

チェビシェフ展開形で表わされたレゾルベントの多項式 によるフィルタの伝達特性の調整

村上 弘

HIROSHI MURAKAMI

首都大学東京 数理情報科学専攻

DEPARTMENT OF MATHEMATICS AND INFORMATION SCIENCES,

Tokyo Metropolitan University *

要 約

フィルタ対角化法を用いて、実対称定値一般固有値問題の指定区間に固有値がある固有対の近似を求める。単一のレゾルベントの多項式をフィルタとして用いることで、複数のレゾルベントの線形結合を用いる場合に比べて演算量と記憶量が低減できることが期待できる。固有値分布の内側の固有値を持つ固有対を求める場合にはレゾルベントのシフトには虚数を選ぶが、固有値分布の端にある固有値を持つ固有対を求める場合にはシフトに実数を選ぶことができる。シフトに実数を選ぶと、すべての計算を実数演算だけを用いて複素数を用いずに行なえるので、演算量と記憶量が減らせて有利になる。

単一のレゾルベントの多項式をフィルタとして採用する場合、その多項式を特にチェビシェフ多項式とすると、阻止域では伝達率の大きさの上限を容易に小さく抑えることができる。しかし通過域では伝達率の最大最小比が小さくないため、必要な近似対の精度は均一性が良くない可能性がある。そこでフィルタに用いる単一のレゾルベントの多項式を、一つのチェビシェフ多項式で表わされるものから、複数のチェビシェフ多項式の和で表わされるチェビシェフ展開形へと拡張することにより阻止域に於ける減衰率を十分に確保しながら、通過域に於ける伝達率の最大最小比をできるだけ抑えることを試みた。

ABSTRACT

For a real symmetric definite generalized eigenproblem, by using the filter diagonalization method, we solve approximations of those eigenpairs whose eigenvalues are in a specified interval. We can expect to reduce both amounts of computation and memory space when we use a filter which is a polynomial of a single resolvent rather than to use a linear combination of many resolvents. When we solve eigenpairs with interior eigenvalues for the shift of the resolvent we choose a complex number. We can choose a real number for the shift when we solve pairs with exterior eigenvalues. If a real number is chosen for the shift, the calculation is made only by real arithmetics rather than complex ones, we have an advantage to reduce both amounts of computation and memory space.

When a filter is a polynomial of a single resolvent, if the polynomial is chosen as a Chebyshev polynomial then it is easy to obtain a small upper-bound for the magnitude of the transfer rate of the filter in the stopband. However, for such a filter, the ratio of the maximum and the minimum of the magnitude of the transfer rates in the passband cannot be made so small, it is likely that the uniformity of approximations of required eigenpairs is not so good. Therefore, in this report of research, to reduce the ratio, we tried to extend the polynomial for the filter from a single Chebyshev polynomial of a certain degree to a sum of Chebyshev polynomials up to a certain degree.

*mrkmhrsh@tmu.ac.jp

1 はじめに

いま係数行列 A と B が実対称で、 B は正定値である一般固有値問題

$$A\mathbf{v} = \lambda B\mathbf{v}$$

の、下端固有対であって固有値が指定された区間 $[a, b]$ にあるものを解くのにフィルタ対角化法を用いることにする（区間の左端 a は最小固有値以下の値とする）。そのためのフィルタとして、固有値が $[a, b]$ 近傍にある固有ベクトルは良く通過させるが、 $[a, b]$ から離れた固有ベクトルは強く減衰させる線形作用素を用いる。

2 フィルタ対角化法の概要

フィルタ対角化法の概要は、以下のようになる。

1. まず固有値が $[a, b]$ 近傍の固有ベクトルを良く通過させるが、 $[a, b]$ から離れた固有ベクトルは強く阻止するフィルタ $\mathcal{F}$ を用意しておく。
2. ランダムな N 次ベクトル m 個の組を作り、それを B -正規直交化して X ($N \times m$ 行列) とする ($X^T B X = I$) 。
3. フィルタ $\mathcal{F}$ をいま作った X へ作用させて、濾過されたベクトルの組 $Y \leftarrow \mathcal{F} X$ ($N \times m$ 行列) を作成する。
4. 得られた Y の列の適切な線形結合の組により「 $[a, b]$ 近傍にある固有値すべてに対応する不変部分空間」を近似する空間の基底を構成する（その構成は X , Y および伝達関数の特性を参照して決める）。
5. 構成した基底にレイリー・リッツ法を適用して、得られたリッツ対を一般固有値問題の近似対にする。

3 レゾルベントの多項式のフィルタとその伝達関数

今回は、下端固有対を解くためのフィルタとして、「実数シフトのレゾルベント」の多項式を採用する。今扱っている一般固有値問題に対応して、シフトが ρ のレゾルベントを

$$\mathcal{R}(\rho) \equiv (A - \rho B)^{-1} B \quad (1)$$

と定義する。そうして実数 ρ と実多項式 $\mathcal{P}$ をうまく選んで作ったレゾルベントの多項式をフィルタとする：

$$\mathcal{F} \equiv \mathcal{P}(\mathcal{R}(\rho)). \quad (2)$$

ここで、 $\mathcal{P}$ は n 次の実多項式である。シフト ρ が最小固有値未満であればフィルタは有界作用素になる。

いまの固有値問題の場合は、固有対 $(\lambda, \mathbf{v})$ はすべて実にとれる。そうしてレゾルベントの固有ベクトルへの作用は

$$\mathcal{R}(\rho) \mathbf{v} = \frac{1}{\lambda - \rho} \mathbf{v} \quad (3)$$

となる。すると、フィルタ $\mathcal{F} = \mathcal{P}(\mathcal{R}(\rho))$ の固有ベクトルに対する作用は

$$\mathcal{F} \mathbf{v} = f(\lambda) \mathbf{v}. \quad (4)$$

となる。ここでフィルタの伝達関数 $f(\lambda)$ は実有理関数で

$$f(\lambda) = \mathcal{P}\left(\frac{1}{\lambda - \rho}\right) \quad (5)$$

となる。

求めたい下端固有対の固有値の区間が $\lambda \in [a, b]$ であるとき、固有値 λ の正規化座標 t を $\lambda \in [a, b]$ から $t \in [0, 1]$ への 1 次変換 $\lambda = a + (b - a)t$ により定義する。その逆変換は $t = \frac{1}{b-a}(\lambda - a)$ となる。そうして正規化座標 t の伝達関数を $g(t) = f(\lambda)$ で定義する。そうしていま 1 より大きいパラメタ μ を導入して $0 \leq t \leq 1$ を通過域、 $1 < t < \mu$ を遷移域、 $\mu \leq t < \infty$ を阻止域とする。(図 1 参照)。

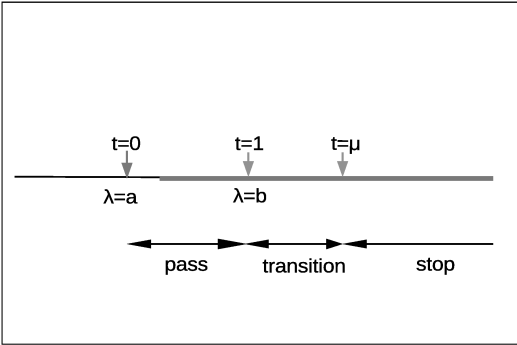


図 1: 固有値 λ の区間 $[a, b]$ と正規化座標 t の関係
通過域 $\lambda \in [a, b]$, $t \in [0, 1]$; 遷移域 $t \in (1, \mu)$; 阻止域 $[mu, \inf)$

伝達関数 $g(t)$ の形状について満たすべき条件を、パラメタ g_p, g_s を $1 > g_p \gg g_s > 0$ として、

- 阻止域では $|g(t)| \leq g_s$ である。
- 通過域では $g(t) \geq g_p$ で、 t が非負の範囲では逆もなりたつ。
- 通過域での $g(t)$ の最大値は 1 (と規格化)。

であるとする (図 2)。

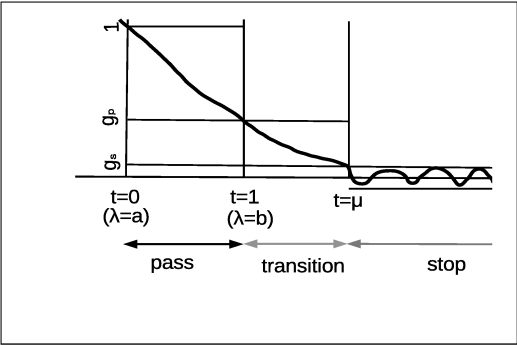


図 2: 伝達関数 $g(t)$ の概形

固有値の座標 λ と正規化座標 t との関係から, $\lambda = \rho$ ($< a$) に $t = -\sigma$ (< 0) が対応すれば, n 次実多項式 $\mathcal{P}$ により 実有理関数 $f(\lambda)$ が

$$f(\lambda) = \mathcal{P} \left(\frac{1}{\lambda - \rho} \right) \quad (6)$$

の形のとき, $g(t)$ も実有理関数で, ある n 次実多項式 $\mathcal{Q}$ を用いて

$$g(t) = \mathcal{Q} \left(\frac{1}{t + \sigma} \right) \quad (7)$$

の形でかける (λ を t に変換すると $\mathcal{P}$ から $\mathcal{Q}$ が決まる). 逆にこの形の式の $g(t)$ が決まれば, $f(\lambda)$ の表式も決まる.

4 伝達関数の三つの形状パラメタについて

伝達関数の三つの形状パラメタ μ , g_s , g_p について, まず前提条件から $\mu > 1$, $1 > g_p \gg g_s > 0$ であることが必要で, その上で形状の良さ・望ましさについては以下のように考える.

- μ は 1 に近いほど良い. なぜならば μ が大きいと遷移域が広くなり, それによって遷移域に固有値が多く入ればそれだけ (不変部分空間を近似する空間の基底を作るためには) 多くのベクトルを濾過する必要があるからである.
- 阻止域での伝達率の大きさの上限である g_s は, 微小であるほど良い. なぜならばこの値が微小でなければ, 固有値が阻止域にある固有ベクトルが十分には阻止されずに混入してしまい, 不変部分空間の近似が悪くなるからである.
- 通過域での伝達率の最小値である g_p は, 1 に近いほど良い. なぜならばこの値が 1 よりもずっと小さいと, 固有値が通過域にある固有対の相互間で伝達率の最大最小比が大きい場合は, 得られる近似対はその伝達率が小さいものはそれだけ精度が落ちてしまうからである.

たとえば, いま次数 n を固定したときに, 伝達関数

$$g(t) = \mathcal{Q} \left(\frac{1}{t + \sigma} \right) \quad (8)$$

の極の位置 $t = -\sigma$ と n 次実多項式 $\mathcal{Q}$ を調整して, なるべく $g(t)$ の三つの形状パラメタ μ , g_s , g_p を良いものにしようと試みることになる. ただし, これら三つの形状パラメタはそれらすべてを同時に良くすることができないトレードオフの関係にあるため, バランスを考えて最適化をすることになる. $\mathcal{Q}$ がまったく一般の実多項式とする場合には, 極の位置 $-\sigma$ も含めた最適化の計算は数値的な手法になり, 最適な極の位置や多項式の係数は得られてもただ数値としてだけ得られることになる.

そこで, チェビシェフ多項式の性質を利用することで, 最適ではないが, 簡単な式計算により制約を満たす伝達関数 $g(t)$ が求まる設計法を導入する. それにより得られる $g(t)$ は阻止域では値の大きさを小さく抑えられるので良い特性を持つが, 通過域の方では値の最大最小比を小さくできず特性が良くない.

5 単一のチェビシェフ多項式で表わされた伝達関数

いま伝達関数 $g(t)$ を n 次のチェビシェフ多項式で表わされた以下の形の式に制限する:

$$g(t) = g_s T_n(y), \quad y = 2x - 1, \quad x = \frac{\mu + \sigma}{t + \sigma}. \quad (9)$$

これはパラメタの三つ組 (n, μ, σ) で決定される (定数 g_s は規格化条件 $g(0) = 1$ を満たすように決める). 阻止域 $\mu \leq t < \infty$ には $-1 < y \leq 1$ の全域が対応する. この $g(t)$ は阻止域では $|g(t)| \leq g_s$ を満たし, チェビシェフ多項式の性質から阻止域以外の $0 \leq t < \mu$ では単調減少となるので, 残りの2条件 $g(0) = 1$, $g(1) = g_p$ から

$$\frac{1}{g_s} = T_n \left(1 + 2 \frac{\mu}{\sigma} \right), \quad \frac{g_p}{g_s} = T_n \left(1 + 2 \frac{\mu-1}{\sigma+1} \right) \quad (10)$$

が得られる. これを逆双曲線関数を用いて表せば

$$\cosh^{-1} \frac{1}{g_s} = 2n \cdot \sinh^{-1} \sqrt{\frac{\mu}{\sigma}}, \quad \cosh^{-1} \frac{g_p}{g_s} = 2n \cdot \sinh^{-1} \sqrt{\frac{\mu-1}{\sigma+1}} \quad (11)$$

となる.

5.1 三つ組 (n, μ, σ) からの g_s と g_p の計算

パラメタの三つ組 (n, μ, σ) が与えられたとき, それから g_s と g_p の値を求めるのには, 以下の二つの式の右辺をそれぞれ計算すれば良い.

$$\begin{cases} \frac{1}{g_s} \leftarrow \cosh \left(2n \cdot \sinh^{-1} \sqrt{\frac{\mu}{\sigma}} \right), \\ \frac{g_p}{g_s} \leftarrow \cosh \left(2n \cdot \sinh^{-1} \sqrt{\frac{\mu-1}{\sigma+1}} \right). \end{cases} \quad (12)$$

5.2 パラメタの三つ組として n, g_s, g_p を指定する場合

いま n, g_s, g_p が与えられたとき, それらから σ と μ の値を求めるのには,

$$\sqrt{\frac{\mu}{\sigma}} = \sinh \left(\frac{1}{2n} \cosh^{-1} \frac{1}{g_s} \right), \quad \sqrt{\frac{\mu-1}{\sigma+1}} = \sinh \left(\frac{1}{2n} \cosh^{-1} \frac{g_p}{g_s} \right). \quad (13)$$

上の各式の右辺を計算して, それぞれ w_1, w_2 とおくと

$$\frac{\mu}{\sigma} = w_1^2, \quad \frac{\mu-1}{\sigma+1} = w_2^2 \quad (14)$$

という二つの関係の組を得るが, これは σ と μ について

$$\sigma \leftarrow \frac{w_2^2 + 1}{(w_1 - w_2)(w_1 + w_2)}, \quad \mu \leftarrow \sigma w_1^2 \quad (15)$$

により簡単に解ける. こうして $g(t)$ に含まれるパラメタの三つ組 (n, μ, σ) が得られる. たとえば, $g_p = 10^{-7}$ と $g_s = 10^{-15}$ と指定したときに, 次数 n の値を 10 から 50 まで 5 刻みで変化させた場合の μ と σ の値を表 1 に示す. 次数 n をかなり増やしてみても μ の値のは緩慢にしか減少しない.

5.3 形状パラメタの三つ組に μ, g_p, g_s を指定する場合

形状パラメタの三つ組 μ, g_p, g_s が指定されたときに, それを (なるべく) 満たすように σ と n を決めれば, 伝達関数を決定する本来のパラメタの三つ組 (n, μ, σ) が揃うことになる. それには, 二つの制約

表 1: 例: $g_p = 10^{-7}$, $g_s = 10^{-15}$ の場合の, 次数 n の各場合に対する μ と σ の値

n	μ	σ
10	2.63	0.330
15	1.87	0.872
20	1.65	1.66
25	1.56	2.68
30	1.52	3.93
35	1.49	5.41
40	1.47	7.12
45	1.46	9.06
50	1.45	11.2

条件から得られる以下の n を表わす 2 通りの式:

$$n = \frac{\cosh^{-1} \frac{1}{g_s}}{2 \sinh^{-1} \sqrt{\frac{\mu}{\sigma}}}, \quad n = \frac{\cosh^{-1} \frac{g_p}{g_s}}{2 \sinh^{-1} \sqrt{\frac{\mu-1}{\sigma+1}}} \quad (16)$$

これら二つが一致する条件として, σ の非線形方程式 $G(\sigma) = r$ が得られる. その左辺と右辺の式は:

$$G(\sigma) \equiv \frac{\sinh^{-1} \sqrt{\frac{\mu-1}{\sigma+1}}}{\sinh^{-1} \sqrt{\frac{\mu}{\sigma}}}, \quad r \equiv \frac{\cosh^{-1} \frac{g_p}{g_s}}{\cosh^{-1} \frac{1}{g_s}}. \quad (17)$$

この方程式 $G(\sigma) = r$ は (たとえば 2 分法を用いて) 解くことができ, それにより σ の値を得る. そのとき n を表わす二つの式の値は当然一致するが, ただし一般には整数とならない. 次数は正整数でなければならないので, 計算により得られた実数 n の値を切り捨て (あるいは切り上げ) て整数化した値を次数 n として設定することにする.

すると, 既に解いて得られた σ と整数化で得た次数 n , 最初に与えた μ を併せて得られた三つ組 (n, μ, σ) から, 前述の方法で新たな g_s と g_p の値を求めることができる. このようにして得られた新たな g_s と g_p の値は最初に指定したときの値には一致しないが, その違いが応用上許容できるならば, 逆にあたかも最初からそれら変更した後の値を指定していたかのように扱えばつじつまを合わせることができる. それにより, チェビシェフ多項式で表わされた $g(t)$ はこのようにして決めたパラメタの三つ組 (n, μ, σ) で与えられる.

6 伝達関数 $g(t)$ からのフィルタ $\mathcal{F}$ の構成

伝達関数 $g(t)$ からフィルタ $\mathcal{F}$ を構成するには, 以下のようにする.

正規化座標 t と固有値 λ の間の関係である $t = \frac{1}{b-a}(\lambda - a)$ を用いると,

$$x = \frac{\mu + \sigma}{t + \sigma} = \frac{\ell}{\lambda - \rho}, \quad (18)$$

である. ただしここで $\ell \equiv (b-a)(\sigma + \mu)$, $\rho \equiv a - (b-a)\sigma$ である ($\sigma > 0$ より $\rho < a$ である).

伝達関数 $g(t) = g_s T_n(y)$ のチェビシェフ多項式の引数 $y = 2x - 1$ は

$$y = 2 \frac{\ell}{\lambda - \rho} - 1, \quad (19)$$

となるので、正規化座標 t による伝達関数である $g(t)$ を、固有値の座標 λ で表わした伝達関数 $f(\lambda)$ は、

$$f(\lambda) = g_s T_n \left(2 \frac{\ell}{\lambda - \rho} - 1 \right), \quad (20)$$

となる。

伝達関数 $f(\lambda)$ に対応するフィルタ作用素 $\mathcal{F}$ は、 $x = \frac{\ell}{\lambda - \rho}$ を作用素 $\ell \mathcal{R}(\rho)$ に置換し、また 1 を恒等作用素 I に置換すれば

$$\mathcal{F} = g_s T_n (2\ell \mathcal{R}(\rho) - I) \quad (21)$$

として得られる（注：複素エルミート定値一般固有値問題の場合でもフィルタの式は同じ形になる）。

6.1 レゾルベントの作用の実装

レゾルベントのベクトル $\mathbf{x}$ への作用 $\mathbf{y} \leftarrow \mathcal{R}(\rho)\mathbf{x}$ は係数の行列が $C = A - \rho B$ である連立 1 次方程式 $C\mathbf{y} = B\mathbf{x}$ を解くことで実装する。行列 A, B は実対称で B は正定値であり、シフト ρ は実数で最小固有値未満であることから、行列 C は実対称正定値である。係数行列が実対称正定値である連立 1 次方程式は、ピボット交換を行なわない修正コレスキ法で実数演算だけを用いて安定に解くことができる。さらに A, B が帯行列であれば C も帯行列になるので、帯行列用の効率的な解法が使用できる。

6.2 フィルタの作用の実装

レゾルベント $\mathcal{R}(\rho) = (A - \rho B)^{-1}B$ をベクトルの組 W に作用させる処理 $Z \leftarrow \mathcal{R}(\rho)W$ はまず W から右辺ベクトルの組 BW を作り、係数 $C \equiv A - \rho B$ の連立 1 次方程式の組 $CZ = BW$ を解くことで実現する。レゾルベントの n 次多項式を作用させる処理の中では、行列が C の連立 1 次方程式の組を解く処理が n 回現れる。そこでまず最初に C の行列分解を 1 度行なっておけば、その分解結果を利用することで連立 1 次方程式は容易に解くことができる。

いま X と Y が N 次ベクトル m 個の組 ($N \times m$ 行列) であるとき、 X にフィルタ $\mathcal{F}$ を作用して Y を作る処理である $Y \leftarrow \mathcal{F}X$ の算法を図 3 に示す。この算にはチェビシェフ多項式の三項漸化式を利用している (Z, V, W は作業用の $N \times m$ 行列である)。

7 伝達関数を設計した例

単一のチェビシェフ多項式で表された下端固有値用の伝達関数を設計した例を示す。正規化座標 t での伝達関数は、 $g(t) = g_s T_n(2x - 1)$ 、ただし $x \equiv \frac{\mu + \sigma}{t + \sigma}$ である。通過域は $0 \leq t \leq 1$ で、阻止域は $\mu \leq t < \infty$ である。三つ組 (n, μ, σ) を指定すれば、簡単な式計算により g_p, g_s が求まる。

- **フィルタ I** 設定した $g(t)$ のパラメタの三つ組は $n = 18, \mu = 2.0, \sigma = 1.8$ である。 $g_p = 3.10 \times 10^{-6}$, $g_s = 8.53 \times 10^{-15}$ となった。伝達関数の大きさ $|g(t)|$ のグラフを図 4 に示す。
- **フィルタ II** 設定した $g(t)$ のパラメタの三つ組は $n = 24, \mu = 1.5, \sigma = 3.0$ である。 $g_p = 3.15 \times 10^{-7}$, $g_s = 3.75 \times 10^{-14}$ となった。伝達関数の大きさ $|g(t)|$ のグラフを図 5 に示す。

```

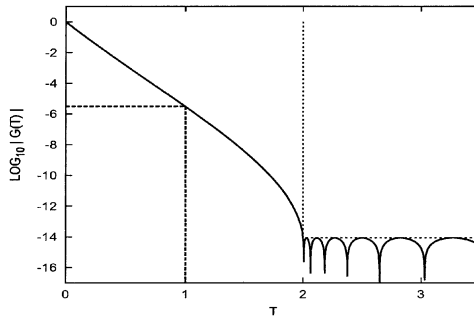
三つ組  $(n, \mu, \sigma)$  を与えて
 $\rho \leftarrow a - (b - a)\sigma$ ;    $\ell \leftarrow (b - a)(\sigma + \mu)$ ;

 $Z \leftarrow \mathcal{R}(\rho) X$ ;
 $V \leftarrow X$ ;
 $W \leftarrow 2\ell Z - X$ ;
for  $k := 2$  to  $n$  do begin
     $Z \leftarrow \mathcal{R}(\rho) W$ ;
     $Y \leftarrow 4\ell Z - 2W - V$ ;
     $V \leftarrow W$ ;
     $W \leftarrow Y$ 
end;
 $Y \leftarrow g_s Y$ 

```

図 3: 算法：フィルタのベクトルの組への作用 $Y \leftarrow \mathcal{F} X$

- **例フィルタ III** 設定した $g(t)$ のパラメタの三組は $n = 32$, $\mu = 2.0$, $\sigma = 6.11$ である. $g_p = 1.13 \times 10^{-5}$, $g_s = 1.45 \times 10^{-15}$ となった. 伝達関数の大きさ $|g(t)|$ のグラフを図 6 に示す.

図 4: (フィルタ I) 伝達関数の大きさ $|g(t)|$

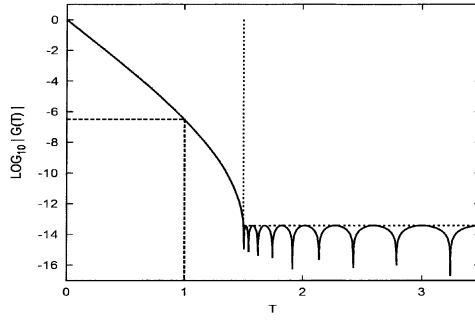
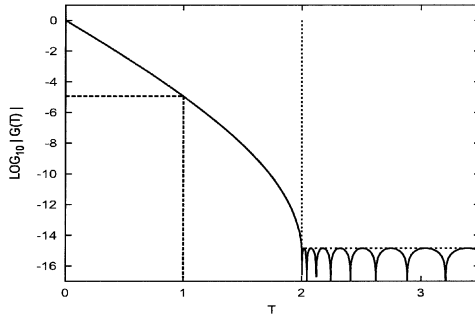
8 通過域に於ける伝達特性の改善案

さて, 単一のチェビシェフ多項式を用いた伝達関数の式

$$g(t) = g_s T_n(2x - 1), \quad x \equiv \frac{\mu + \sigma}{t + \sigma} \quad (22)$$

の, 通過域 $t \in [0, 1]$ に於ける値の最大最小比を減らす改善をしたい. 用いるレゾルベントは 1 個で, $g(t)$ の極は $t = \sigma$ だけにあるとする.

阻止域 $t \in [\mu, \infty)$ ではチェビシェフ多項式 T_n の引数 $2x - 1$ は $(-1, 1]$ にある. 各区間の端と対応する x の値は, $t = \infty$ は $x_\infty = 0$ に, $t = \mu$ は $x_\mu = 1$ に, $t = 1$ は $x_0 \equiv 1 + (\mu - 1)/(1 + \sigma)$ に, $t = 0$ は

図 5: (フィルタ II) 伝達関数の大きさ $|g(t)|$ 図 6: (フィルタ III) 伝達関数の大きさ $|g(t)|$

$x_1 \equiv 1 + \mu/\sigma$ に、それぞれ対応する。それゆえ $x_\infty \equiv 0 < x_\mu \equiv 1 < x_1 < x_0$ であつて、 $x \in (0, 1]$ が阻止域に、 $x \in (1, x_1)$ が遷移域に、 $x \in [x_1, x_0]$ が通過域に、それぞれ対応する。

8.1 チェビシエフ展開形への拡張

方法の自然な拡張方法の一つは、伝達関数を単一のチェビシエフ多項式で表わされたものから次数の異なるチェビシエフ多項式の和であるチェビシエフ展開形

$$g(t) \equiv \sum_{k=0}^n c_k T_k(2x - 1) \quad (23)$$

にすることであり、それと対応してフィルタ作用素

$$\mathcal{F} \equiv \sum_{k=0}^n c_k T_k(2\ell \mathcal{R}(\rho) - I) . \quad (24)$$

を採用することである。展開の係数 c_k は、あらかじめ（たとえば以下で示すような最小 2 乗法に類似した方法で）決定しておくものとする（簡単のためにシフト ρ は変更しないとする）。このようにすると、利用するレゾルベントは単一のままでよくて、ベクトル $\mathbf{v}$ に対するフィルタ $\mathcal{F}$ の作用の計算は、 $T_k(2\ell \mathcal{R}(\rho) - I) \mathbf{v}$ をチェビシエフ多項式に対する 3 項漸化式を用いて k を 0 から n まで上昇させて作りながら、そのつどそれに係数 c_k を乗じた累和を作ることにより実現できる。

8.2 最小2乗法による最適化

阻止域 $x \in [0, 1]$ で (重み付きの) 「伝達関数の 0 からのずれ」の 2 乗平均の 2 乗は以下の式になる：

$$\begin{aligned}
 J_{\text{stop}} &\equiv \int_0^1 \frac{\{g(t)\}^2}{\sqrt{1-(2x-1)^2}} dx \\
 &= \sum_{i,j=0}^n c_i c_j \int_0^1 \frac{T_i(2x-1)T_j(2x-1)}{\sqrt{1-(2x-1)^2}} dx \\
 &= \frac{1}{2} \sum_{i,j=0}^n c_i c_j \int_{-1}^1 \frac{T_i(y)T_j(y)}{\sqrt{1-y^2}} dy \\
 &= \frac{1}{2} \left(2c_0^2 + \sum_{j=1}^n c_j^2 \right).
 \end{aligned} \tag{25}$$

また、通過域 $x \in [x_1, x_0]$ に於ける「伝達関数の値の 1 からのずれ」の 2 乗平均の 2 乗は：

$$\begin{aligned}
 J_{\text{pass}} &\equiv \int_{x_1}^{x_0} \{1-g(t)\}^2 dx \\
 &= \int_{x_1}^{x_0} \{g(t)\}^2 dx - 2 \int_{x_1}^{x_0} g(t) dx + \text{Const} \\
 &= \sum_{i,j=0}^n \mathcal{A}_{i,j} c_i c_j - 2 \sum_{j=0}^n \mathbf{b}_j c_j + \text{Const}
 \end{aligned} \tag{26}$$

ここで、

$$\begin{cases} \mathcal{A}_{i,j} &\equiv \int_{x_1}^{x_0} T_i(2x-1)T_j(2x-1) dx, \\ \mathbf{b}_j &\equiv \int_{x_1}^{x_0} T_j(2x-1) dx = \mathcal{A}_{j,0}. \end{cases} \tag{27}$$

係数 $\mathcal{A}_{i,j}$ の定積分の計算 $y = 2x - 1$ から $y_1 = 2x_1 - 1 = 1 + 2 \cdot \frac{\mu - 1}{1 + \sigma}$, $y_0 = 2x_0 - 1 = 1 + 2 \cdot \frac{\mu}{\sigma}$ とおくと、公式 $T_i(y)T_j(y) = \{T_{i+j}(y) + T_{|i-j|}(y)\}/2$ を使うと

$$\begin{aligned}
 \mathcal{A}_{i,j} &= \frac{1}{2} \int_{y_1}^{y_0} T_i(y)T_j(y) dy \\
 &= \frac{1}{4} \int_{y_1}^{y_0} \{T_{i+j}(y) + T_{|i-j|}(y)\} dy \\
 &= \frac{1}{8} (K_{i+j} + K_{|i-j|}).
 \end{aligned} \tag{28}$$

定積分 K_ℓ の計算法 $K_\ell \equiv 2 \int_{y_1}^{y_0} T_\ell(y) dy$ ($\ell \geq 0$) に対応する不定積分は：

$$2 \int T_\ell(y) dy = \begin{cases} 2y & \ell = 0 \text{ の場合,} \\ \frac{1}{2} T_2(y) & \ell = 1 \text{ の場合,} \\ \frac{1}{\ell+1} T_{\ell+1}(y) - \frac{1}{\ell-1} T_{\ell-1}(y) & \ell \geq 2 \text{ の場合.} \end{cases} \tag{29}$$

よって定積分 K_ℓ ($\ell \geq 0$) は,

$$K_\ell = \begin{cases} 2(y_0 - y_1) & \ell = 0 \text{ の場合,} \\ y_0^2 - y_1^2 & \ell = 1 \text{ の場合,} \\ \frac{1}{\ell+1} \{T_{\ell+1}(y_0) - T_{\ell+1}(y_1)\} - \frac{1}{\ell-1} \{T_{\ell-1}(y_0) - T_{\ell-1}(y_1)\} & \ell \geq 2 \text{ の場合.} \end{cases} \quad (30)$$

上記の方法を用いて $\mathcal{A}_{i,j}$ と b_j を計算すれば:

$$\begin{cases} J_{\text{stop}} & \equiv \frac{1}{2} \left(2c_0^2 + \sum_{j=1}^n c_j^2 \right), \\ J_{\text{pass}} & \equiv \sum_{i,j=0}^n \mathcal{A}_{i,j} c_i c_j - 2 \sum_{j=0}^n b_j c_j + \text{Const}, \end{cases} \quad (31)$$

となる. いま阻止域に対する制約条件 $2J_{\text{stop}} = \varepsilon^2$ を与えて, 通過域に対する値 J_{pass} を最小にする $\mathbf{c} = [c_0, c_1, \dots, c_n]$ を決める. それには, ラグランジュの未定乗数法を用いる.

8.2.1 制約付き最小化

まず便利のために, J_{stop} が変数の平方和になるように上記のベクトル $\mathbf{c}$ の第 0 要素である c_0 だけをスケール変換したものを $\mathbf{c}'$ として, それに対応して行列 $\mathcal{A}$ の第 0 行と第 0 列だけをスケール変換したものを $\mathcal{A}'$ とし, ベクトル $\mathbf{b}$ の第 0 要素 b_0 をスケール変換して $\mathbf{b}'$ としておく. それはいま $n+1$ 次の対角行列を $\Delta \equiv \text{diag}(\sqrt{2}, 1, 1, \dots, 1)$ と置いて, それにより

$$\begin{aligned} \mathbf{c}' & \equiv \Delta \mathbf{c}, \\ \mathcal{A}' & \equiv \Delta^{-1} \mathcal{A} \Delta^{-1}, \\ \mathbf{b}' & \equiv \Delta^{-1} \mathbf{b} \end{aligned} \quad (32)$$

と書ける. それにより

$$\begin{cases} J_{\text{stop}} & = \frac{1}{2} \mathbf{c}'^T \mathbf{c}', \\ J_{\text{pass}} & = \mathbf{c}'^T \mathcal{A}' \mathbf{c}' - 2 \mathbf{b}'^T \mathbf{c}' + \text{Const}, \end{cases} \quad (33)$$

となるので, これに対してラグランジュの未定乗数法により制約付き最小化を行なう. そうして目的関数を

$$\mathcal{L}(\mathbf{c}', \eta) \equiv \frac{1}{2} \mathbf{c}'^T \mathcal{A}' \mathbf{c}' - \mathbf{b}'^T \mathbf{c}' + \frac{1}{2} \eta (\mathbf{c}'^T \mathbf{c}' - \varepsilon^2) \quad (34)$$

とすると, その最小化条件は:

$$\begin{cases} (\mathcal{A}' + \eta I) \mathbf{c}' = \mathbf{b}', \\ \mathbf{c}'^T \mathbf{c}' = \varepsilon^2, \end{cases} \quad (35)$$

となる. すると, η を未知数として $\mathbf{c}'$ を表わせば,

$$\mathbf{c}' = (\mathcal{A}' + \eta I)^{-1} \mathbf{b}' \quad (36)$$

であるから, それを制約条件の式 $\mathbf{c}'^T \mathbf{c}' = \varepsilon^2$ に代入すると, 以下の η 単独の方程式:

$$\mathbf{b}'^T (\mathcal{A}' + \eta I)^{-2} \mathbf{b}' = \varepsilon^2, \quad (37)$$

が得られる. この方程式を解いて得られた η の値を用いて $\mathbf{c}'$ を

$$\mathbf{c}' \leftarrow (\mathcal{A}' + \eta I)^{-1} \mathbf{b}' \quad (38)$$

と計算する. このようにして得られた $\mathbf{c}'$ の最初の要素を $\mathbf{c} \leftarrow \Delta^{-1} \mathbf{c}'$ によりスケール変更して $\mathbf{c}$ を作る.

8.2.2 固有値分解を利用した η の解法

あらかじめ実対称正定値行列 A' の対角化 $A' \rightarrow UDU^T$, $U^TU = I$ を行い, それを用いて $\beta \equiv U^T \mathbf{b}'$ とおくと, η を決めるための方程式は:

$$\varepsilon^2 = \mathbf{b}'^T U (D + \eta I)^{-2} U^T \mathbf{b}' = \beta^T (D + \eta I)^{-2} \beta \quad (39)$$

となる. この関係をベクトルの成分を用いて書き直すと:

$$\sum_{j=0}^n \left(\frac{\beta_j}{d_j + \eta} \right)^2 - \varepsilon^2 = 0 \quad (40)$$

となる. この等式 (40) の左辺を $h(\eta)$ と書けば, 対角行列 D は数学的には正值なので, $\eta > 0$ で $h(\eta)$ は連続な狭義の単調減少関数であり, $\eta \rightarrow \infty$ のとき $h(\eta)$ は負の値 $-\varepsilon^2$ にちかづく. すると $\eta = 0$ のときの値 $h(0) \equiv \sum_{j=0}^n (\beta_j/d_j)^2 - \varepsilon^2$ が正であれば, 方程式 $h(\eta) = 0$ の正の根 η が唯一存在する. (なお実際には数値計算の誤差の影響で D の対角要素 d_j に負の値が現れたらそれをすべて零に置き換えることにする. その場合は $\eta = 0$ は $h(\eta)$ の極になるが, それでもやはり $h(\eta)$ は $\eta > 0$ で連続で狭義単調減少であり, $\eta \rightarrow 0+$ のとき $h(\eta) \rightarrow +\infty$ であるから, $h(\eta) = 0$ の正の根 η が唯一存在する.) すると非線形方程式 $h(\eta) = 0$ に対してたとえば二分法を用いることで唯一の正根 η を必ず求めることができ, それにより

$$\mathbf{c}' = A'^{-1} \mathbf{b}' = U (D + \eta I)^{-1} U^T \mathbf{b}' = U (D + \eta I)^{-1} \beta \quad (41)$$

によりベクトル $\mathbf{c}'$ が求まる. その $\mathbf{c}'$ から第 0 要素だけのスケール変更 $\mathbf{c} \leftarrow \Delta^{-1} \mathbf{c}'$ により $\mathbf{c}$ を作る.

8.2.3 チェビシエフ展開形へ拡張したフィルタの状況

上記の方法で構成されたフィルタがはたして期待どおりに機能するかどうかについては実際に試験をして確かめてみる必要がある. なぜならば, 通過域に於ける伝達関数の値の最大最小比を削減できたとしても, それは通過域に於ける強い相殺により実現されているので, 実際には丸め誤差の拡大が起きていて, 固有値が通過域にある求めたい固有ベクトルは良い精度では得られない可能性があるからである.

その危惧について, 以下のグラフで見えていく.

- 図 7 は, $n=30$, $\mu=2.0$, $\sigma=3.0$, $\varepsilon=10^{-13}$ としてチェビシエフ展開の係数を最小 2 乗法的方法で決めて構成した伝達関数 (その 1) の大きさのグラフである. $g_p = 9.2 \times 10^{-4}$, $g_s = 1.4 \times 10^{-13}$ となった.
- 図 8 は, $n=30$, $\mu=1.5$, $\sigma=3.0$, $\varepsilon=10^{-12}$ としてチェビシエフ展開の係数を最小 2 乗法的方法で決めて構成した伝達関数 (その 2) の大きさのグラフである. $g_p = 2.9 \times 10^{-5}$, $g_s = 1.9 \times 10^{-12}$ となった.
- 図 9 は, $n=40$, $\mu=1.25$, $\sigma=1.5$, $\varepsilon=10^{-14}$ としてチェビシエフ展開の係数を最小 2 乗法的方法で決めて構成した伝達関数 (その 3) の大きさのグラフである. $g_p = 3.4 \times 10^{-7}$, $g_s = 1.5 \times 10^{-14}$ となった.

これらをみると, いずれも通過域で伝達関数の頂上付近が値が押し潰されて振動しており, 通過域での伝達率の最大最小比を低減できていることが分かる. しかし, $g(t)$ の式計算の中では各項の数値は丸めを行いながら加算がされていることを思い出す必要がある. たとえば $g(t) = \sum_{k=0}^n c_k T_k(2x-1)$ の計算で, 横軸 t の値に対して, チェビシエフ多項式の三項漸化式を利用して項番号 k を 0 から n まで上げながら係数 c_k を乗じて和をとりながら計算していくときに, 途中で部分和の値が到達した絶対値最大の値をグラフにプロットしたものを書き添えると, 以下ようになる (通過域に於いて上側にあるグラフが書き加えたものである).

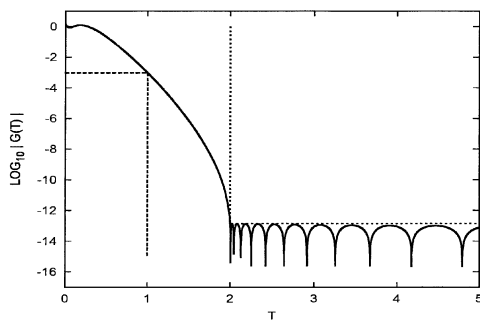


図 7: 最小 2 乗の方法による伝達関数 (その 1): $|g(t)|$ ($n=30$, $\mu=2.0$, $\sigma=3.0$, $\epsilon=10^{-13}$) $g_p = 9.2 \times 10^{-4}$, $g_s = 1.4 \times 10^{-13}$

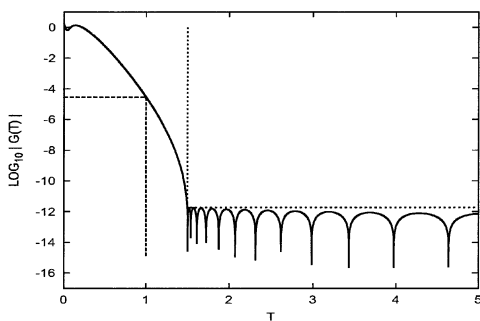


図 8: 最小 2 乗の方法による伝達関数 (その 2): $|g(t)|$ ($n=30$, $\mu=1.5$, $\sigma=3.0$, $\epsilon=10^{-12}$) $g_p = 2.9 \times 10^{-5}$, $g_s = 1.9 \times 10^{-12}$

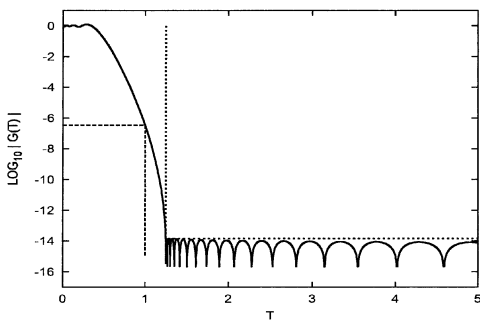


図 9: 最小 2 乗の方法による伝達関数 (その 3): $|g(t)|$ ($n=40$, $\mu=1.25$, $\sigma=1.5$, $\epsilon=10^{-14}$) $g_p = 3.4 \times 10^{-7}$, $g_s = 1.5 \times 10^{-14}$

- 図 10 は図 7 に部分和がとった最大の絶対値のグラフを書き加えたものである。
- 図 11 は図 8 に部分和がとった最大の絶対値のグラフを書き加えたものである。

- 図 12 は図 9 に部分和がとった最大の絶対値のグラフを書き加えたものである。

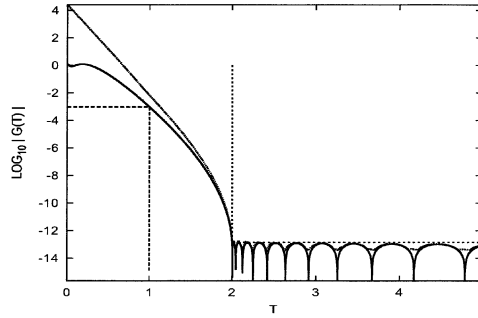


図 10: 最小 2 乗の方法による伝達関数 (その 1): $|g(t)|$ ($n=30, \mu=2.0, \sigma=3.0, \epsilon=10^{-13}$) $g_p = 9.2 \times 10^{-4}$, $g_s = 1.4 \times 10^{-13}$

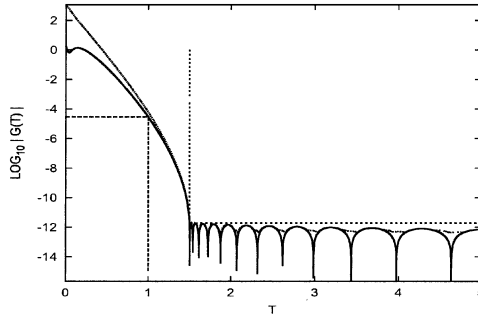


図 11: 最小 2 乗の方法による伝達関数 (その 2): $|g(t)|$ ($n=30, \mu=1.5, \sigma=3.0, \epsilon=10^{-12}$) $g_p = 2.9 \times 10^{-5}$, $g_s = 1.9 \times 10^{-12}$

8.3 チェビシェフ展開の低次項を省く近似法

チェビシェフ展開に含まれる各項の基底多項式 T_k のうちで、通過域付近 ($1 < x$) では高次側のものが相対的に値が大きくて影響が支配的となる。そこで、展開係数を並べたベクトル $\mathbf{c}$ の要素のうちで、高次側の s 個 ($s \leq n$) の係数 $c_{n-s+1}, \dots, c_n$ の自由度だけを残して、それ以外の低次側の係数 $c_0, c_1, \dots, c_{n-s}$ は零に制限する「近似」を導入してみる。そうしてここでは零に制限しない高次側 s 個の係数を集めた s 次のベクトルを $\mathbf{c}' = \{c_{n-s+1}, \dots, c_n\}$ と表わし、またそれと対応するように $\mathbf{A}$ の行と列の番号がどちらも $n-s+1$ 以上の部分の要素だけを並べた s 次の行列を $\mathbf{A}'$ とし、また同様に s 次のベクトル $\mathbf{b}' = \{b_{n-s+1}, \dots, b_n\}$ も作る。するとラグランジュの未定乗数法による制約付き最小化のための目的関数の式は、

$$\mathcal{L}(\mathbf{c}', \eta') \equiv \frac{1}{2} \mathbf{c}'^T \mathbf{A}' \mathbf{c}' - \mathbf{b}'^T \mathbf{c}' + \frac{1}{2} \eta' (\mathbf{c}'^T \mathbf{c}' - \epsilon^2) \quad (42)$$

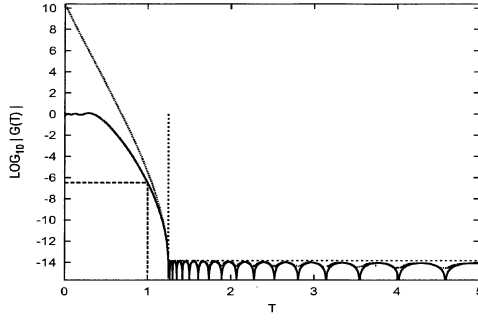


図 12: 最小 2 乗的方法による伝達関数 (その 3): $|g(t)|$ ($n=40$, $\mu=1.25$, $\sigma=1.5$, $\epsilon=10^{-14}$) $g_p = 3.4 \times 10^{-7}$, $g_s = 1.5 \times 10^{-14}$

となるので, その最小化条件は以前と同様に

$$\begin{cases} (\mathcal{A}' + \eta' I) \mathbf{c}' = \mathbf{b}', \\ \mathbf{c}'^T \mathbf{c}' = \epsilon^2. \end{cases} \quad (43)$$

となる. すると, これもまた同様に η' を正の未知数として

$$\mathbf{c}' = (\mathcal{A}' + \eta' I)^{-1} \mathbf{b}' \quad (44)$$

と書けるので, それを $\mathbf{c}'^T \mathbf{c}' = \epsilon^2$ に代入すると, 単独の変数 η' に対する以下の方程式が得られる:

$$\mathbf{b}'^T (\mathcal{A}' + \eta' I)^{-2} \mathbf{b}' = \epsilon^2. \quad (45)$$

この非線形方程式を解いて η' の唯一の正の値を得たら, それを用いて

$$\mathbf{c}' \leftarrow (\mathcal{A}' + \eta' I)^{-1} \mathbf{b}' \quad (46)$$

により $\mathbf{c}'$ を計算する. 前と同様に η' や $\mathbf{c}'$ を求める実際の計算には, 実対称行列 $\mathcal{A}'$ の固有値分解が利用できる.

9 まとめ

実対称定値一般固有値問題で固有値が指定した区間 $[a, b]$ にある下端固有対をフィルタ対角化法で求める (但し a は最小固有値以下の値とする). そのために用いるフィルタを, シフトが実数の単一のレゾルベントの多項式によって構成する方法について考察を行った. 用いるレゾルベントの実数値のシフトは最小固有値より小さい値になる. レゾルベントの作用は実対称正定値の行列を係数とする連立 1 次方程式を解いて実現する. その連立 1 次方程式を行列分解を利用して解く場合は, 行列分解を最初に 1 度だけ行えば良く, 既に得られた分解結果を保持しておくそれをレゾルベントの作用を行なうたびに再利用することができる. 単一のレゾルベントを用いることにより, 実対称正定値の行列の対称分解を 1 度だけ行えばよい点が記憶容量の制約が強い場合には有利であり, また分解に費やす計算量がレゾルベントを多数用いる場合より少なくて済むことも利点である.

レゾルベントのチェビシェフ多項式により構成されるフィルタの伝達特性は, 通過域での伝達率の最大最小比が大きいのので得られる固有対の精度が不均一になるが, ごく簡単な式を用いてフィルタの設計が行な

える。実際に得られたフィルタを用いて対角化を試してみることで、この種類のフィルタがある程度うまく働くことは確認してある。

今回はその改良の方向として、単一のレゾルベントのチェビシェフ多項式から一般化して、引数が共通の異なる次数のチェビシェフ多項式の和（チェビシェフ展開形）を用いるフィルタの伝達関数を設計する方法の検討を加えた。一つの次数ではなくて異なる次数の複数のチェビシェフ多項式の線形和を採用することにより増えた自由度を用いて、通過域に於ける伝達率の最大最小比を低減することを狙った。

なお今回は省略したが、実対称定値の固有値問題で固有値が一般的な位置にある「中間固有対」をフィルタ対角化を用いて求めることもできる。それには今回 $x \equiv \frac{\mu + \sigma}{t + \sigma}$ としている箇所を $x \equiv \frac{\mu^2 + \sigma^2}{t^2 + \sigma^2}$ で置き換えて、その後は x の実多項式としてフィルタの伝達関数を扱えば、ほぼ同様の議論に沿ってフィルタの伝達特性の調整や設計ができることを示せる。その場合には、得られた伝達関数から構成されるフィルタは「ある複素数をシフトとする単一のレゾルベントの虚部」の実多項式になるが、その実多項式の形として（今回の場合と同様に）ある次数のチェビシェフ多項式や、通過域に於ける特性を自由度を増やすことで改善できるチェビシェフ展開形（複数のチェビシェフ多項式の和）が採用できる。

参 考 文 献

- [1] 村上弘: 固有値が指定された区間内にある固有対を解くための対称固有値問題用のフィルタの設計, 情報処理学会論文誌: コンピューティングシステム (ACS31), Vol.3, No.3 (2010年9月), pp.1-21.
- [2] 村上弘: 対称一般固有値問題のフィルタ作用素を用いた不変部分空間の近似構成, 情報処理学会論文誌: コンピューティングシステム (ACS35), Vol.4, No.4 (2011年10月), pp.1-14.
- [3] 村上弘: レゾルベントを用いたフィルタによる固有値問題の解法について, 情報処理学会研究報告, Vol.2012-HPC-133, No.22 (2012年3月), pp.1-8.
- [4] 村上弘: 実対称定値一般固有値問題の最小側固有値を持つ固有対に対する実数シフトのレゾルベントを組み合わせたフィルタによる解法, 先進的計算基盤システムシンポジウム論文集 2012, (2012年5月), pp.81-82.
- [5] 村上弘: Hermite 対称な定値一般固有値問題のフィルタ対角化法について, 情報処理学会研究報告, Vol.2012-HPC-134, No.1 (2012年6月), pp.1-8.
- [6] 村上弘: レゾルベントの線形結合をフィルタに用いたエルミート定値一般固有値問題のフィルタ対角化法, 情報処理学会論文誌: コンピューティングシステム (ACS45), Vol.7, No.1 (2014年3月), pp.57-72.
- [7] 村上弘: フィルタ対角化法について, 日本応用数理学会 2014 年度年会予稿集 (2014年8月), pp.329-330.
- [8] 村上弘: レゾルベントの多項式をフィルタとして用いる対角化法について, 情報処理学会研究報告, Vol.2014-HPC-146, No.13 (2014年9月), pp.1-4.
- [9] 村上弘: 実対称定値一般固有値問題に対するレゾルベントの多項式によるフィルタの構成法の検討, 情報処理学会研究報告, Vol.2014-HPC-147, No.2 (2014年12月), pp.1-10.
- [10] 村上弘: 実数シフトのレゾルベントを組み合わせたフィルタによる実対称定値一般固有値問題の下端付近の固有値を持つ固有対の解法, HPCS2015 シンポジウム論文集, Vol.2015 (2015年5月), pp.38-51.
- [11] 村上弘: 一つのレゾルベントから構成されたフィルタを用いた実対称定値一般固有値問題に対するフィルタ対角化法の実験, 情報処理学会研究報告, Vol.2015-HPC-149, No.7 (2015年6月), pp.1-16.

- [12] 村上弘:「実数シフトのレゾルベントの多項式をフィルタに用いた実対称定値一般固有値問題の下端付近の固有値を持つ固有対の解法」, **日本応用数学会 2015 年度年会予稿集 (統合版)** (2015 年 9 月 2 日), pp.442–443.
- [13] Hiroshi Murakami: "Filter diagonalization method for a real symmetric definite generalized eigenproblem whose filter is a polynomial of a resolvent", poster presentation [P-13] at EPASA2015 (International Workshop on Eigenvalue Problems: Algorithms, Software and Applications, in Petascale Computing), at International Congress Center, EPOCAL TSUKUBA, Tsukuba, Japan. (Sep.15th,2015). in abstracts, p.28 (single page poster abstract).
- [14] 村上弘:「レゾルベントの多項式によるフィルタの伝達特性の調整」, RIMS 共同研究「数式処理とその周辺分野の研究」, 於京都大学益川ホール (2015 年 12 月 4 日) に対する *RIMS 講究録原稿*, 14 頁分 (submitted).
- [15] 村上弘:「固有値問題の解法に用いるレゾルベントの多項式型のフィルタの設計」, **情報処理学会研究報告**, Vol.2016-HPC-153, No.38(2016 年 3 月 3 日), pp.1–13.
- [16] 村上弘:「虚数シフトのレゾルベントの多項式の実部をフィルタに用いた実対称定値一般固有値問題の中間固有対の解法」, **HPCS2016 シンポジウム論文集** (ポスタ発表論文) (2016 年 6 月 6 日), p.49(全 1 頁).
- [17] 村上弘:「実対称定値一般固有値問題の最小側固有対を解くための実数シフトのレゾルベントの多項式によるフィルタの簡易な設計法」, **情報処理学会研究報告集**, Vol.2016-HPC-155, No.44 (2016 年 8 月 10 日), pp.1–27.
- [18] 村上弘:「レゾルベントの多項式をフィルタに用いた対称定値一般固有値問題のフィルタ対角化法」, **日本応用数学会 2016 年度年会予稿集** (2016 年 9 月 13 日)2 頁分.