T(n) - C_T(n-1),\ \forall n,\ n-1\in B_i$. Using (10), we get the following analytic expression for the difference function ΔC_T:

$$\Delta C_T(n,\ i,\ h,\ b)=\begin{cases} S+\dfrac{D}{2}[z(z-1)h-(i+z)(i-z+1)b] & (h,b)\in B_i(z,\ 1),\ z=1,...,i \\ S-\dfrac{D}{2}[(z-1)(2i-z+2)h-(i-z+1)(i-z+2)b] & (h,b)\in H_i(z,\ 1),\ z=2,...,i+1 \end{cases} \tag{11}$$

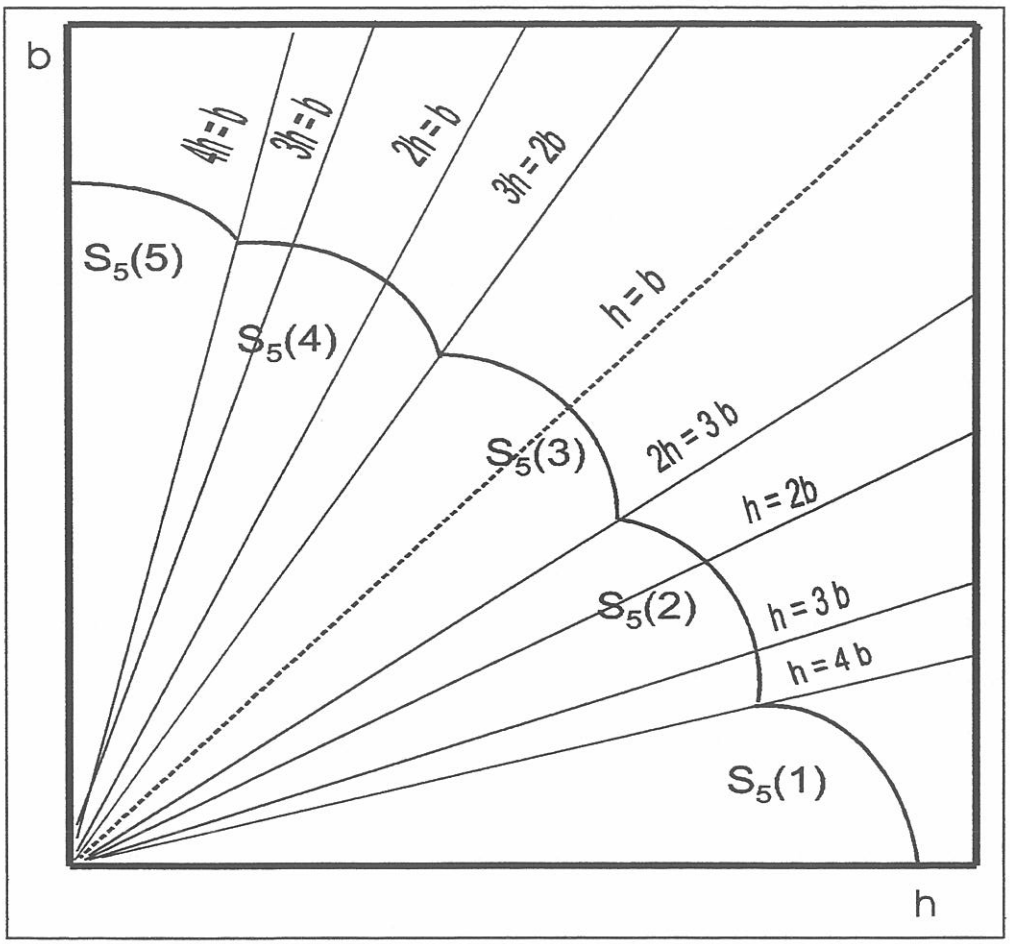


Figure 2. Graphical representation of the sets $S_5(z)$ for $z = 1, \ldots, 5$

Theorem 4. The optimal policy in the class of policies P(n), $n \in B_i$, for any set B_i containing at least two n values, is the one with n_{opt} cycles, where

- $n_{opt} = \min B_i$, for any set of cost parameters S, D, h, and b such that $\Delta C_T > 0$
- $n_{opt} = \max B_i$, for any set of cost parameters S, D, h, and b such that $\Delta C_T < 0$
- $n_{opt} \equiv B_i$, for any set of cost parameters S, D, h, and b such that $\Delta C_T = 0$

We shall now examine some special cases, where the overall n_{opt} value may be easily determined.

Let us set $f_1(z) = S + \frac{D}{2}[z(z-1)h-(i+z)(i-z+1)b]$, $\forall (h,b) \in B_i(z, 1)$. It is easy to verify that the difference of the function $f_1(z)$ w.r.t z is given by $\Delta f_1(z) = D(z-1)(h+b)$ $z = 2,\ldots,i$, $\forall$ h, b $\in B_i(z-1)$, which is positive and increasing $\forall z$. This proves that the function $f_1(z)$ is convex and increasing w.r.t z.

1st Case: For $z=2$, $f_1(2) = S + \frac{D}{2}[2h-(i+2)(i-1)b]$, $\forall(h,b) \in B_i(2, 1)$. If the values of the parameters D, S, h and b are such that $f_1(2) > 0$, then $f_1(z) > 0$ for all $z = 3,\ldots,i$ and $\Delta C_T > 0$. Therefore, $n_{opt} = \min B_i$ for all D, S, h and b satisfying $f_1(z) > 0$.

2nd Case: For $z = i$, $f_1(i) = S + \frac{D}{2} i[(i-1)h-2b]$, $\forall(h,b) \in B_i(i, 1)$. If $f_1(i) < 0$, then $f_1(z) < 0$ for all $z = 2,\ldots,i$ and $\Delta C_T < 0$. Therefore, $n_{opt} = \max B_i$.

3rd Case: If $f_1(2) < 0$ and $f_1(i) > 0$, then there exists some z_o such that $f_1(z_0) < 0$ and $f_1(z_0+1) \geq 0$. In this case $n_{opt} = \min B_i$ for all $z = 2,\ldots,z_0$ and $n_{opt} = \max B_i$ for all $z = z_0+1,\ldots,i$ $\forall(h,b) \in B_i(z, 1)$.

Similar results can be obtained if we examine the function ΔC_T defined on $H_i(z, 1)$.

4.1 Example

Analytical cost expressions for any policy with n setups, for the problem with T=12, are given in Table 1. Let us search for the optimal policy in the class of policies P(n), n $\in$ B_2, and let us suppose that $\frac{b}{2} \leq h \leq b$. Without loss of generality we assume that $D=1$. In this case $\Delta C_T = S + h - 2b$, and using theorem 4, we obtain

- $n_{opt} = 5$, for all S, h, and b values such that $S < 2b - h$, and
- $n_{opt} = 4$, for all S, h, and b values such that $S > 2b - h$

Therefore, if we are restricted to apply policies with $n=5$ cycles, since $12 \bmod(5) \neq 0$, this policy will consist of $x=2$ cycles of length $i+1=3$, and $y=3$ cycles of length $i=2$.

Using theorem 3, the optimal policy is determined if we:

order in any cycle from the x 's, $(i+1)D=3D$, and

order in any cycle from the y 's, $iD=2D$,

while λ_2^* and λ_3^* are calculated from the following relations

$$\frac{2b}{h+b} \leq \lambda_2^* \leq 1 + \frac{2b}{h+b} \Rightarrow 1 < \frac{2b}{h+b} (\text{since } \frac{b}{2} \leq h < b) \Rightarrow 1 < \lambda_2^* < 3 \Rightarrow \lambda_2^* = 2, \text{ and}$$

$$\frac{3b}{h+b} \leq \lambda_3^* \leq 1 + \frac{3b}{h+b} \Rightarrow 1 < \frac{2b}{h+b} (\text{since } \frac{b}{2} \leq h < b) \Rightarrow \frac{3}{2} < \frac{3b}{h+b} \Rightarrow \frac{3}{2} < \lambda_3^* < 3 \Rightarrow \lambda_3^* = 2$$

In case $h = b$ we obtain $\lambda_2^* = 1$ and $\lambda_2^* = 2$, and we have two equivalent policies. This always happens at the boundaries of the sets $S_i(z)$.

In this example the optimal policy allows for backlogging in the x cycles, while no backlogging is allowed in the y cycles.

The function $C_T(n, i, h, b)$ defined on the sets B_i has two branches designated by the (h, b) values. Furthermore, its form changes when moving from the set B_i to the closest non empty set B_{i+k}, $i < j < i+k$. This behavior adds difficulties to the analytical determination of $\Delta C_T(n, i, h, b)$ because it happens that $n \in B_i$ and $n-1 \in B_{i+k}$. To cope with this difficulty, we define, what we have called the *jump* of the function $C_T(n, i, h, b)$ between the sets B_i, B_{i+k} as follows [5]: let us take the sets $B_i \neq \varnothing$, $B_j = \varnothing$, $B_{i+k} \neq \varnothing$, $i < j < i+k$ and set $n_0 = \min B_i$ and $\nu_0 = \max B_i$. Obviously $n_0 - \nu_0 = 1$. For these sets, we define the function

$$J(B_i, B_{i+k}) = C_T(n_0, i, h, b) - C_T(\nu_0, i+k, h, b),$$

and we call it the *jump* of the function $C_T(n, i, h, b)$ between the sets B_i and B_{i+k}. To derive an analytical expression for $J(B_i, B_{i+k})$ we need to know the sets (h, b) on which it is defined. Based on the properties of the sets $S_i(z)$, it can be proved that, the only non empty sets where the jump of $C_T(n, i, h, b)$ is defined, are:

$$\begin{aligned}
&B_{i+k}(z, 1) \cap B_i(z-t, 1), \text{ for } t = 0, 1, \ldots, k-1 \\
&H_{i+k}(z, 1) \cap B_i(z-t, 1), \text{ for } t = 1, 2, \ldots, k \\
&B_{i+k}(z, 1) \cap H_i(z-t, 1), \text{ for } t = 0, 1, \ldots, k-1 \\
&H_{i+k}(z, 1) \cap H_i(z-t, 1), \text{ for } t = 1, 2, \ldots, k
\end{aligned} \tag{12}$$

More detailed analysis for these sets is given in the Appendix.

Substituting $C_T(n, i, h, b)$ from (10), we obtain the following analytical expression for the function $J(B_i, B_{i+k})$:

$$J(B_i, B_{i+k}) = S + \frac{D}{2} z(z-1)h + \frac{D}{2}\{-2kT + k(n_0-1)(2i+k+1) - (i+z)(i-z+1)\}b$$
$$\text{when } (h,b) \in B_{i+k}(z,1) \cap B_i(z,1)$$

$$J(B_i, B_{i+k}) = S + \frac{D}{2}[z(z-1) - n_0 t(2z-t-1)]h + \frac{D}{2}\{2T(t-k) - (i+z)(i-z+1) - n_0 t(2z-1-t) + (n_0-1)k(2i+k+1)\}b$$
$$\text{when } (h,b) \in B_{i+k}(z,1) \cap B_i(z-t,1) \quad t = 1, \ldots, k-1 \tag{13}$$

$$J(B_i, B_{i+k}) = S + \frac{D}{2}\{-2T(z-1) + n_0(z-1)(z-2) + (n_0-1)(z-1)[2(i+k+1) - z]\}h +$$
$$\frac{D}{2}\{2T(i-z+2) - (i-z+2)(2n_0 i - i + z - 1) - 2k(n_0-1)(i-z+1) - k(k+1)(n_0-1)\}b$$
$$\text{when } (h,b) \in H_{i+k}(z,1) \cap B_i(z-1,1)$$

$$J(B_i, B_{i+k}) = S + \frac{D}{2}\{(z-1)\{2[(n_0-1)(i+k+1) - T] + z\} - n_0 t(2z-t-1)\}h +$$
$$\frac{D}{2}\{(i-z+1)[2T - 2k(n_0-1) + (i-z+2)] + 2tT - 2n_0(i+1) + n(-2z+t+1) - (n_0-1)k(k+1)\}b$$
$$\text{when } (h,b) \in H_{i+k}(z,1) \cap B_i(z-t,1) \quad t = 2, \ldots, k \tag{14}$$

$$J(B_i,B_{i+k}) = S+\frac{D}{2}(z-1)\{2T-2n_0(i+1)+z\}h+$$
$$\frac{D}{2}\{-2T(i+k-z+1)+(n_0-1)(i+k+z)(i+k-z+1)+n_0(i-z+1)(i-z+2)\}b$$
$$\text{when } (h,b)\in B_{i+k}(z,1)\cap H_i(z,1)$$
$$J(B_i,B_{i+k}) = S+\frac{D}{2}\{2(z-1)[T-n_0(i+1)]-2tT+z(z-1)+n_0t[2(i+1-z)+1]\}h+$$
$$\frac{D}{2}\{-2T(i+k-z+1)+n_0(i-z+t+1)(i-z+t+2)+(n_0-1)(i+k+z)(i+k-z+1)\}b$$
$$\text{when } (h,b)\in B_{i+k}(z,1)\cap H_i(z-t,1)\quad t=1,\ \ldots\ ,k-1 \qquad (15)$$
$$J(B_i,B_{i+k}) = S+\frac{D}{2}\{-2T-(n_0-1)z(z-1)+2(n_0-1)(z-1)(i+k+1)-n_0(z-2)(2i-z+3)\}h+$$
$$\frac{D}{2}\{(i-z+1)(i-z+2)+2(i-z+2)[n_0-k(n_0-1)]-(n_0-1)k(k-1)\}b$$
$$\text{when } (h,b)\in H_{i+k}(z,1)\cap H_i(z-1,1)$$
$$J(B_i,B_{i+k}) = S+\frac{D}{2}\{(z-1)\{2T-n_0(2i-z+2)-2T+(n_0-1)[2(i+k)-z+2]\}-t[2T-n_0(2i-z+t+2)-tn_0(z-1)]\}h+$$
$$\frac{D}{2}\{(i-z+1)(i-z+2)+n_0t[2(i-z+1)+(t+1)]+k(n_0-1)[2(i-z+1)+(k+1)]\}b$$
$$\text{when } (h,b)\in H_{i+k}(z,1)\cap H_i(z-t,1)\quad t=2,\ \ldots\ ,k \qquad (16)$$

For k=1,

$$J(B_i,B_{i+1})=\begin{cases} S+\frac{D}{2}z(z-1)h+\frac{D}{2}\{-2T+2n_0(i+1)-(i+1+z)(i+2-z)\}b & (h,b)\in B_{i+1}(z,1)\cap B_i(z,1)\\ S+\frac{D}{2}(z-1)\{-2T+2n_0(i+1)-2i+z-4\}h+\frac{D}{2}(i-z+2)\{2T-2n_0(i+1)+i-z+3\}b & (h,b)\in H_{i+1}(z,1)\cap B_i(z-1,1)\\ S+\frac{D}{2}(z-1)\{2T-2n_0(i+1)+z\}h+\frac{D}{2}(i+2-z)\{-2T+2n_0(i+1)-(i+z+1)\}b & (h,b)\in B_{i+1}(z,1)\cap H_i(z,1)\\ S+\frac{D}{2}\{-2T+2n_0(i+1)-2i(z-1)+z(z-5)+4\}h+\frac{D}{2}(i-z+2)(i-z+3)b & (h,b)\in H_{i+1}(z,1)\cap H_i(z-1,1)\end{cases}$$

From the above results, we are now able to prove the following theorem:

Theorem 5. The function $C_T(n, i, h, b)$, $n\in N_T$, $i\in I$, is convex with respect to n.

Proof. The convexity of the function $C_T(n, i, h, b)$ will be established if we prove the following:

First $\Delta B_i > J(B_i,B_{i+k_1}) \qquad \forall\, i\in I \qquad (17)$

when B_i has at least two points, $B_{i+k_1}\neq\varnothing$, and $B_j=\varnothing$, $i<j<i+k_1$. This means that any difference is greater than the jump which may follows immediately.

Second $J(B_i,B_{i+k_1})\geq \Delta B_{i+k_1} \qquad \forall\, i\in I \qquad (18)$

when $B_i\neq\varnothing$, B_{i+k_1} has at least two points, and $B_j=\varnothing$, $i<j<i+k_1$. This means that any jump is greater than the difference which may follows immediately. This relation has meaning only if $k_1 = 1$, because as it is proved in [5], any non empty set B_j which follows an empty B_i set, $i < j$, has exactly one point.

Third $J(B_i, B_{i+k_1}) > J(B_{i+k_1}, B_{i+k_1+k_2}) \quad \forall\, i \in I$ (19)

when $B_i \neq \varnothing$, $B_{i+k_1}, B_{i+k_1+k_2}$ have exactly one point, and $B_j = \varnothing$, $i<j<i+k_1+k_2$, $j \neq i+k_1$. This means that any jump must be greater than the jump which may follows immediately. Obviously the above relations must be valid for any feasible combination of i, k_1, k_2.

The approach used to prove this theorem is similar to that used in [5]. The detailed proof is presented in the Appendix. •

Chand and Sethi [2], have also proved the convexity of the function $C_T(n, i, h, b)$ with respect to n, for the problem with varying demand and holding cost. However, the approach used in this paper, leads to the construction of an efficient and easily applicable algorithm to determine the optimal policy and its corresponding stability regions, for any given set of problem parameters, which is suitable for practical use.

5. Optimal Policy and Stability Regions

In this section, we present an algorithm to determine the optimal policy for the problem under consideration and its corresponding stability regions, i.e. the set of all problem parameters values for which an optimal solution remains valid [2, 6]. The convexity of the function $C_T(n, i, h, b)$ over N_T, guarantees the existence of a global minimum which may be determined by examining the sign of the differences $\Delta C_T(n, i, h, b)$ and the jumps $J(B_i, B_{i+k})$. In order to calculate the sign of these functions we only need to compute

$$S(B_i) = \Delta C_T(n, i, h, b) = \delta_i$$

$$S(B_i, B_{i+k}) = J(B_i, B_{i+k}) = j_{i+k}$$

To facilitate the presentation we call these functions, the "*sign functions*" ([5]). The proposed algorithm is proceeding as follows:

Step 0: Partition N_T into the sets B_i

Step 1: For any consecutive non empty sets B_i and B_{i+k}, $i < j < i+k$, calculate progressively starting from $i = 1$, their corresponding *sign functions* δ_i and j_{i+k}, using relations (11), (13), (14), (15) and (16).

Step 2: For any given set of problem parameters, check the *sign functions*, starting from δ_1 until you find the first one, if any exists, which is either negative or equal to zero. If such a *sign function* exists, this must be either δ_i or j_{i+k}. In this case

2.1 If $\delta_i \leq 0$ then $\quad n_{opt} = \begin{cases} \max B_i, & \text{if } \delta_i < 0 \\ B_i, & \text{if } \delta_i = 0 \end{cases}$

2.2 If $j_{i+k} < 0$ then $\quad n_{opt} = \min B_i$

2.3 If $j_{i+k} = 0$ then $n_{opt} = \{\max B_{i+k},\ \min B_i\}$, $k > 1$

2.4 If $j_{i+k} = 0$, $k = 1$, and B_{i+1} has at least two points, then due to the fact that in this case strict inequality may not hold in (18), which holds only if $k = 1$, we have to check if $\delta_{i+1} = 0$. If $\delta_{i+1} = 0$ then we have to include in the optimal policy, all integers contained in B_{i+1}. Therefore,

$$n_{opt} = \{B_{i+1},\ \min B_i\} \text{ if } \begin{cases} j_{i+1} = 0 \\ B_{j+1} \text{ has at least two points} \\ \delta_{i+1} = 0 \end{cases}$$

Step 3: If no *sign function* exists which is either negative or equal to zero, e.g. if $\delta_i > 0$ and $j_{i+k} > 0$, $\forall i \in I$ then $n_{opt} = 1$

In case that for any given set of problem parameters values there exists more than one optimal value, the symbol n_{opt} will be also used to represent the set of all these values. The optimal policy may be straightforward determined based on Theorem 3, using n_{opt} previously calculated using the proposed algorithm. The stability region for the n_{opt} may be determined by analyzing the sign functions inequalities with the problem parameters S, h, and b, which ensure that the value associated with the calculated solution is not higher than that for any other feasible solution [6].

5.1 Example

The sign functions for the problem with $T = 12$, are presented in Table 1. We now briefly explain how the proposed algorithm may be easily applied to this problem providing the optimal policy and stability regions for any set of problem parameters values.

Partitioning N_T into the sets B_i gives: $B_1 = \{12, 11, ..., 6\}$, $B_2 = \{5, 4\}$, $B_3 = \{3\}$, $B_4 = \varnothing$, $B_5 = \{2\}$, $B_6 = B_7 = B_8 = B_9 = B_{10} = \varnothing$, $B_{11} = \{1\}$, $B_{12} = \varnothing$. Using relations (11), (13), (14), (15) and (16) the sign functions may be easily calculated for all non empty sets B_i. If we assume without loss of generality, that $D = 1$, then for any set of S, h, b values with $b < h < 2b$

1. If $\delta_1 = S - b < 0$ then $n_{opt} = \max B_1 = \{12\}$
2. If $\delta_1 = S - b > 0, j_2 = S - 2h + b = \delta_2 < 0$ then $n_{opt} = \min B_1 = \{6\}$
3. If $j_2 = \delta_2 = 0$ then since the set B_2 has two integers, $n_{opt} = \{6, 5, 4\}$
4. If $\delta_1 = S - b > 0, j_2 = S - 2h + b = \delta_2 > 0$, and $j_3 = S + h - 5b < 0$, then

$$n_{opt} = \min B_2 = \{4\}$$

Using the same approach, the number of setups in the optimal policy may be easily determined for any set of problem parameters, using the sign functions calculated in Table 1. The set of problem parameters values, for which the previous relations hold, constitutes the stability region of the optimal policy.

The optimal policy may be determined as follows:

1. we consider initially the case where n_{opt} is an integer divisor of T. For any set of S, h, b values with $b < h < 2b$, $\delta_1 = S - b > 0$, $j_2 = S - 2h + b = \delta_2 > 0$ and $j_3 = S + h - 5b < 0$, $n_{opt} = 4$. Therefore, the optimal policy consists of *x=4* cycles of length *(i+1)=3* (*y=0*), and may be determined if we order in any cycle *(i+1)D=3D*.

$$\frac{3b}{h+b} \le \lambda_3^* \le 1 + \frac{3b}{h+b} \Rightarrow 1 < \frac{3b}{h+b} \text{ and } 2 < 1 + \frac{3b}{h+b} \text{ (since } b < h < 2b) \Rightarrow$$

$$1 < \lambda_3^* < 3 \Rightarrow \lambda_3^* = 2$$

Therefore, for any cycle in the optimal policy, backlogging is allowed in the first period.

2. for any set of S, h, b values with $b < h < 2b$ and $j_2 = \delta_2 = S - 2h + b = 0$, $n_{opt} = 5$. Therefore, using the results calculated in the example of section 4.1, the optimal policy will consist of *x=2* cycles of length *i+1=3*, and *y=3* cycles of length *i=2*, while it may be determined if we:

   order in any cycle from the *x 's*, *(i+1)D=3D*, and

   order in any cycle from the *y 's*, *iD=2D*

$$\frac{2b}{h+b} \le \lambda_2^* \le 1 + \frac{2b}{h+b} \Rightarrow \frac{2b}{h+b} < 1 \text{ and } 1 + \frac{2b}{h+b} > 1 \text{ (since } b < h < 2b) \Rightarrow \lambda_2^* = 1$$

$$\frac{3b}{h+b} \le \lambda_3^* \le 1 + \frac{3b}{h+b} \Rightarrow \frac{3b}{h+b} > 1 \text{ and } 1 + \frac{2b}{h+b} > 2 \text{ (since } b < h < 2b) \Rightarrow \lambda_3^* = 2$$

Therefore, for any cycle in the optimal policy backlogging is allowed only in the first period.

5. Conclusions

In this paper we have provided new insights into the problem of determining optimal solutions for the single product lot sizing problem with backlogging. The main contribution of the paper is the derivation of an analytical expression for the cost of the optimal policy as a function of all parameters in the model. This expression is utilized in proving the convexity of the total cost function in the number of cycles, constituting a completely different approach than that of Chand and Voros [2]. Another valuable contribution is the development of an efficient algorithm, analogous to that presented in [5], determining the optimal number of cycles, which enables the derivation of the optimal policy and the construction of stability regions for any set of problem parameters values.

References

[1] J. D. Blackburn, H. Kunreuther, Planning Horizons for the Dynamic Lot Size Model with Backlogging, Management Science 21 (3) (1974) 251 – 255.

[2] S. Chand, J. Voros, Setup Cost Stability Region for the Dynamic Lot Sizing Problem with Backlogging, European Journal of Operational Research 58 (1992) 68 – 77.

[3] A. Federgruen, M. Tzur, The Dynamic Lot Size Model with Backlogging: A Simple O(nlogn) Algorithm and Minimal Forecast Horizon Procedure, Naval Research Logistics 28 (1993) 213 - 228.

[4] T. E. Morton, An Improved Algorithm for the Stationary Cost Dynamic Lot Size Model with Backlogging, Management Science 24 (8) (1978) 869 - 873.

[5] S. Papachristos, I. Ganas, Optimal Policy and Stability Regions for the Single Product Periodic Review Inventory Product, with Stationary Demands, Journal of the Operational Research Society 49 (1998) 165 - 175.

[6] K. Richter, Sequential Stability of the Constant Cost Dynamic Lot Size Model, International Journal of Production Economics 35 (1994) 359 - 363.

[7] W. I. Zangwill, A Backlogging Model and a Multi-Echelon Model of a Dynamic Economic Lot Size Production System: A Network Approach, Management Science 13 (1) (1969) 105 - 119.

APPENDIX

Proof of Theorem 1.

The proof of this theorem will be established in two steps. The technique used here is similar to that used in Papachristos et al [5]. We shall distinguish the following cases:

Case 1: T mod (n) = 0, and $a = \frac{T}{n}$. In this case $y = an - T = 0$, $x = na - n(a-1) = n$. We must prove that $C_T(n, h, b) = nS + n\bar{g}_\alpha = n(S + \bar{g}_\alpha)$.

Let us set $f(m_i, \lambda_i, \nu_i) = \sum_{i=1}^{t} m_i[\lambda_i(\lambda_i - 1)h + \nu_i(\nu_i + 1)b]$. Now consider the problem

$$\min_{m_i, \lambda_i, \nu_i} f(m_i, \lambda_i, \nu_i) \tag{20}$$

$$\text{s.t.} \quad \sum_{i=1}^{t} m_i(\lambda_i + \nu_i) = km \tag{21}$$

$$\sum_{i=1}^{t} m_i = m \tag{22}$$

$$\lambda_i + \nu_i = \kappa_i$$

with k, m, t fixed positive integers, m_i integers (not necessarily positive) summing up to m, and λ_i, ν_i integer variables. In this problem $km \bmod(m) = 0$. To solve it, we shall first solve its relaxation, letting m_i, λ_i and ν_i to be real variables, using the Lagrange Multipliers method. If we set

$$F = f(m_i, \lambda_i, \nu_i) - w\left[\sum_{i=1}^{t} m_i(\lambda_i + \nu_i) - km\right] - z\left[\sum_{i=1}^{t} m_i - m\right] - \sum_{i=1}^{t} t_i(\lambda_i + \nu_i - \kappa_i),$$

take derivatives and equate them to zero we obtain the following equations:

$$\frac{\partial F}{\partial \lambda_i} = 2m_i\lambda_i h - m_i h - m_i w - t_i = 0 \tag{23}$$

$$\frac{\partial F}{\partial \nu_i} = 2m_i\nu_i b + m_i b - m_i w - t_i = 0 \tag{24}$$

$$\frac{\partial F}{\partial w} = \sum_{i=1}^{t} m_i(\lambda_i + \nu_i) - km = 0 \tag{25}$$

$$\frac{\partial F}{\partial z} = \sum_{i=1}^{t} m_i - m = 0 \tag{26}$$

$$\frac{\partial F}{\partial t_i} = \lambda_i + \nu_i - k_i = 0$$

$$\frac{\partial F}{\partial \kappa_i} = t_i = 0$$

If we multiply relations (23) and (24) by b and h respectively, we obtain

$$\begin{aligned} &2m_i\lambda_i hb - m_i hb - m_i wb = 0 \\ &2m_i\nu_i hb + m_i hb - m_i wh = 0 \end{aligned} \tag{27}$$

Summing up these two equations we obtain

$$2hbm_i(\lambda_i+\nu_i)-m_iw(h+b)=0 \tag{28}$$

From (28) taking sum over i we obtain

$$2hb\sum_{i=1}^{t}m_i(\lambda_i+\nu_i)-w(h+b)\sum_{i=1}^{t}m_i=0$$

Combing this relation with (21) and (22) we result with

$$2hbkm-w(h+b)m=0\Rightarrow w=\frac{2hb}{h+b}k \tag{29}$$

Replacing w into (28), we obtain the very basic relation $\lambda_i^\circ+\nu_i^\circ=k$, while replacing w into (23), and (24), we obtain the minimizing values for λ_i, ν_i:

$$\lambda_i^\circ=\frac{1}{2}+\frac{b}{h+b}k,\qquad \nu_i^\circ=-\frac{1}{2}+\frac{h}{h+b}k \tag{30}$$

From the above relation we have that the λ_i, ν_i giving the minimum in (20) are all equal, are independent from m_i and the most important their sum is equal to k. Relations (23), (24), (25), and (26) are valid for any m_i values (even negative) provided that their sum is equal to m. So, we can take $m_1 = m$, $m_i = 0$ for $i = 2,\ldots,t$. Therefore, the minimum in (20) is equal to

$$\sum_{i=1}^{t}m_i\left[\lambda_1^\circ(\lambda_1^\circ-1)h+\nu_1^\circ(\nu_1^\circ+1)b\right]=m\left[\lambda_1^\circ(\lambda_1^\circ-1)h+\nu_1^\circ(\nu_1^\circ+1)b\right] \tag{31}$$

Now consider the function $\sigma(\lambda,\nu)=\left[\lambda(\lambda-1)h+\nu(\nu+1)b\right]$, with λ, ν real variables, and $\lambda+\nu=\kappa$. Replacing $\nu=\kappa-\lambda$, we obtain $\sigma(\lambda)=(h+b)\lambda^2-[(h+b)+2kb]\lambda+k^2b+kb$. This function is strictly convex w.r.t λ, takes its minimum at λ_1°, and it is symmetric around λ_1°. Therefore, if we search for its minimum in the set of integer λ values, we can easily conclude that the minimizing point is the integer λ^* closest to λ_1°. It is easy to see that this is the integer contained in the interval

$$\frac{b}{h+b}k\le\lambda^*\le 1+\frac{b}{h+b}k \tag{32}$$

If $\frac{b}{h+b}k$ is an integer then there are two λ^* values giving the same minimum, but to avoid confusion, we shall always keep the smallest one.

The above discussion proves that $\min_{m_i,\lambda_i,\nu_i} f(m_i,\lambda_i,\nu_i)=m\bar{g}_k$, and so Theorem 1 is true in case that n divides T.

Case 2. Tmod(n) $\neq 0$, and $\alpha=\left\lceil\frac{T}{n}\right\rceil$, $\beta=\left\lfloor\frac{T}{n}\right\rfloor=\alpha-1$, $x=T-n\beta$, $y=\alpha n-T$.

In this case we claim that

$$M=\min_{m_i,\lambda_i,\nu_i} f(m_i,\lambda_i,\nu_i)=x\bar{g}_\alpha+y\bar{g}_\beta$$

$$\text{s.t.} \quad \sum_{i=1}^{t} m_i(\lambda_i + \nu_i) = T$$

$$\sum_{i=1}^{t} m_i = n \tag{33}$$

$$\lambda_i + \nu_i = \kappa_i$$

and m_i, λ_i, ν_i integers

First we shall prove that $x\bar{g}_\alpha + y\bar{g}_\beta$ is a lower bound for M, i.e. we shall prove that $M \geq x\bar{g}_\alpha + y\bar{g}_\beta$. If this is not true for all m_i, λ_i, ν_i satisfying (33), then there will exist at least one set of $\bar{m}_i, \bar{\lambda}_i, \bar{\nu}_i$ values such that

$$f(\bar{m}_i, \bar{\lambda}_i, \bar{\nu}_i) = \sum_{i=1}^{t} \bar{m}_i \left[\bar{\lambda}_i(\bar{\lambda}_i - 1)h + \bar{\nu}_i(\bar{\nu}_i + 1)b \right] < x\bar{g}_\alpha + y\bar{g}_\beta \Rightarrow f(\bar{m}_i, \bar{\lambda}_i, \bar{\nu}_i) - y\bar{g}_\beta < x\bar{g}_\alpha \tag{34}$$

with $\sum_{i=1}^{t} \bar{m}_i(\bar{\lambda}_i + \bar{\nu}_i) = T$, $\sum_{i=1}^{t} \bar{m}_i = n$, and $\bar{m}_i, \bar{\lambda}_i, \bar{\nu}_i$ integers.

Now let us consider the auxiliary problem

$$N = \min_{\lambda_i,\, \nu_i,\, \lambda_{t+1},\, \nu_{t+1}} f(m_i,\ \lambda_i,\ \nu_i) + m_{t+1}[\lambda_{t+1}(\lambda_{t+1} - 1)h + \nu_{t+1}(\nu_{t+1} + 1)b)] \tag{35}$$

$$\text{s.t.} \quad \sum_{i=1}^{t} m_i(\lambda_i + \nu_i) + m_{t+1}(\lambda_{t+1} + \nu_{t+1}) = x(\beta + 1) = \bar{T} \tag{36}$$

$$\sum_{i=1}^{t} m_i + m_{t+1} = x = \bar{n},$$

and m_{t+1} any integer (not necessarily positive)

In this case we have that $\bar{n} = x$, divides $\bar{T} = x(\beta + 1)$ and according to Case 1 the minimum is equal to $\bar{n}\,\bar{g}_{\beta+1} = x\bar{g}_\alpha$. This minimum is achieved at $\lambda_i^\circ = \lambda^*, \nu_i^\circ = (\beta + 1) - \lambda^*$, $i = 1, \ldots, t+1$ where λ^* is obtained from (32) for *k=(β+1)*, and is valid for any m_i, m_{t+1} satisfying the second condition of (36). So we can take $m_{t+1} \in Z - \{0\}$ without affecting the minimum.

If we now take $\bar{m}_{t+1} = -y$ and $\bar{\lambda}_{t+1} + \bar{\nu}_{t+1} = \beta$, with $\bar{\lambda}_{t+1}$, $\bar{\nu}_{t+1}$ chosen so that $g(\bar{\lambda}_{t+1}, \bar{\nu}_{t+1}) = \bar{g}_\beta$ then we have

$$\sum_{i=1}^{t} \bar{m}_i(\bar{\lambda}_i + \bar{\nu}_i) - y\beta = T - y\beta = T - (n - x)\beta = T - n\beta + x\beta = x + x\beta = x(\beta + 1)$$

$$\sum_{i=1}^{t} \bar{m}_i - y = n - y = x$$

So, the numbers, $\bar{m}_i, \bar{\lambda}_i, \bar{\nu}_i$, i = 1, ..., t, and $\bar{m}_{t+1}, \bar{\lambda}_{t+1}, \bar{\nu}_{t+1}$ with $\bar{m}_{t+1} = -y$ and $\bar{\lambda}_{t+1} + \bar{\nu}_{t+1} = \beta$ satisfy constraints (36) of the auxiliary problem. If we substitute them into the

objective function (35) we obtain $f(\bar{m}_i,\bar{\lambda}_i,\bar{\nu}_i)-y\,\bar{g}_\beta \geq N = x\,\bar{g}_\alpha$. Combining this and (34), we obtain $x\,\bar{g}_\alpha \geq f(\bar{m}_i,\bar{\lambda}_i,\bar{\nu}_i)-y\,\bar{g}_\beta > x\,\bar{g}_\alpha$ which is a contradiction. Therefore, $x\,\bar{g}_\alpha + y\,\bar{g}_\beta$ is a lower bound for M.

Now, we shall see that are values of λ_i, ν_i, m_i which satisfy (33) while the corresponding value of $f(m_i, \lambda_i, \nu_i)$ is equal to the lower bound. This may be achieved as follows: take any from the m_i and give them values summing up to x. For all these i indexes set $\lambda_i + \nu_i = \alpha$. Then, to all these λ_i's assign the value λ^* resulting from (32) for $k = \alpha$, and then set $\nu_i = \nu^* = \alpha - \lambda^*$. This part of $f(m_i, \lambda_i, \nu_i)$ will take the value $x\,\bar{g}_\alpha$. Do the same for the remaining m_i, i.e. give them values summing up to $n - x = y$. For all these i indexes set $\lambda_i + \nu_i = \beta$. Then, to all these λ_i's assign the value ν^* resulting from (32) for $k = \beta$, and then set $\nu_i = \nu^* = \beta - \lambda^*$. This part of $f(m_i, \lambda_i, \nu_i)$ will take the value $y\,\bar{g}_\beta$.

The above discussion proves Theorem 1 in case that n is not an integer divisor of T.

Proof of Theorem 5

The first condition will be established if we prove the following:

A1. $S + \frac{D}{2}[z(z-1)h-(i+z)(i-z+1)b] > S+\frac{D}{2}z(z-1)h+$

$\frac{D}{2}\{-2k_1T+k_1(n_0-1)(2i+k_1+1)-(i+z)(i-z+1)\}b \quad (h,b)\in B_{i+k_1}(z,1)\cap B_i(z,1) \Rightarrow$

$k_1\{-2T+(n_0-1)(2i+k_1+1)\}<0 \Rightarrow (n_0-1)(2i+k_1+1)-2T<0$

This is valid because since $n_0 = \min B_i$, $(n_0-1) < \frac{T}{i+1} \Rightarrow (n_0-1)(i+1) < T$,

and $(n_0-1)\in B_{i+k_1} \Rightarrow (n_0-1) < \frac{T}{i+k_1} \Rightarrow (n_0-1)(i+k_1) < T$ •

A2. $S + \frac{D}{2}[(z-1)(z-2)h-(i+z-1)(i-z+2)b] >$

$S+\frac{D}{2}\{-2T(z-1)+n_0(z-1)(z-2)+(n_0-1)(z-1)[2(i+k_1+1)-z]\}h+$

$\frac{D}{2}\{2T(i-z+2)-(i-z+2)[2n_0i-(i-z+1)+k_1(n_0-1)]-k_1(n_0-1)[(i-z+1)+k_1]\}b$

$(h,b)\in H_{i+k_1}(z,1)\cap B_i(z-1,1) \Rightarrow$

$(z-1)\{-2T+(n_0-1)(z-2)+2(n_0-1)(i+k_1+1)-(n_0-1)z\}h+$

$\{2T(i-z+2)-(i-z+2)[2n_0i-(i-z+1)-(i+z-1)+k_1(n_0-1)]-k_1(n_0-1)[(i-z+1)+k_1]\}b<0 \Rightarrow$

$$(z-1)\{-2T+(n_0-1)(2i+k_1)+2(n_0-1)(2i+k_1+1)-2(n_0-1)\}h+$$

$$\{2(i-z+2)[T-(n-1)(i+k_1)]+k_1(n_0-1)(i-z+2)-k_1(n_0-1)(i-z+1)-k_1^2(n_0-1)\}b<0\Rightarrow$$

$$2(z-1)[-T+(n_0-1)(i+k_1)]h+\{2(i-z+2)[T-(n-1)(i+k_1)]-k_1(k_1-1)(n_0-1)\}b<0\Rightarrow$$

$$[(i-z+2)b-(z-1)h][T-(n-1)(i+k_1)]-k_1(k_1-1)(n_0-1)b\le 0$$

which holds since $n_0=\min B_i$, $(n_0-1)\in B_{i+k_1}\Rightarrow(n_0-1)<\dfrac{T}{i+k_1}\Rightarrow(n_0-1)(i+k_1)<T$, and

$(i-z+2)b\le(z-1)h$, since $(h,b)\in H_{i+k_1}(z,1)\cap B_i(z-1,1)$ •

A3. $S-\dfrac{D}{2}[(z-1)(2i-z+2)h-(i-z+1)(i-z+2)b]>S+\dfrac{D}{2}(z-1)\{2T-2n_0(i+1)+z\}h+$

$$\frac{D}{2}\{-2T(i+k_1-z+1)+(n_0-1)(i+k_1+z)(i+k_1-z+1)+n_0(i-z+1)(i-z+2)\}b$$

$(h,b)\in B_{i+k_1}(z,1)\cap H_i(z,1)\Rightarrow$

$$(z-1)\{2T-2n(i+1)+z+2(i+1)-z\}h+$$

$$\{-2T(i+k_1-z+1)+(n_0-1)(i+k_1+z)(i+k_1-z+1)+n_0(i-z+1)(i-z+2)-(i-z+1)(i-z+2)\}b<0\Rightarrow$$

$$(z-1)\{2T-2(i+1)(n_0-1)\}h+$$

$$\{-2T(i+k_1-z+1)+(n_0-1)[(i-z+1)+k_1][i+(k_1+z)]+(n_0-1)(i-z+1)[(i+2)-z]\}b<0\Rightarrow$$

$$(z-1)\{2T-2(i+1)(n_0-1)\}h+\{-2T(i+k_1-z+1)+i(n_0-1)(i-z+1)+(i+2)(n_0-1)(i-z+1)$$

$$-z(n_0-1)(i-z+1)+(n_0-1)[ik_1+(k_1+z)(i-z+1)+k_1(k_1+z)\}b<0\Rightarrow$$

$$(z-1)\{2T-2(i+1)(n_0-1)\}h+b\{[-2T+2(i+1)(n_0-1)](i-z+1)-k_12T+k_1(n_0-1)(i-z+1)+$$

$$+k_1(n_0-1)(i+k_1+z)\}b<0\Rightarrow$$

$$[h(z-1)-b(i-z+1)][2T-2(i+1)(n_0-1)]+bk_1[-2T+(n_0-1)(2i+k_1+1)]<0\Rightarrow$$

$$[h(z-1)-b(i-z+k_1+1)][2T-2(i+1)(n_0-1)]-bk_1(n_0-1)(k_1-1)\le 0$$

This is valid because since

$$n_0=\min B_i,\ (n_0-1)<\frac{T}{i+1}\Rightarrow(n_0-1)(i+1)<T\Rightarrow 2T-2(n_0-1)(i+1)>0,$$

$h(z-1)-b(i-z+k_1+1)<0$, since $(h,b)\in B_{i+k}(z,1)\cap H_i(z,1)$,

while $-bk_1(n_0-1)(k_1-1)$ is obviously negative •

A4. In case that $(h,b)\in H_{i+k_1}(z,1)\cap H_i(z-1,1)$,

$$S-\frac{D}{2}[(z-2)(2i-z+3)h-(i-z+2)(i-z+3)b]>$$

$$S+\frac{D}{2}\{-2T-(n_0-1)z(z-1)+2(n_0-1)(z-1)(i+k_1+1)-n_0(z-2)(2i-z+3)\}h+$$

$$\frac{D}{2}\{(i-z+1)(i-z+2)+2[n_0-k_1(n_0-1)](i-z+2)-(n_0-1)k_1(k_1-1)\}b\ \Rightarrow$$

$$S+\frac{D}{2}\{-2T-(n_0-1)z(z-1)+2(n_0-1)(z-1)(i+k_1+1)-n_0(z-2)(2i-z+3)\}h+$$

$$\frac{D}{2}\{(i-z+1)(i-z+2)+2[n_0-k_1(n_0-1)](i-z+2)-(n_0-1)k_1(k_1-1)\}b<0\ \Rightarrow$$

$$\{-2T-(n_0-1)z(z-1)+2(n_0-1)(z-1)(i+k_1+1)-(n_0-1)(z-2)(2i-z+3)\}h+$$

$$\{(i-z+1)(i-z+2)-(i-z+2)(i-z+3)+2[n_0-k_1(n_0-1)](i-z+2)-(n_0-1)k_1(k_1-1)\}<0\Rightarrow$$

$$\{-2T-(n_0-1)z(z-1)+2(n_0-1)(z-1)(i+k_1+1)-(n_0-1)(z-2)(2i-z+3)\}h+$$

$$\{-2(i-z+2)+2n_0(i-z+2)-2k_1(n_0-1)(i-z+2)-(n_0-1)k_1(k_1-1)\}<0\Rightarrow$$

$$\{-2T-(n_0-1)[z(z-1)+(z-2)[2(i+1)-(z-1)]]+2(n_0-1)(z-1)(i+k_1+1)\}h+$$

$$\{2(n_0-1)(i-z+2)(1-k_1)-(n_0-1)k_1(k_1-1)\}b<0\Rightarrow$$

$$\{-2T+2(n_0-1)(z-1)(i+k_1)-2(n_0-1)(z-2)(i+1)\}h+\{2(n_0-1)(i-z+2)(1-k_1)-(n_0-1)k_1(k_1-1)\}b<0\Rightarrow$$

$$\{-2T+2(n_0-1)[(z-1)(i+k_1)-(z-2)(i+1)]\}h+\{2(n_0-1)(i-z+2)(1-k_1)-(n_0-1)k_1(k_1-1)\}b<0\Rightarrow$$

$$[-2T+2(n_0-1)(i+1)+2(n_0-1)(z-1)(k_1-1)]h-\{2(n_0-1)(i-z+2)(k_1-1)+(n_0-1)k_1(k_1-1)\}b<0\Rightarrow$$

$$[-2T+2(n_0-1)(i+1)]h+2(n_0-1)(k_1-1)[(z-1)h-(i-z+2)b]-b(n_0-1)k_1(k_1-1)\le 0$$

This is valid because since $n_0=\min B_i$, $(n_0-1)<\dfrac{T}{i+1}\Rightarrow (n_0-1)(i+1)-T<0$ and

$h(z-1)-b(i-z+2)<0$, since $(h,b)\in H_{i+k_1}(z,1)\cap H_i(z-1,1)$ •

The second condition will be established if we prove the following:

B1.

$$S+\frac{D}{2}z(z-1)h+\frac{D}{2}\{-2T+2(n_0-1)(i+1)-(i+z)(i-z+1)\}b\ge S+\frac{D}{2}[z(z-1)h-(i+1+z)(i+2-z)b]$$

$$(h,b)\in B_{i+1}(z,1)\cap B_i(z,1)\Rightarrow$$

$$-2T+2(n_0-1)(i+1)-(i+z)(i-z+1)+(i+1+z)(i+2-z)\ge 0\Rightarrow$$

$$-2T+2(n_0-1)(i+1)-(i+z)(i-z+1)+(i+z)(i-z+1)+2(i+1)\ge 0\Rightarrow -2T+2n_0(i+1)\ge 0$$

This is valid because since $n_0=\min B_i$, $\dfrac{T}{i+1}\le n_0\Rightarrow n_0(i+1)-T\ge 0$ •

B2.

$$S+\frac{D}{2}(z-1)\{-2T+n_0(z-2)+2(i+2)(n_0-1)-z(n_0-1)\}h+$$

$$\frac{D}{2}\{2T(i-z+2)-(i-z+2)(2n_0i-i+z-1)-2(n_0-1)(i-z+2)\}b\ge$$

$$S-\frac{D}{2}\{(z-1)[2(i+1)-z+2]h-(i-z+2)(i-z+3)b\}\qquad (h,b)\in H_{i+1}(z,1)\cap B_i(z-1,1)\Rightarrow$$

$$(z-1)\{-2T+n_0(z-2)+2(i+2)(n_0-1)-z(n_0-1)+2(i+1)-z+2\}h+$$

$$\{2T(i-z+2)-(i-z+2)[2n_0i-(i-z+1)+2(n_0-1)-(i-z+3)\}b\ge 0\Rightarrow$$

$$(z-1)\{-2T+2n_0(i+1)\}h+\{2T(i-z+2)-(i-z+2)[2n_0i+2(n_0-1)-2(i-z+2)\}b\ge 0\Rightarrow$$

$$(z-1)\{-2T+2n_0(i+1)\}h+\{[2T-2n_0(i+1)](i-z+2)+2(i-z+2)(i-z+3)\}b\ge 0\Rightarrow$$

$$(z-1)[-2T+2n_0(i+1)]h-[-2T+2n_0(i+1)](i-z+2)b+2(i-z+2)(i-z+3)b\ge 0\Rightarrow$$

$[(z-1)h-(i-z+2)b][-2T+2n_0(i+1)]+2(i-z+2)(i-z+3)b \geq 0$

This is valid because since $n_0 = \min B_i$, $\frac{T}{i+1} \leq n_0 \Rightarrow 2n_0(i+1)-2T \geq 0$, and

$(h,b) \in H_{i+1}(z,1) \cap B_i(z-1,1)$, $(z-1)h-(i-z+2)b \geq 0$ •

B3.

$$S+\frac{D}{2}(z-1)\{2T-2n_0(i+1)+z\}h+\frac{D}{2}(i-z+2)\{-2T+2n_0(i+1)-(i+z+1)\}b \geq$$

$$S+\frac{D}{2}\{z(z-1)h-(i+z+1)(i-z+2)b\} \qquad (h,b) \in B_{i+1}(z,1) \cap H_i(z,1) \Rightarrow$$

$(z-1)\{2T-2n_0(i+1)\}h-(i-z+2)\{2T-2n_0(i+1)\} \geq 0 \Rightarrow \{(z-1)h-(i-z+2)b\}\{2T-2n_0(i+1)\} \geq 0$

This is valid because since $n_0 = \min B_i$, $\frac{T}{i+1} \leq n_0 \Rightarrow 2T-2n_0(i+1) \leq 0$,

and since $(h,b) \in B_{i+1}(z,1) \cap H_i(z,1)$, $(z-1)h-(i-z+2)b \leq 0$ •

B4.

$$S+\frac{D}{2}\{-2T-(n_0-1)z(z-1)+2(n_0-1)(z-1)(i+2)-n_0(z-2)(2i-z+3)\}h+$$

$$\frac{D}{2}\{(i-z+1)(i-z+2)+2(i-z+2)\}b \geq S-\frac{D}{2}\{(z-1)[2(i+1)-z+2]h-(i-z+2)(i-z+3)b\}$$

$$(h,b) \in H_{i+1}(z,1) \cap H_i(z-1,1) \Rightarrow$$

$\{-2T-n_0z(z-1)+2n_0(z-1)(i+2)-n_0(z-2)(2i-z+3)\}h \geq 0 \Rightarrow$

$\{-2T-n_0z(z-1)+2n_0(z-1)(i+2)-2n_0(z-2)(i+1)+n_0(z-1)(z-2)\}h \geq 0 \Rightarrow$

$\{-2T+2n_0(z-1)(i+1)-2n_0(z-2)(i+1)\}h \geq 0 \Rightarrow$

$\{-2T+2n_0(i+1)[(z-1)-(z-2)]\}h \geq 0 \Rightarrow [-2T+2n_0(i+1)]h \geq 0$

This is valid because since $n_0 = \min B_i$, $\frac{T}{i+1} \leq n_0 \Rightarrow 2n_0(i+1)-2T \geq 0$ •

Before proving the third condition, it can be easily shown using relations (11) through (14) that the function $J(B_{i+k_1}, B_{i+k_1+k_2})$ is defined as follows:

$$J(B_{i+k_1}, B_{i+k_1+k_2}) = S+\frac{D}{2}z(z-1)h+\frac{D}{2}\{-2k_2T+k_2(n_0-1)[2(i+k_1)+(k_2+1)]-(i+k_1+z)(i+k_1-z+1)\}b$$

$$\text{when } (h,b) \in B_{i+k_1+k_2}(z,1) \cap B_{i+k_1}(z,1)$$

$$J(B_{i+k_1}, B_{i+k_1+k_2}) = S+\frac{D}{2}(z-1)\{-2T+2(n_0-1)(i+k_1+k_2+1)-2n_0+z\}h+$$

$$\frac{D}{2}\{(-2k_2T+[2T-2n_0(i+k_1)+(i+k_1+k_2-z+1)])(i+k_1+k_2-z+2)-n_0k_2(k_2-2z+3)\}b$$

$$\text{when } (h,b) \in H_{i+k_1+k_2}(z,1) \cap B_{i+k_1}(z-1,1)$$

$$J(B_{i+k_1}, B_{i+k_1+k_2}) = S + \frac{D}{2}(z-1)\{2T - 2n_0(i+k_1+1)+z\}h +$$

$$\frac{D}{2}\{[-2T + 2n_0(i+k_1) + n_0 - (i+k_1+k_2+z)](i+k_1+k_2-z+1) + n_0k_2(k_2-1)\}b$$

$$\text{when } (h,b) \in B_{i+k_1+k_2}(z,1) \cap H_{i+k_1}(z,1)$$

$$J(B_{i+k_1}, B_{i+k_1+k_2}) = S + \frac{D}{2}\{-2T - n_0(z-1)(z-2) - 2n_0(z-2)(i+k_1+1) + 2(n_0-1)(z-1)(i+k_1+k_2+1) - z(z-1)(n_0-1)\}h +$$

$$\frac{D}{2}\{(i+k_1+k_2-z+1)(i+k_1+k_2-z+2) - n_0k_2(k_2-1)\}b$$

$$\text{when } (h,b) \in H_{i+k_1+k_2}(z,1) \cap H_{i+k_1}(z-1,1)$$

Therefore, the third condition will be established if we prove the following:

C1.

$$S + \frac{D}{2}z(z-1)h + \frac{D}{2}\{-2k_1T + k_1(n_0-1)(2i+k_1+1) - (i+z)(i-z+1)\}b \geq$$

$$S + \frac{D}{2}z(z-1)h + \frac{D}{2}\{-2k_2T + k_2(n_0-1)[2(i+k_1) + (k_2+1)] - (i+k_1+z)(i+k_1-z+1)\}b$$

$$\text{when } (h,b) \in \left(B_{i+k_1}(z,1) \cap B_i(z,1)\right) \cap \left(B_{i+k_1+k_2}(z,1) \cap B_{i+k_1}(z,1)\right) \equiv B_{i+k_1+k_2}(z,1) \cap B_{i+k_1}(z,1) \cap B_i(z,1) \Rightarrow$$

$$2(k_2-k_1)T + k_1(n_0-1)(2i+k_1+1) - k_2(n_0-1)(k_2+1) - 2k_2(n_0-1)(i+k_1) + k_1(i+z) + k_1(i-z+1) + k_1^2 \geq 0 \Rightarrow$$

$$2(k_2-k_1)T + k_1n_0(2i+k_1+1) - k_2(n_0-1)(k_2+1) - 2k_2(n_0-1)(i+k_1) + (2ik_1 + k_1 + k_1^2) \geq 0 \Rightarrow$$

(for a formal proof see [5])

C2.

$$S + \frac{D}{2}\{-2T(z-1) + n_0(z-1)(z-2) + (n_0-1)(z-1)[2(i+k_1+1) - z]\}h +$$

$$\frac{D}{2}\{2T(i-z+2) - (i-z+2)(2n_0i - i + z - 1) - 2k_1(n_0-1)(i-z+1) - k_1(k_1+1)(n_0-1)\}b \geq$$

$$S + \frac{D}{2}(z-1)\{-2T + 2(n_0-1)(i+k_1+k_2) + z - 2]\}h +$$

$$\frac{D}{2}\{(-2k_2T + [2T - 2n_0(i+k_1) + (i+k_1+k_2-z+1)])(i+k_1+k_2-z+2) - n_0k_2(k_2-2z+3)\}b$$

$$\text{when } (h,b) \in \left(H_{i+k_1}(z,1) \cap B_i(z-1,1)\right) \cap \left(H_{i+k_1+k_2}(z,1) \cap B_{i+k_1}(z-1,1)\right)$$

$$A = h(z-1)\{2(n_0-1)[(i+k_1+1) - (i+k_1+k_2)] + n_0(z-2) - z(n_0-1) - (z-2)\} =$$

$$h(z-1)\{(n_0-1)(z-2) - z(n_0-1) - 2(n_0-1)(k_2-1)\} = -2h(z-1)(n_0-1)k_2$$

$$B = \{2T[(i-z+2) + (k_2-1)(i+k_1+k_2-z+1)] - n_0[2i(i-z+2) - 2(i+k_1)(i+k_1+k_2-z+2)] +$$

$$(i-z+1)(i-z+2) - (i+k_1+k_2-z+1)(i+k_1+k_2-z+2) - k_1(n_0-1)[(i-z+2) + (i-z+1) + k_1] + n_0k_2(k_2-2z+3)\}b =$$

$$\{2T[1 + (k_2-1)(k_1+k_2) + k_2(i-z+1)] + 2n_0[(i+k_1)(k_1+k_2) + k_1(i-z+2)] - (k_1+k_2)[2(i-z+1) + k_1 + k_2 + 1] -$$

$$k_1(n_0-1)[2(i-z+1) + k_1 + 1] + n_0k_2(k_2-2z+3)\}b =$$

$$\{2T[1 + (k_2-1)(k_1+k_2) + k_2(i-z+1)] + n_0[2(i+k_1)(k_1+k_2) + 2k_1(i-z+2) - 2k_1(i-z+1) - k_1(k_1+1) + k_2(k_2+1)$$

$$-2k_2(z-1)] - (k_1+k_2)[2(i-z+1) + k_1 + k_2 + 1] - k_1[2(i-z+1) + k_1 + 1]\}b =$$

$$\{2T[1 + (k_2-1)(k_1+k_2) + k_2(i-z+1)] + n_0[2i(k_1+k_2) + 2k_1(k_1+k_2) + 2k_1 - k_1(k_1+1) + k_2(k_2+1) - 2k_2(z-1)] -$$

$$k_2[2(i-z+1) + k_1 + k_2 + 1] + k_1[2(i-z+1) + k_1 + 1 - 2(i-z+1) - k_1 - k_2 - 1]\}b =$$

$$\{2T[1 + (k_2-1)(k_1+k_2) + k_2(i-z+1)] + n_0[2i(k_1+k_2) + k_2(k_2+1) - 2k_2z + 2(k_1+k_2) + 2k_1(k_1+k_2) -$$

$$k_1(k_1+1)] - k_2[2(i+k_1-z+1) + k_2 + 1]\}b =$$

$$\{2T[1+(k_2-1)(k_1+k_2)+k_2(i-z+1)]+(n_0-1)[k_2(k_2+1)-2k_2z+2k_1k_2+2ik_2+2k_2+nk_1(2i+k_1+1)\}b=$$
$$\{2T[1+(k_2-1)(k_1+k_2)+k_2(i-z+1)]+(n_0-1)k_2[2(i+k_1+2-z)+k_2-1\}b=$$
$$\{2T[1+(k_2-1)(k_1+k_2)+k_2(i-z+1)]+(n_0-1)k_2(k_2-1)+2(n_0-1)k_2(i+k_1+2-z)\}b$$

Therefore, $A+B=2(n_0-1)k_2\{(i+k_1+2-z)b-h(z-1)\}+$
$$\{2T[1+(k_2-1)(k_1+k_2)+k_2(i-z+1)]+(n_0-1)k_2(k_2-1)\}b\geq 0$$
since $(h,b)\in\left(H_{i+k_1}(z,1)\cap B_i(z-1,1)\right)\cap\left(H_{i+k_1+k_2}(z,1)\cap B_{i+k_1}(z-1,1)\right)$ •

C3.

$$S+\frac{D}{2}(z-1)\{2T-2n_0(i+1)+z\}h+$$
$$\frac{D}{2}\{-2T(i+k_1-z+1)+(n_0-1)(i+k_1+z)(i+k_1-z+1)+n_0(i-z+1)(i-z+2)\}b\geq$$
$$S+\frac{D}{2}(z-1)\{2T-2n_0(i+k_1+1)+z\}h+$$
$$\frac{D}{2}\{[-2T+2n_0(i+k_1)+n_0-(i+k_1+k_2+z)](i+k_1+k_2-z+1)+n_0k_2(k_2-1)\}b$$
when $(h,b)\in\left(B_{i+k_1}(z,1)\cap H_i(z,1)\right)\cap\left(B_{i+k_1+k_2}(z,1)\cap H_{i+k_1}(z,1)\right)$

$2n_0k_1(z-1)h+\{-2T[(i+k_1-z+1)-(i+k_1+k_2-z+1)]-n_0[2(i+k_1)+1](i+k_1+k_2-z+1)+$
$(i+k_1+k_2+z)(i+k_1+k_2-z+1)+(n_0-1)(i+k_1+z)(i+k_1-z+1)+n_0(i-z+1)(i-z+2)-n_0k_2(k_2-1)\}b=$
$2n_0k_1(z-1)h+\{2k_2T-n_0[2(i+k_1)+1-(i-z+2)-(i+k_1+z)](i-z+1)-n_0(k_1+k_2)[2(i+k_1)+1]+$
$(n_0-1)k_1(i+k_1+z)-(i+k_1+z)(i-z+1)-n_0k_2(k_2-1)\}b=$
$2n_0k_1(z-1)h+\{2k_2T-n_0(k_1-1)(i-z+1)-n_0(k_1+k_2)[2(i+k_1)+1]+(n_0-1)k_1(i+k_1+z)+(k_1+k_2)(i+k_1+z)+$
$k_2(i-z+1)+k_2(k_1+k_2)-n_0k_2(k_2-1)\}b=$
$2n_0k_1(z-1)h+\{2k_2T-n_0k_1[(i-z+1)+2(i+k_1)+1-(i+k_1+z)]+n_0(i-z+1)-k_1(i+k_1+z)-$
$n_0k_2[2(i+k_1)+1+k_2-1]+k_1(i+k_1+z)+k_2(2i+1+k_2+2k_1)\}b=$
$2n_0k_1(z-1)h+\{2k_2T-2n_0k_1(i-z+1)-n_0k_1^2+n_0(i-z+1)-2n_0k_2(i+k_1)-n_0k_2^2+2k_2(i+k_1)+k_2(k_2+1)\}b=$
$2n_0k_1(z-1)h+\left\{k_2[2T-2(n_0-1)(i+k_1)]-n_0[(i-z+1)(2k_1-1)+k_1^2+k_2^2]+k_2(k_2+1)\right\}b=$
$2n_0k_1\{(z-1)h-(i-z+1)b\}+\left\{k_2[2T-2(n_0-1)(i+k_1)]+n_0[(i-z+1)-k_1^2-k_2^2]+k_2(k_2+1)\right\}b=$
$2n_0k_1\{(z-1)h-(i-z+k_1+1)b\}+\left\{k_2[2T-2(n_0-1)(i+k_1)]+n_0(i-z+1)+n_0k_1^2-(n_0-1)k_2^2+k_2\right\}b\geq 0$

since $n_0=\min B_i$, $n_0-1\in B_{i+k_1}\Rightarrow n_0-1<\frac{T}{i+k_1}\Rightarrow 2T-2(n_0-1)(i+k_1)>0$,
$(h,b)\in\left(B_{i+k_1}(z,1)\cap H_i(z,1)\right)\cap\left(B_{i+k_1+k_2}(z,1)\cap H_{i+k_1}(z,1)\right)\Rightarrow(z-1)h-(i-z+k_1+1)b\geq 0$,
while $n_0(i-z+1)+n_0k_1^2-(n_0-1)k_2^2+k_2$ is obviously positive •

C4.

$$S+\frac{D}{2}\{-2T-(n_0-1)z(z-1)+2(n_0-1)(z-1)(i+k_1+1)-n_0(z-2)(2i-z+3)\}h+$$
$$\frac{D}{2}\{(i-z+1)(i-z+2)+2(i-z+2)[n_0-k_1(n_0-1)]-(n_0-1)k_1(k_1-1)\}b\geq$$

$$S+\frac{D}{2}\{-2T-n_0(z-1)(z-2)-2n_0(z-2)(i+k_1+1)+2(n_0-1)(z-1)(i+k_1+k_2+1)-z(z-1)(n_0-1)\}h+$$
$$\frac{D}{2}\{(i+k_1+k_2-z+1)(i+k_1+k_2-z+2)-n_0k_2(k_2-1)\}b$$
when $(h,b)\in\left(H_{i+k_1}(z,1)\cap H_i(z-1,1)\right)\cap\left(H_{i+k_1+k_2}(z,1)\cap H_{i+k_1}(z-1,1)\right)$

$$A = \{n_0(z-1)(z-2)+2(n_0-1)(z-1)(i+k_1+1)-n_0(z-2)(2i-z+3)+2n_0(z-2)(i+k_1+1)-$$
$$2(n_0-1)(z-1)(i+k_1+k_2+1)\}h =$$
$$\{(n_0-1)(z-1)[(z-2)+2(i+k_1+1)-2(i+k_1+k_2+1)]+(z-1)(z-2)-n_0(z-2)(2i-z+3)+$$
$$2n_0(z-2)(i+k_1+1)\}h =$$
$$\{(n_0-1)(z-1)[(z-2)-2k_2]+(z-1)(z-2)+n_0(z-2)[2(i+k_1+1)-(2i-z+3)\}h =$$
$$\{(n_0-1)(z-1)[(z-2)-2k_2]+(z-1)(z-2)+n_0(z-2)[(z-1)+2k_1]\}h =$$
$$\{2n_0(z-1)(z-2)-2k_2(n_0-1)(z-1)+2n_0k_1(z-2)\}h$$

$$B = \{-(k_1+k_2)[(i-z+1)+(i-z+2)+(k_1+k_2)]+2n_0(i-z+2)-2k_1(n_0-1)(i-z+2)-(n_0-1)k_1(k_1-1)+$$
$$n_0k_2(k_2-1)\}b =$$
$$\{-(k_1+k_2)[2(i-z+k_1+k_2+1)-(k_1+k_2-1)]+2(n_0+k_1)(i-z+2)-2n_0k_1(i-z+2)+n_0k_2(k_2-1)-$$
$$(n_0-1)k_1(k_1-1)\}b =$$
$$\{(k_1+k_2)(k_1+k_2-1)-2(k_1+k_2)(i-z+k_1+k_2+1)+2(n_0+k_1)(i-z+2)-2n_0k_1(i-z+2)+n_0k_2(k_2-1)-$$
$$(n_0-1)k_1(k_1-1)\}b =$$
$$\{(k_1+k_2)(k_1+k_2-1)-2k_2(i-z+k_1+k_2+1)+2k_1[(i-z+2)-(i-z+k_1+k_2+1)]+2n_0(i-z+2)-$$
$$2n_0k_1(i-z+2)+n_0k_2(k_2-1)-(n_0-1)k_1(k_1-1)\}b =$$
$$\{(k_1+k_2)(k_1+k_2-1)-2k_2(i-z+k_1+k_2+1)-2k_1(k_1+k_2-1)+2n_0(i-z+2)-2n_0k_1(i-z+2)+$$
$$n_0k_2(k_2-1)-(n_0-1)k_1(k_1-1)\}b =$$
$$\{(k_2-k_1)(k_1+k_2-1)-2k_2(i-z+k_1+k_2+1)+2n_0(i-z+2)-2n_0k_1(i-z+2)+$$
$$n_0k_2(k_2-1)-(n_0-1)k_1(k_1-1)\}b$$

Therefore, $A+B = 2n_0k_1\{(z-2)h-(i-z+2)b\}+2k_2\{(z-1)h-(i-z+k_1+k_2+1)b\}+$

$$\{2n_0(z-1)(z-2)-2k_2n_0(z-1)\}h+$$
$$\{(k_2-k_1)(k_1+k_2-1)+2n_0(i-z+2)+n_0k_2(k_2-1)-(n_0-1)k_1(k_1-1)\}b \geq 0$$

since $(h,b) \in \left(H_{i+k_1}(z,1) \cap H_i(z-1,1)\right) \cap \left(H_{i+k_1+k_2}(z,1) \cap H_{i+k_1}(z-1,1)\right) \Rightarrow$
$(z-2)h-(i-z+2)b \geq 0$ and $(z-1)h-(i-z+k_1+k_2+1)b \geq 0$,

while the remaining part of the above relation is obviously positive

Determining the domain of the jump of the function $C_T(n, i, h, b)$

1.
$$B_i(z,1) = \{(z-1)h \leq (i+1-z)b \leq zh\} = \left\{\frac{i+1-z}{z}b \leq h \leq \frac{i+1-z}{z-1}b\right\} \text{ and}$$
$$B_{i+k}(z,1) = \{(z-1)h \leq (i+k+1-z)b \leq zh\} = \left\{\frac{i+k+1-z}{z}b \leq h \leq \frac{i+k+1-z}{z-1}b\right\}$$

The intersection $B_{i+k}(z,1) \cap B_i(z,1)$ is valid if

$$\frac{i+k+1-z}{z-1} \geq \frac{i+1-z}{z} \text{ and } \frac{i+1-z}{z-1} \geq \frac{i+k+1-z}{z} \Rightarrow (i+1) \leq z(k+1) \leq (i+k+1) \text{ and}$$

it is defined as follows: $B_{i+k}(z,1) \cap B_i(z,1) = \{(i+k+1-z)b \leq zh,\ (z-1)h \leq (i+1-z)b\}$ •

$$B_i(z-1,1)=\{(z-2)h\le(i+2-z)b\le(z-1)h\}=\left\{\frac{i+2-z}{z-1}b\le h\le\frac{i+2-z}{z-2}b\right\}$$

2.

$$H_{i+k}(z,1)=\{(i+k+1-z)b\le(z-1)h\le(i+k+2-z)b\}=\left\{\frac{i+k+1-z}{z-1}b\le h\le\frac{i+k+2-z}{z-1}b\right\}$$

The intersection $H_{i+k}(z,1)\cap B_i(z-1,1)$ is valid if $\frac{i+k+1-z}{z-1}\le\frac{i+2-z}{z-2}\Rightarrow k(z-2)\le i$,

while it is defined as follows:

- if $\frac{i+2-z}{z-1}\le\frac{i+k+1-z}{z-1}\le\frac{i+2-z}{z-2}\le\frac{i+k+2-z}{z-1}\Rightarrow(i+2-z)\le k(z-2)\le i$ then

  $H_{i+k}(z,1)\cap B_i(z-1,1)=\{(i+k+1-z)b\le(z-1)h,\ (z-2)h\le(i+2-z)b\}$

- if $\frac{i+2-z}{z-1}\le\frac{i+k+1-z}{z-1}\le\frac{i+k+2-z}{z-1}\le\frac{i+2-z}{z-2}\Rightarrow k(z-2)\le(i+2-z)$ then

  $H_{i+k}(z,1)\cap B_i(z-1,1)=\{(i+k+1-z)b\le(z-1)h\le(i+k+2-z)b\}$

$$H_i(z,1)=\{(i+1-z)b\le(z-1)h\le(i+2-z)b\}=\left\{\frac{i+1-z}{z-1}b\le h\le\frac{i+2-z}{z-1}b\right\}$$

3.

$$B_{i+k}(z,1)=\{(z-1)h\le(i+k+1-z)b\le zh\}=\left\{\frac{i+k+1-z}{z}b\le h\le\frac{i+k+1-z}{z-1}b\right\}$$

The intersection $B_{i+k}(z,1)\cap H_i(z,1)$ is valid if $\frac{i+k+1-z}{z}\le\frac{i+2-z}{z-1}\Rightarrow k(z-1)\le(i+1)$,

while it is defined as follows:

- if $\frac{i+1-z}{z-1}\le\frac{i+k+1-z}{z}\le\frac{i+2-z}{z-1}\le\frac{i+k+1-z}{z-1}\Rightarrow i+1-z\le k(z-1)\le i+1$ then

  $B_{i+k}(z,1)\cap H_i(z,1)=\{(i+k+1-z)b\le zh,\ (z\text{-}1)h\le(i+2-z)b\}$

- if $\frac{i+k+1-z}{z}\le\frac{i+1-z}{z-1}\le\frac{i+2-z}{z-1}\le\frac{i+k+1-z}{z-1}\Rightarrow k(z-1)\le i+1-z$ then

  $B_{i+k}(z,1)\cap H_i(z,1)=\{(i+1-z)b\le(z\text{-}1)h\le(i+2-z)b\}$

$$H_i(z-1,1)=\{(i+z-2)b\le(z-2)h\le(i+3-z)h\}=\left\{\frac{i+z-2}{z-2}b\le h\le\frac{i+3-z}{z-2}b\right\}$$

4.

$$H_{i+k}(z,1)=\{(i+k+1-z)b\le(z-1)h\le(i+k+2-z)b\}=\left\{\frac{i+k+1-z}{z-1}b\le h\le\frac{i+k+2-z}{z-1}b\right\}$$

The intersection $H_{i+k}(z,1)\cap H_i(z-1,1)$ is valid if

$$\frac{i+k+2-z}{z-1}\ge\frac{i+2-z}{z-2}\text{ and }\frac{i+3-z}{z-2}\ge\frac{i+k+1-z}{z-1}\Rightarrow(i-z+2)\le k(z-2)\le(i+z-1),$$

while it is defined as follows:

- if $\frac{i+k+1-z}{z-1} \le \frac{i+2-z}{z-2} \le \frac{i+k+2-z}{z-1} \le \frac{i+3-z}{z-2} \Rightarrow i+2-z \le k(z-2) \le i$ then

  $H_{i+k}(z,1) \cap H_i(z-1,1) = \{(i+2-z)b \le (z-2)h,\ (z-1)h \le (i+k+2-z)b\}$

- if $\frac{i+2-z}{z-2} \le \frac{i+k+1-z}{z-1} \le \frac{i+3-z}{z-2} \le \frac{i+k+2-z}{z-1} \Rightarrow i \le k(z-2) \le i+1-z$ then

  $H_{i+k}(z,1) \cap H_i(z-1,1) = \{(i+k+1-z)b \le (z-1)h,\ (z-2)h \le (i+3-z)b\}$

i	B_i	$C_T(n, i, h, b)$		z	ΔC_T	$J(B_t, B_{t+k})$	
1	*{12, 11, ..., 6}*	$nS+(12-n)Db$	$\{b \le h\}$	1	$S-Db$	$S-3Db$	$\{2b \le h\}$
						$S-2Dh+Db$	$\{b \le h \le 2b\}$
		$nS+(12-n)Dh$	$\{h \le b\}$	2	$S-Dh$	$S+Dh-2Db$	$\{\frac{b}{2} \le h \le b\}$
						$S-3Dh$	$\{h \le \frac{b}{2}\}$
2	*{5, 4}*	$nS+(24-3n)Db$	$\{2b \le h\}$	1	$S-3Db$	$S-6Db$	$\{3b \le h\}$
						$S-3Dh+3Db$	$\{2b \le h \le 3b\}$
		$nS+(12-2n)Dh+nDb$	$\{b \le h \le 2b\}$	2	$S-2Dh+Db$	$S+Dh-5Db$	$\{b \le h \le 2b\}$
		$nS+nDh+(12-2n)Db$	$\left\{\frac{b}{2} \le h \le b\right\}$	2	$S+Dh-2Db$	$S-5Dh+Db$	$\{\frac{b}{2} \le h \le b\}$
		$nS+(24-3n)Dh$	$\left\{h \le \frac{b}{2}\right\}$	3	$S-3Dh$	$S+3Dh-3Db$	$\{\frac{b}{3} \le h \le \frac{b}{2}\}$
						$S-6Dh$	$\{h \le \frac{b}{3}\}$
3	*{3}*	$3S+18Db$	$\{3b \le h\}$	1	---	$S-12Db$	$\{5b \le h\}$
						$S-2Dh-2Db$	$\{3b \le h \le 5b\}$
		$3S+3Dh+9Db$	$\{b \le h \le 3b\}$	2		$S+Dh-11Db$	$\{2b \le h \le 3b\}$
						$S-3Dh-3Db$	$\{\frac{b}{2} \le h \le 2b\}$
		$3S+9Dh+3Db$	$\left\{\frac{b}{3} \le h \le b\right\}$	3		$S-11Dh+Db$	$\{\frac{b}{3} \le h \le \frac{b}{2}\}$
		$3S+18Dh$	$\left\{h \le \frac{b}{3}\right\}$	4		$S-2Dh-2Db$	$\{\frac{b}{5} \le h \le \frac{b}{3}\}$
						$S-12Dh$	$\{h \le \frac{b}{5}\}$

Table 1. Sign functions calculation for the problem with T = 12 (continued)

i	B_i	$C_T(n, i, h, b)$		z	ΔC_T	$J(B_i, B_{i+k})$	
		$2S+30Db$	$\{5b \le h\}$	1		$S-36Db$	$\{11b \le h\}$
						S – Dh – 25Db	$\{3b \le h \le 11b\}$
		$2S+2Dh+20Db$	$\{2b \le h \le 5b\}$	2		S – 4Dh – 16Db	$\{\frac{7}{5}b \le h \le 3b\}$
		$2S+6Dh+12Db$	$\{b \le h \le 2b\}$	3		S – 9Dh – 9Db	$\{\frac{5}{7}b \le h \le \frac{7}{5}b\}$
5	{2}	$2S+12Dh+6Db$	$\left\{\frac{b}{2} \le h \le b\right\}$	4	---	S – 16Dh – 4Db	$\{\frac{b}{3} \le h \le \frac{5}{7}b\}$
		$2S+20Dh+2Db$	$\left\{\frac{b}{5} \le h \le \frac{b}{2}\right\}$	5		$S-15Dh-Db$	$\{\frac{b}{5} \le h \le \frac{b}{3}\}$
		$2S+30Dh$	$\left\{h \le \frac{b}{5}\right\}$	6		$S-25Dh-Db$	$\{\frac{b}{11} \le h \le \frac{b}{5}\}$
						$S-36Dh$	$\{h \le \frac{b}{11}\}$

Table 1. Sign functions calculation for the problem with T = 12 (continued)

i	B_i	$C_T(n, i, h, b)$		z	ΔC_T	$J(B_i, B_{i+k})$
		$S+66Db$	$\{11b \le h\}$	1		
		$S+Dh+55Db$	$\{5b \le h \le 11b\}$	2		
		$S+3Dh+45Db$	$\{3b \le h \le 5b\}$	3		
		$S+3Dh+36Db$	$\{2b \le h \le 3b\}$	4		
		$S+10Dh+28Db$	$\left\{\frac{7}{5}b \le h \le 2b\right\}$	5		
		$S+15Dh+21Db$	$\left\{b \le h \le \frac{7}{5}b\right\}$	6		
11	{1}	$S+21Dh+15Db$	$\left\{\frac{5}{7}b \le h \le b\right\}$	7	---	---
		$S+28Dh+10Db$	$\left\{\frac{b}{2} \le h \le \frac{5}{7}b\right\}$	8		
		$S+36Dh+6Db$	$\left\{\frac{b}{3} \le h \le \frac{b}{2}\right\}$	9		
		$S+45Dh+3Db$	$\left\{\frac{b}{5} \le h \le \frac{b}{3}\right\}$	10		
		$S+55Dh+Db$	$\left\{\frac{b}{11} \le h \le \frac{b}{5}\right\}$	11		
		$S+66Dh$	$\left\{h \le \frac{b}{11}\right\}$	12		

Table 1. Sign functions calculation for the problem with T = 12