RIMS Computer Algebra - Fundation and Applications（オンライン）.
2022年12月21日(水) PM 9:30–10:00（30分間）

フィルタを用いた固有値問題の近似解法について

村上 弘（東京都立大学）

- 固有値問題で，固有値が指定された区間にある少数の固有対を一斉に近似して求める.
- 今回のフィルタはレゾルベントの線形結合の多項式.
- ベクトルの組にフィルタを適用して不要な固有ベクトルを含有する率を減らし，それから近似不変部分空間の基底を構成して，その基底から近似固有対を取り出す.
- 通常はランダムなベクトルの組に対してフィルタを適用するが，前処理として性能は低いが手間の少ないフィルタの適用によりあらかじめ不要な固有ベクトルの含有する率を減らしておけば近似固有対の精度が向上.

はじめに

- 一般固有値問題 $Av = \lambda Bv$ （AもBも対称で，Bは正定値）で固有値 λ が区間 $[a, b]$ に入る固有対の数が数百程度の場合に，それらの近似をフィルタを利用して一斉に求める.

- フィルタ $\mathcal{F}$ は必要な固有ベクトルは良く伝達するが，不要なものを殆ど伝達しない線形作用素としてうまく構成.

- ランダムなベクトルを十分多く生成し，それらを B-正規直交化したベクトルの組 X に，フィルタ $\mathcal{F}$ を適用.

- フィルタ適用後のベクトルの組は，不要な固有ベクトルの含有率が小さくなる.

- 「不要な固有ベクトル」の含有率が小さいベクトルの組から線形結合をうまく選んで，区間 $[a, b]$ にある固有値に対応する不変部分空間 $S_{[a,b]}$ の近似空間の基底を構成する.

- その基底に Rayleigh-Ritz 法を適用すると，
  必要な固有対の近似が一斉に得られる.

- 「必要な固有ベクトル」に対するフィルタ伝達率の不均一さから，近似固有対の精度の不均一さの傾向が生じる.

- ベクトルの組に対して
  「B-正規直交化に続いてフィルタを適用する」操作を
  繰り返せば，精度の不均一さを改善できる（既出）.

- 今回用いるフィルタは
  「レゾルベントの線形結合」とチェビシェフ多項式の合成.
  チェビシェフ多項式が n 次のとき，3項漸化式に従って
  「レゾルベントの線形結合」をベクトルの組に n 回適用する.

- 従来は，フィルタを適用する前のベクトルの組として
  ランダムなベクトルの組を B-正規直交化したものを使用.

- 今回は，フィルタを適用する前のベクトルの組として
  ランダムなベクトルの組に低次のフィルタを適用して
  それを B-正規直交化をしたものを使用.

フィルタの構成

- シフト ρ のレゾルベントを $\mathcal{R}(\rho) \equiv (A - \rho B)^{-1} B$ と定義.
- K 個のレゾルベント $\mathcal{R}(\rho_j)$ の線形結合を $\mathcal{Y}$ とし，
  フィルタ $\mathcal{F}$ は，$\mathcal{Y}$ の n 次チェビシェフ多項式の定数倍.

$$\begin{cases} \mathcal{Y} &= c\mathcal{I} + \displaystyle\sum_{j=1}^{K} \gamma_j \mathcal{R}(\rho_j)\,, \\ \mathcal{F} &= g_{\mathrm{S}} T_n(\mathcal{Y})\,. \end{cases} \tag{1}$$

- $\mathcal{Y}$ が実作用素となるように：
  - 定数 c は実数.
  - シフト ρ_j が実数$\Rightarrow$レゾルベントの線形結合係数 γ_j は実数.
  - 虚数のシフトは複素共役対をなし，互いに複素共役なシフトのレゾルベントの線形結合係数は互いに複素共役.

- 固有対 (λ, v) に対して $\mathcal{F}\, v = f(\lambda)\, v$ が成立.

- フィルタ $\mathcal{F}$ の伝達関数 $f(\lambda)$ は
  実有理関数 $y(\lambda)$ の n 次チェビシェフ多項式の定数 g_s 倍：

$$\begin{cases} y(\lambda) \;\equiv\; c + \displaystyle\sum_{j=1}^{K} \frac{\gamma_j}{\lambda - \rho_j}\,, \\ f(\lambda) \;\equiv\; g_{\mathrm{S}}\, T_n(\, y(\lambda)\,)\,. \end{cases} \tag{2}$$

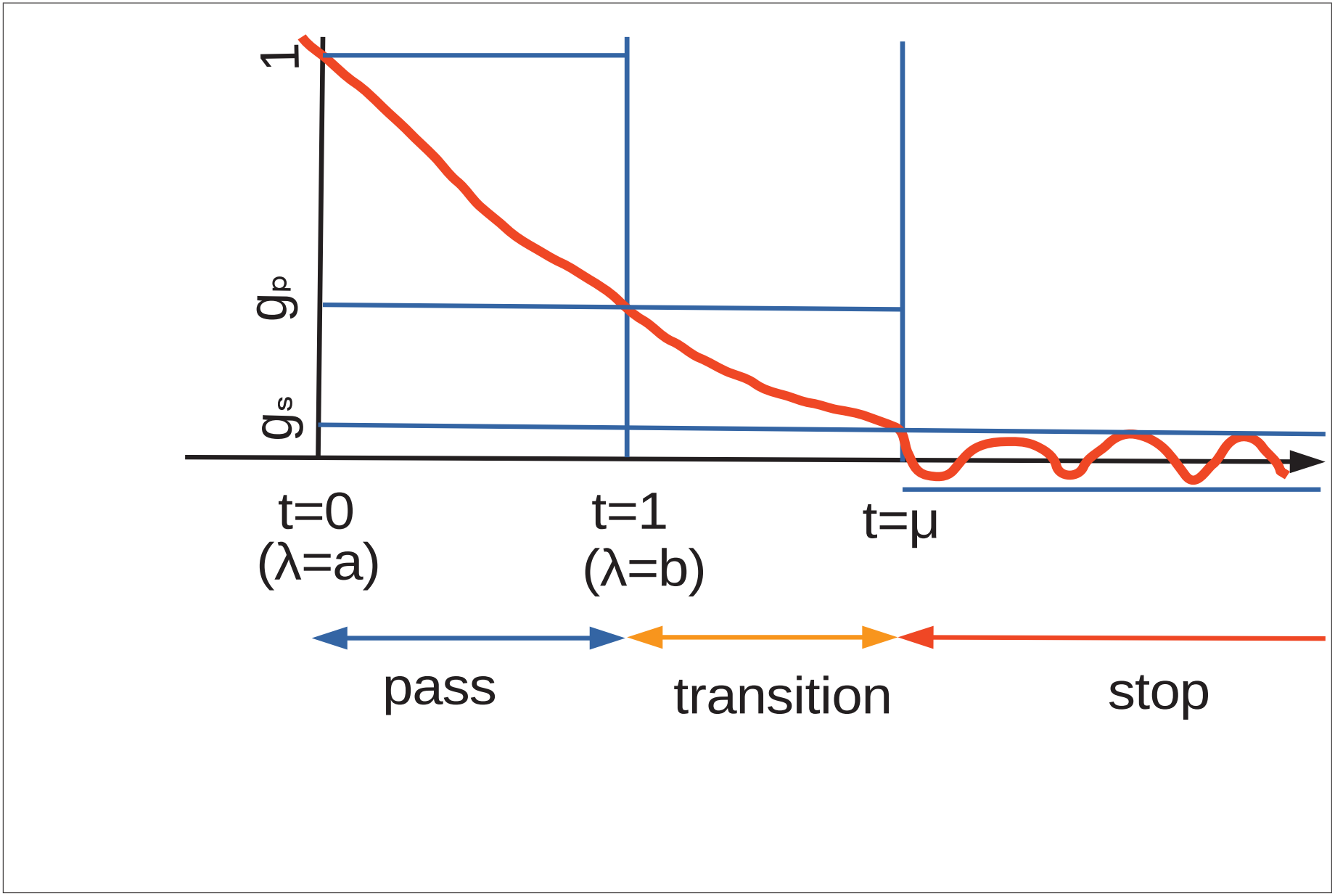
- 下端の固有対を求める場合は，区間 $[a, b]$ の幅を $\mu\,(>1)$ 倍に拡げた区間を $[a, b']$ として，

- 伝達関数 $f(\lambda)$ が以下の条件(3)を満たすように
  結合係数 γ_j とシフト ρ_j，$j = 1, 2, \ldots, K$，
  実数 c，およびチェビシェフ多項式の次数 n をうまく選んで：

$$\begin{cases} \text{通過域 } \lambda \in [a, b] & \text{では } g_{\rm p} \le f(\lambda) \le 1, \\ \text{遷移域 } \lambda \in (b, b') & \text{では } g_{\rm s} < f(\lambda) < g_{\rm p}\,, \\ \text{阻止域 } \lambda \in [b', \infty) & \text{では } |f(\lambda)| \le g_{\rm s}\,. \end{cases} \tag{3}$$

  ただし $0 < g_{\rm s} \ll g_{\rm p} < 1$.

  パラメタ μ，$g_{\rm s}$，$g_{\rm p}$ は，$f(\lambda)$ のグラフの形状の代表値.



伝達関数 $g(t) \equiv f(\lambda)$ の概形

フィルタのベクトルの組への作用

- ランダムなベクトル m 個を B-正規直交化して
  ベクトルの組 X を作成（つまり $X^T B X = I_m$ である）.
  このベクトルの組 X にフィルタ $\mathcal{F}$ を適用.

- チェビシェフ多項式は3項漸化式 (4) を満たす.

$$\begin{cases} T_0(z) \;=\; 1, \quad T_1(z) \;=\; z, \\ T_\ell(z) \;=\; 2z\,T_{\ell-1}(z) - T_{\ell-2}(z), \quad (\ell \geq 2)\,. \end{cases} \tag{4}$$

- 組 X へのフィルタ $\mathcal{F}$ の作用 $\mathcal{F}X = g_{\rm S}\,T_n(\mathcal{Y})X$ の計算は,
  $X^{(j)} \equiv T_j(\mathcal{Y})X$ とおくと，$\mathcal{F}X = g_{\rm S}\,X^{(n)}$.
  $X^{(j)}$ は漸化式 (5) で求める.

$$\begin{cases} X^{(0)} \;=\; X, \quad X^{(1)} \;=\; \mathcal{Y}\,X, \\ X^{(\ell)} \;=\; 2\mathcal{Y}\,X^{(\ell-1)} - X^{(\ell-2)}, \ (\ell \geq 2). \end{cases} \tag{5}$$

フィルタがレゾルベント1つのチェビシェフ多項式の伝達関数は

$$\begin{cases} f(\lambda) \equiv g_{\mathrm{S}}\, T_n(y(\lambda)), \\ y(\lambda) \equiv \dfrac{\gamma}{\lambda - \rho} + c\,. \end{cases}$$

固有値 λ の相対座標 t は，$\lambda \in [a, b]$ と $t \in [0, 1]$ の線形変換.
そうして t を引数とする 伝達関数 $g(t) \equiv f(\lambda)$ を定義.

レゾルベント1つのフィルタの例：次数nと3つの形状パラメタ

フィルタ名	F1-1	F1-2	F1-3	F1-4	F1-5
次数n	27	28	33	25	109
μ	2.0	1.75	1.5	1.5	1.3
g_s	1E-12	1E-15	1E-15	1E-16	1E-15
g_p	1E-4	1E-6	1E-7	1E-8	1E-8

g_p が小さいほど「必要な固有対」の精度が不均一になる傾向.
※ μが2.0や1.75の場合は，ベクトルの数が110では不足気味.

次数の低いフィルタによる前処理

次数の低いフィルタによる前処理

フィルタ $\mathcal{F}$ が，レゾルベントの線形結合 $\mathcal{Y}$ の n 次チェビシェフ多項式の場合：

$$\begin{cases} \mathcal{Y} &= c\mathcal{I} + \displaystyle\sum_{j=1}^{K} c_j \mathcal{R}(\rho_j)\,, \\ \mathcal{F} &= g_s T_n(\mathcal{Y})\,. \end{cases}$$

フィルタ $\mathcal{F}$ を B-正規直交化操作をはさんで2回適用する場合：

$$\begin{cases} X &\leftarrow \text{ランダムな}B\text{−正規直交ベクトルm個の組；} \\ Z &\leftarrow g_s T_n(\mathcal{Y})\, X\text{；}\quad 1\text{回目のn次のフィルタ} \\ X &\leftarrow Z\text{の}B\text{−正規直交化；} \\ Z &\leftarrow g_s T_n(\mathcal{Y})\, X\,.\quad 2\text{回目のn次のフィルタ} \end{cases}$$

作用素 $\mathcal{Y}$ の適用回数は2倍に増えて $2n$ になる.

計算量を2倍にまではせずに，ある程度の精度向上を得たい.

チェビシェフの次数 ν（$\nu < n$）のフィルタを適用して，
B-正規直交化をした後に，本来の次数 nのフィルタを適用すると，
レゾルベントの線形結合 $\mathcal{Y}$ の適用回数は $\nu + n$.

$$
\begin{cases}
X & \leftarrow & \text{ランダムな}B-\text{正規直交ベクトルm個の組；} \\
Z & \leftarrow & \widetilde{g}_s\, T_\nu(\mathcal{Y})\, X\text{；}\quad \text{前処理の}\nu\text{次のフィルタ} \\
X & \leftarrow & Z\text{の}B-\text{正規直交化；} \\
Z & \leftarrow & g_s\, T_n(\mathcal{Y})\, X\,. \quad \text{本来の}n\text{次のフィルタ}
\end{cases}
$$

$\mathcal{Y}$ は $T_\nu(\mathcal{Y})$ と $T_n(\mathcal{Y})$ で共通 $\Rightarrow$ 同じレゾルベントを利用できる.

$$
\text{ここで}\begin{cases}
\widetilde{g}_s & \equiv & 1/\cosh\left(\,\frac{\nu}{n}\,\cosh^{-1}\frac{1}{g_s}\,\right), \\
\widetilde{g}_p & \equiv & \widetilde{g}_s \cosh\left(\,\frac{\nu}{n}\,\cosh^{-1}\frac{g_p}{g_s}\,\right).
\end{cases}
$$

$\widetilde{g}_s$ と $\widetilde{g}_p$ はそれぞれ，ν次のフィルタの伝達率の最大値を1に規格化する定数および通過域での伝達率の最小値.

実 験 環 境

- 計算機：東大情報基盤センター Oakbridge-CX の1ノード.
  ノードあたりのCPUの数は2つ.
- CPUはIntel Xeon Platinum 8280(CascadeLake)(2.7GHz).
  コア数：28/CPU（56/ノード）.
  L3キャッシュ容量：38.5MiB/CPU.
  拡張命令セットのレベル：AVX-512.
- 理論ピーク性能：4.84TFLOPS(DP)/ノード.
  主記憶DDR4メモリの容量：192GiB/ノード,
  ノードのメモリバンド幅：281.6GB/s.
- Fortranはintel ifort，バージョン "19.1.3.304 20200925".
- 計算は1ノード（2CPU）で，OpenMPによりノード内のコア数に等しい56スレッドで並列実行.

固有値問題の例題と，近似固有対の精度の評価法

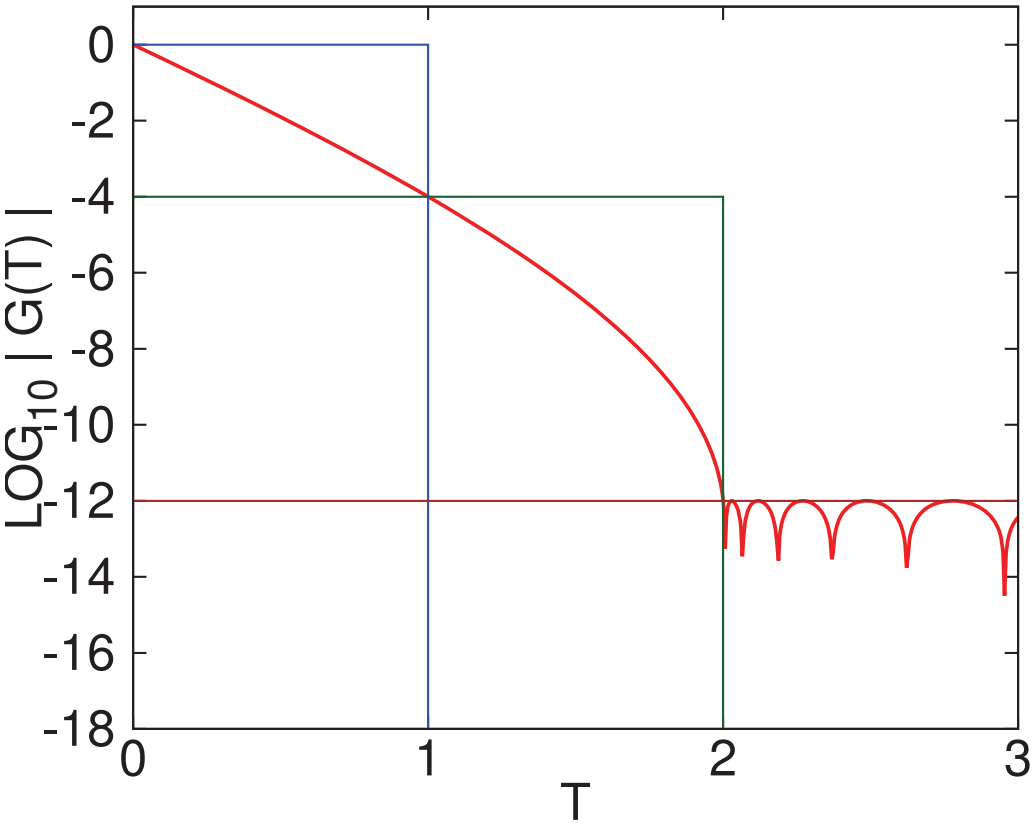
- **1辺の長さ** π **の立方体領域で零境界条件を課したラプラシアンのFEM離散化から生じる一般固有値問題** $Av = \lambda Bv$**.**
- **立方体を各辺方向に** $n_1{+}1$**，** $n_2{+}1$**，** $n_3{+}1$ **に等分した直方体がFEMの要素で，要素内の基底関数は3重線形関数.**
- **要素分割は** $(n_1, n_2, n_3) = (40, 50, 60)$ **.**
  行列 A **と** B **は次数** $N = 120,000$ **（12万）で下帯幅** $w_L = 2,041$**.**
- **固有値が下端付近の区間** $[3,\ 30]$ **にある** 54**対の固有対を近似.**
  数値と演算には倍精度（IEEE 754，binary64）を使用.
- **最初に与えるランダムなベクトルの数は** 110 **個.**
- **近似固有対** $(\lambda,\ v)$ **の精度を「相対残差の大きさ」**Θ **で評価.**

$$\Theta \ \equiv \ \frac{||Av - \lambda Bv||_2}{||\lambda Bv||_2} \, .$$

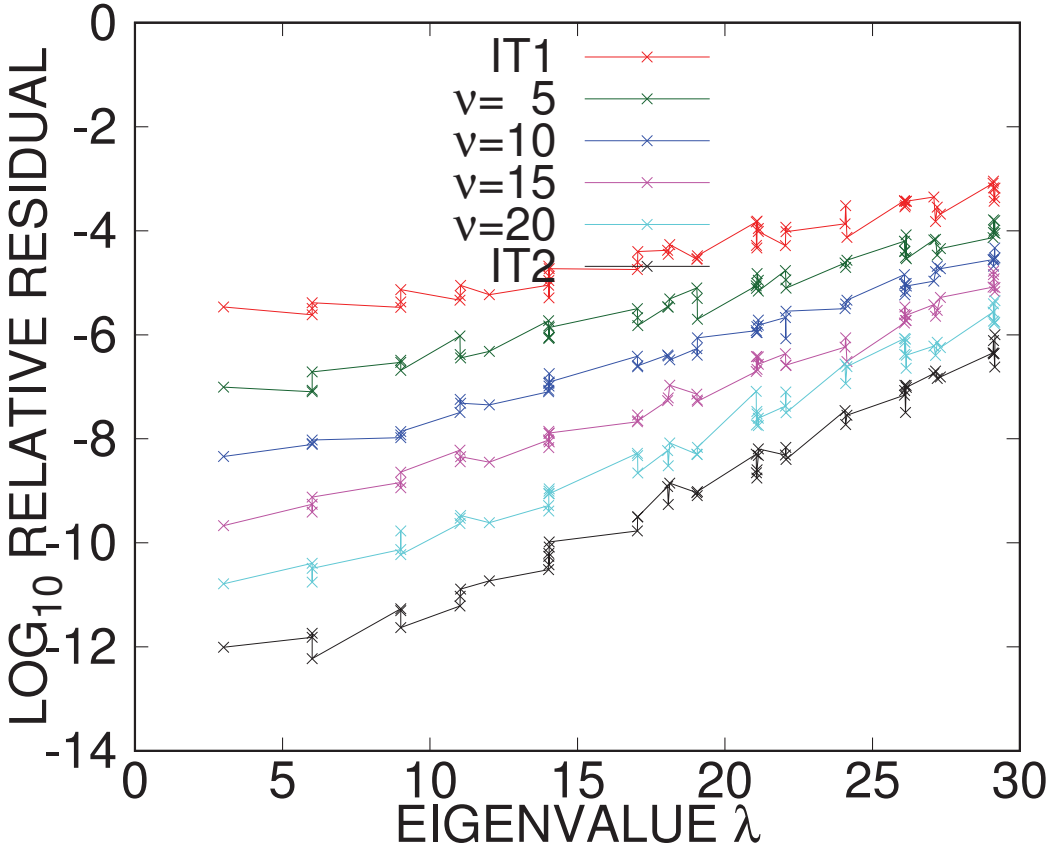
次数の低いフィルタによる前処理の実験例

実数シフトのレゾルベントを1つ用いたフィルタによる例

フィルタF1-1　（$n = 27$，$\mu = 2.0$，g_s =1E-12，g_p =1E-4 ）

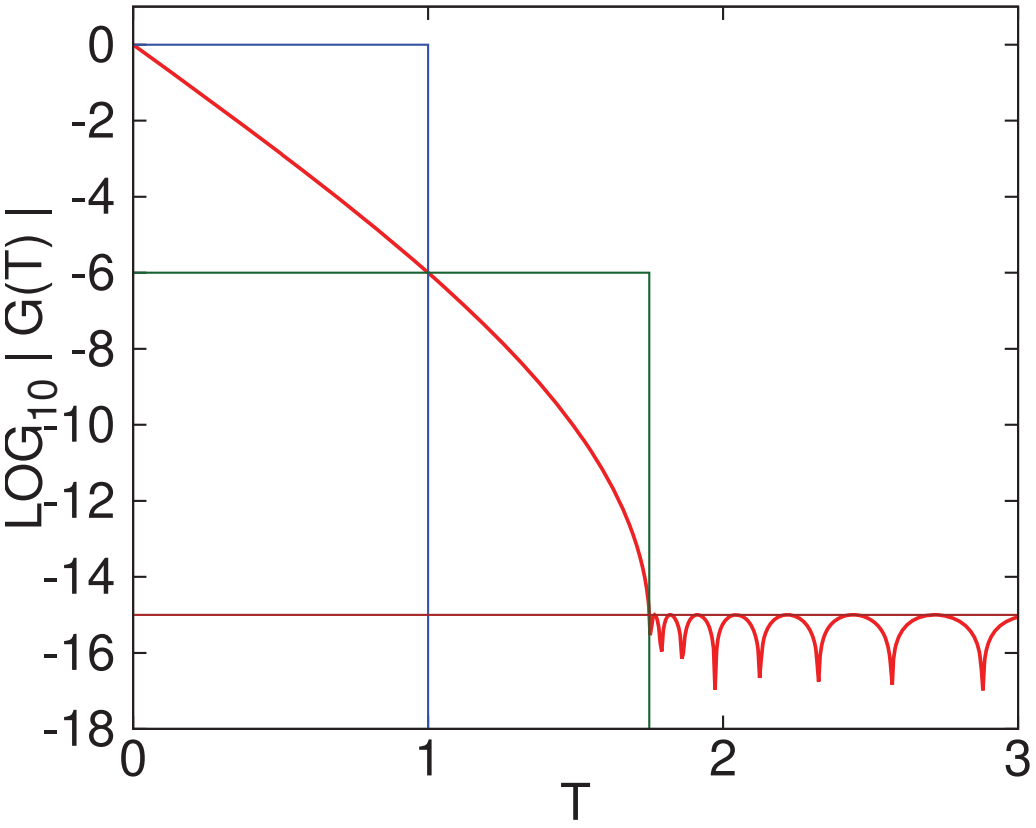


伝達関数の大きさ $|g(t)|$ の対数

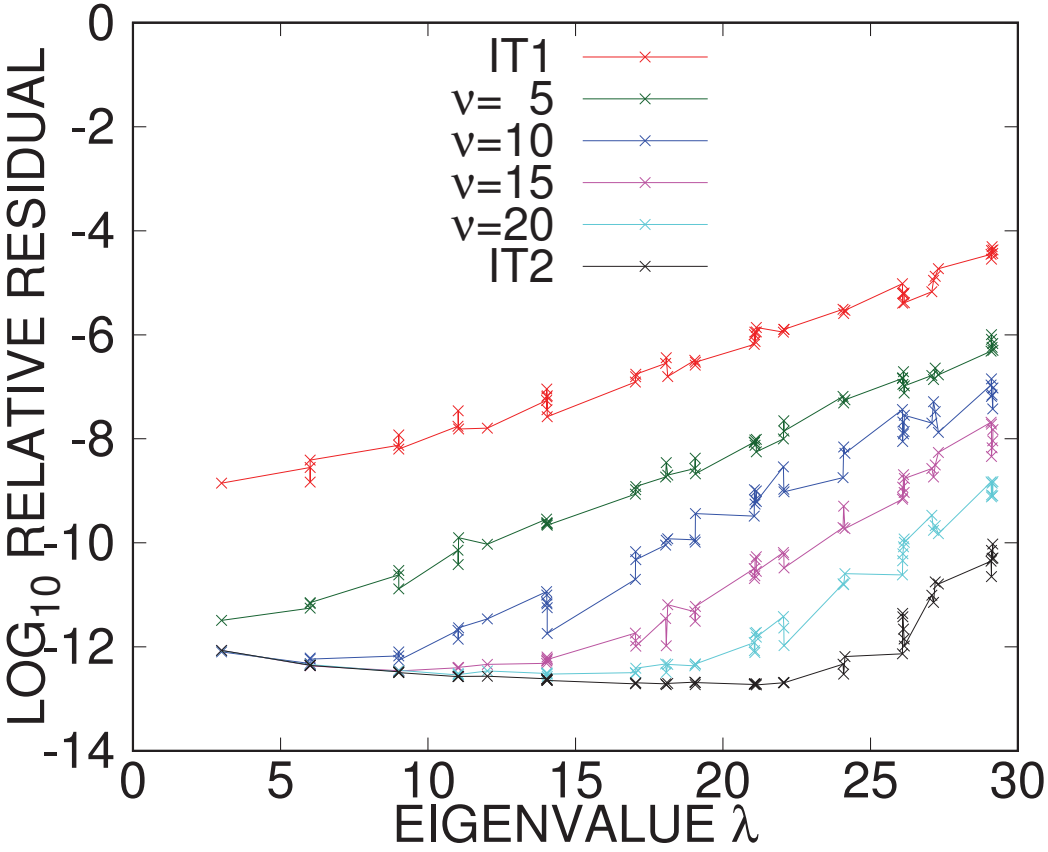


相対残差の大きさ Θ の対数

フィルタF1-2　($n = 28$, $\mu = 1.75$, g_s =1E-15, g_p =1E-6)

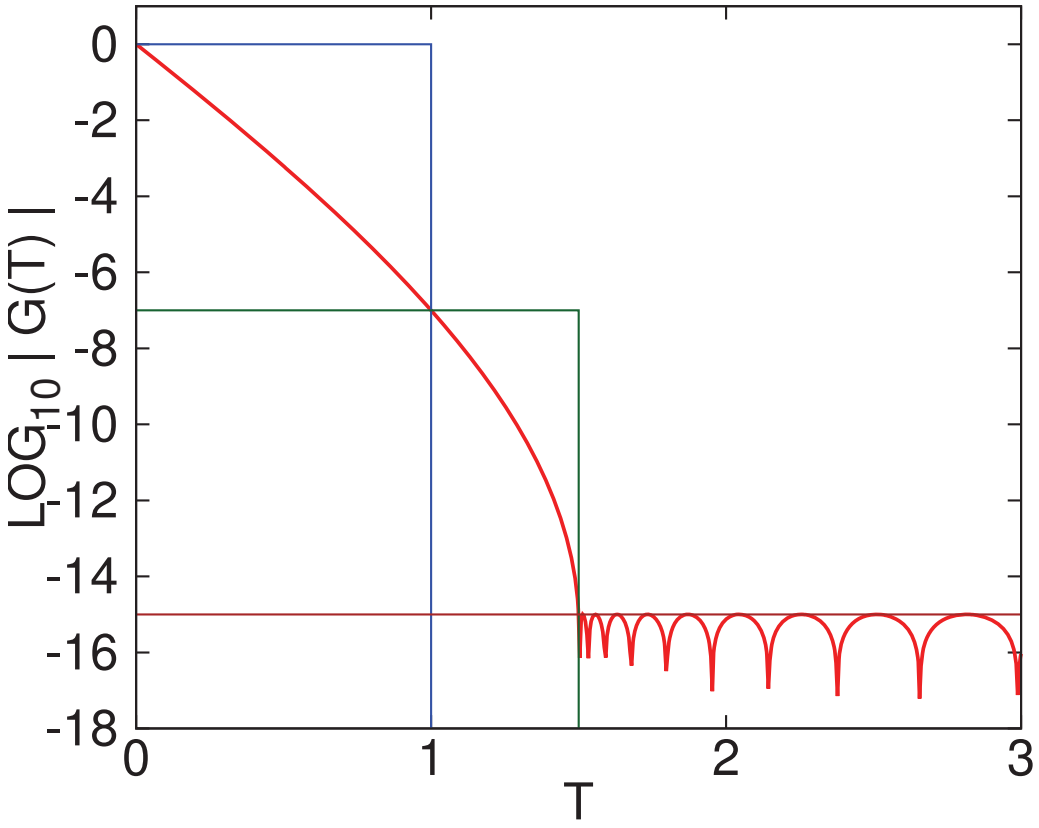


伝達関数の大きさ $|g(t)|$ の対数

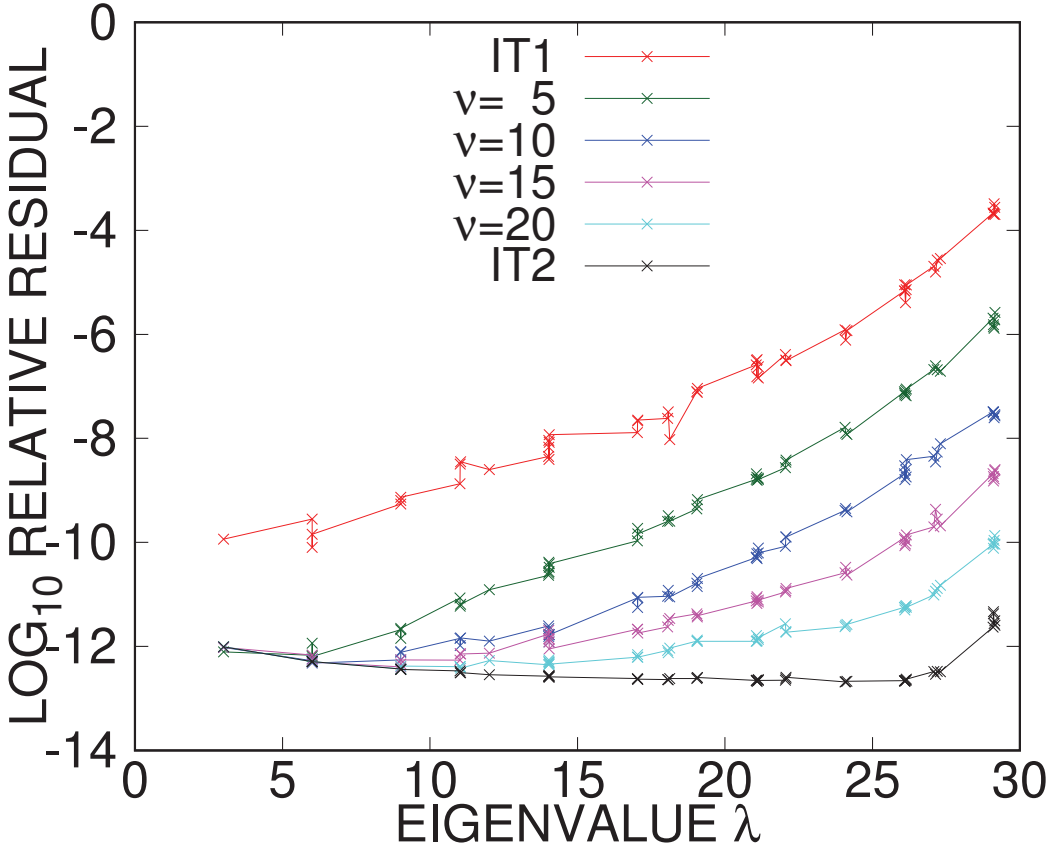


相対残差の大きさ Θ の対数

フィルタF1-3　($n = 33$, $\mu = 1.5$, g_s =1E-15, g_p =1E-7)

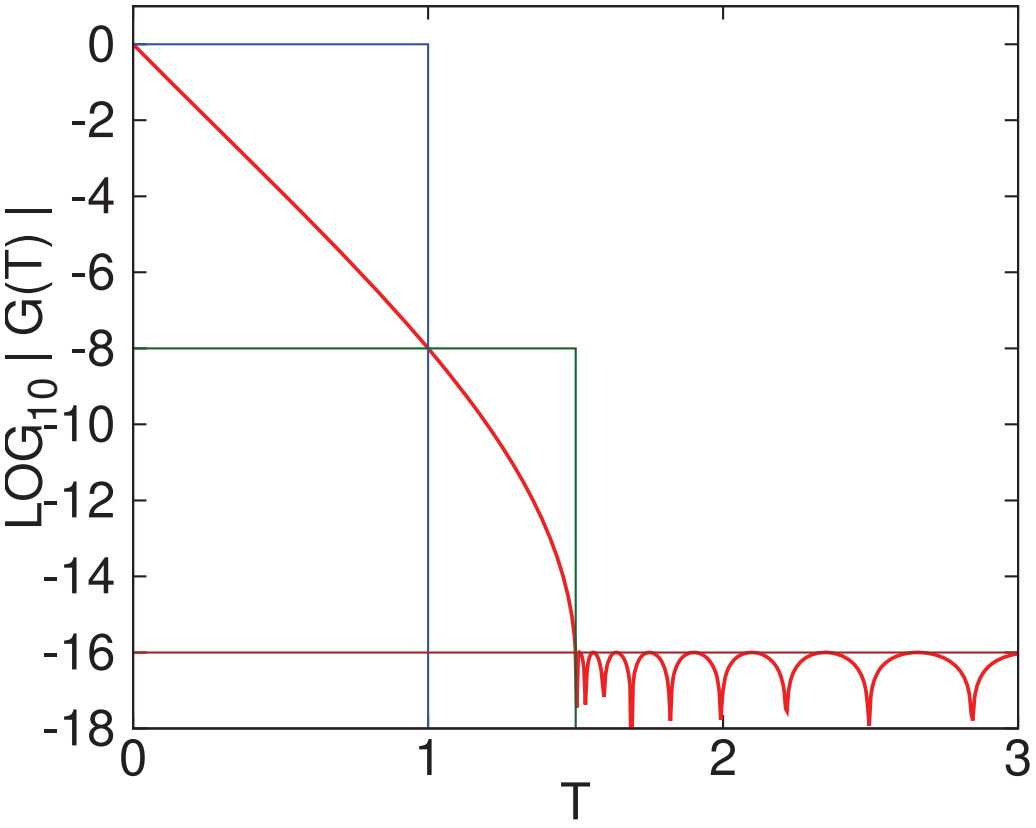


伝達関数の大きさ $|g(t)|$ の対数

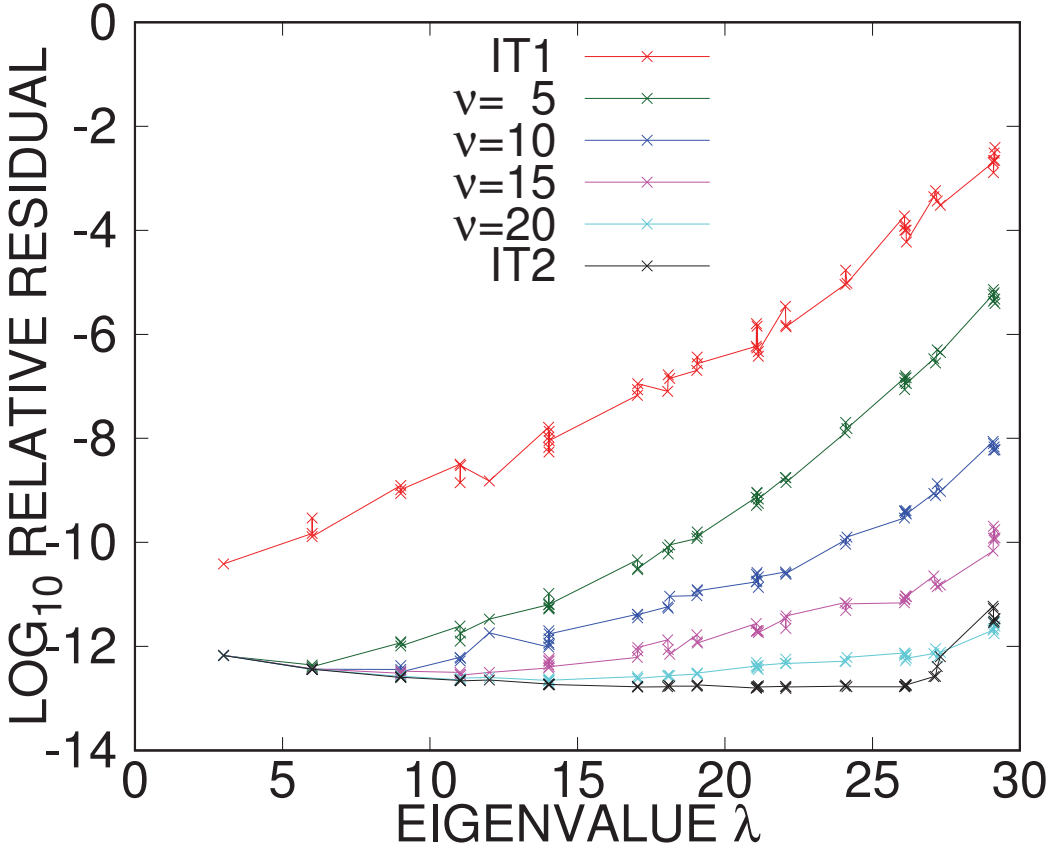


相対残差の大きさ Θ の対数

フィルタF1-4　（$n = 25$，$\mu = 1.5$，g_s =1E-16，g_p =1E-8 ）

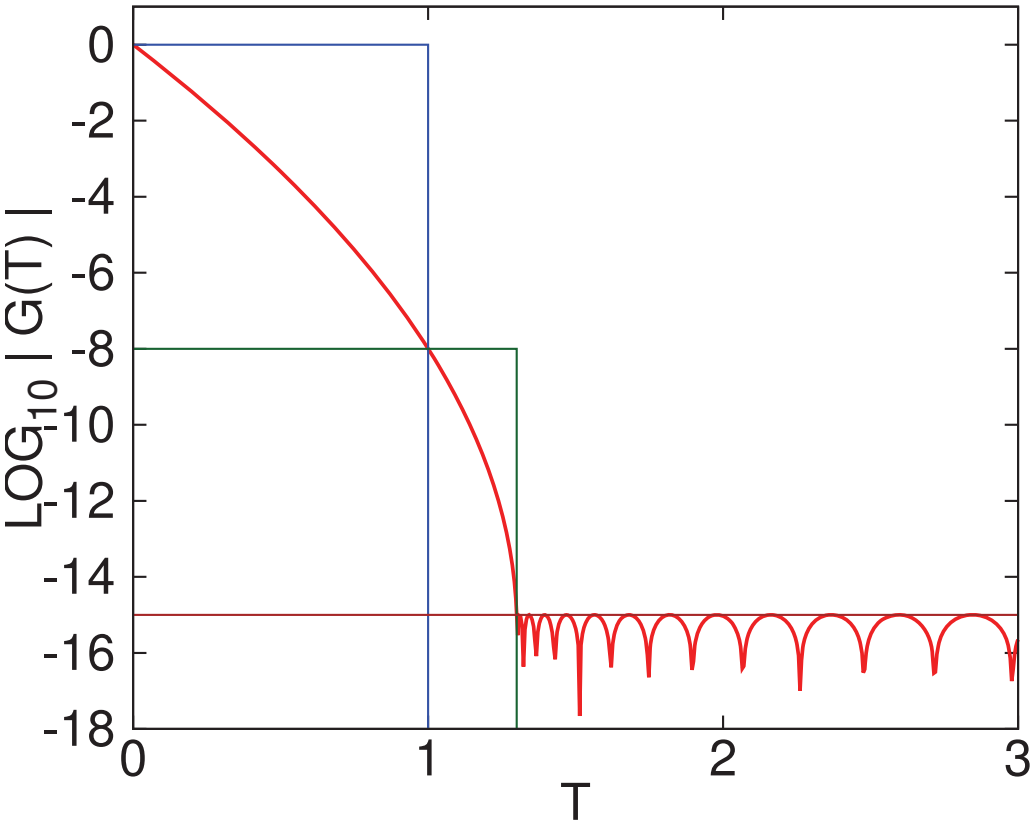


伝達関数の大きさ $|g(t)|$ の対数

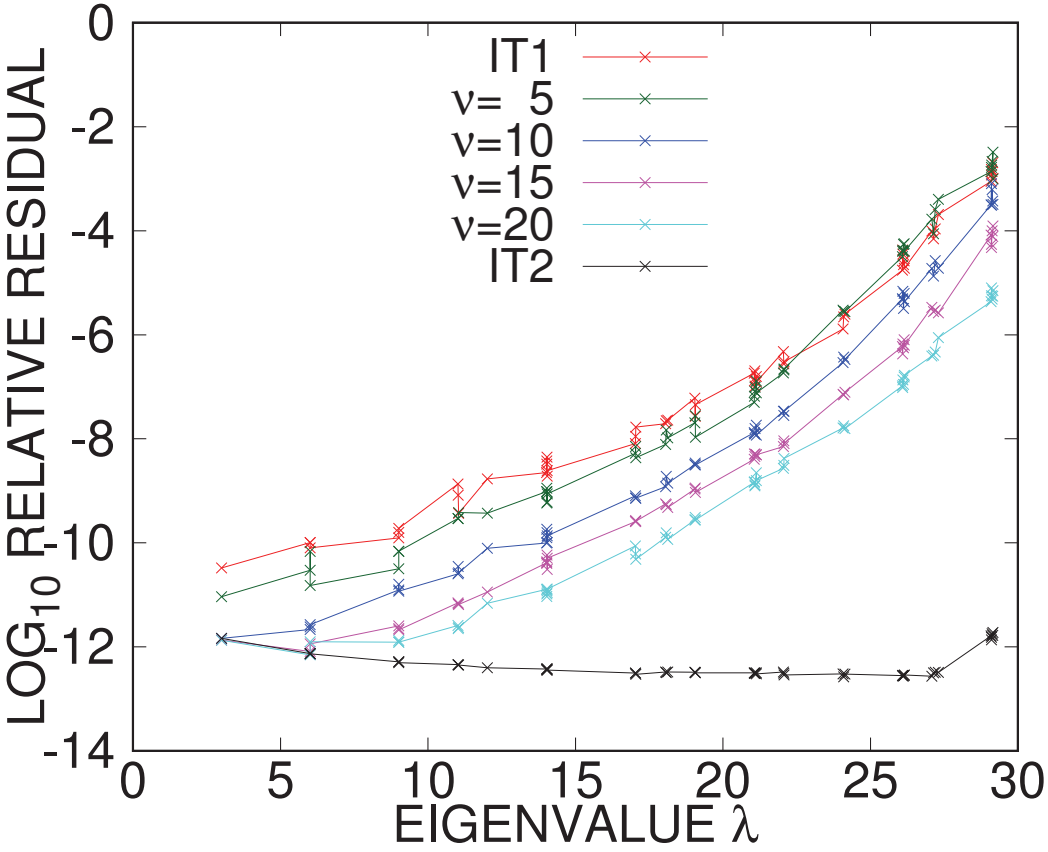


相対残差の大きさ Θ の対数

フィルタ F1-5　（$n = 109$, $\mu = 1.3$, g_s =1E-15, g_p =1E-8 ）



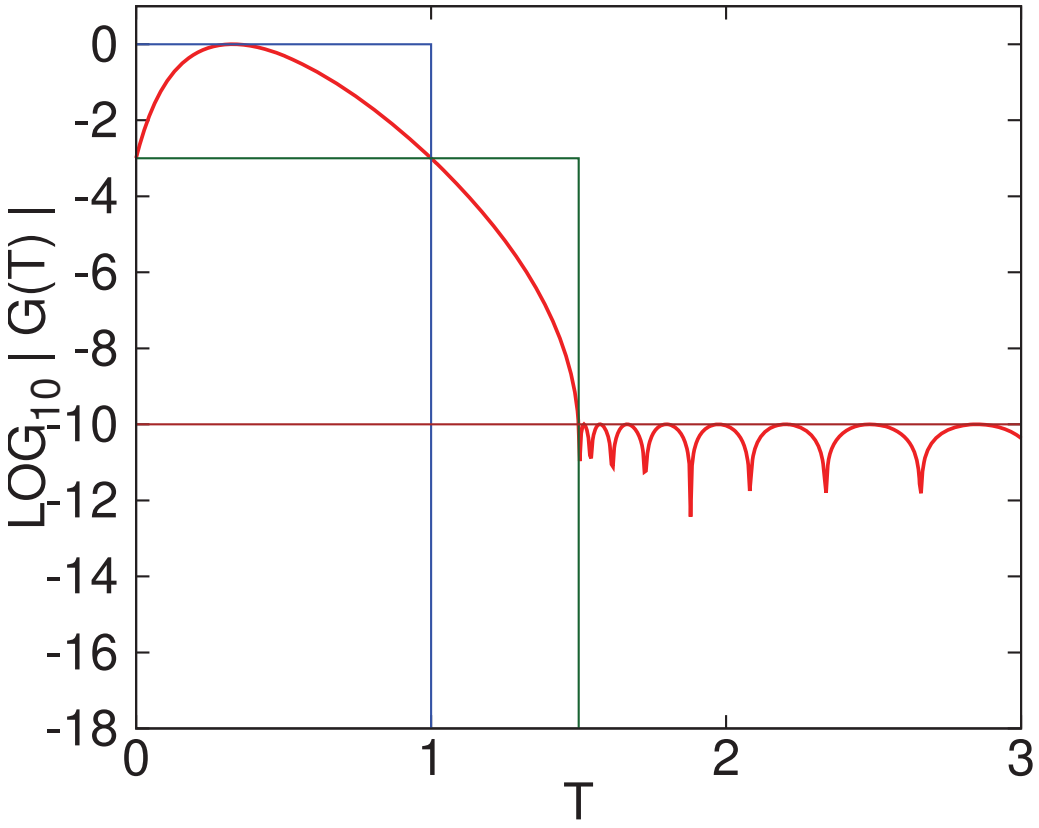
伝達関数の大きさ $|g(t)|$ の対数



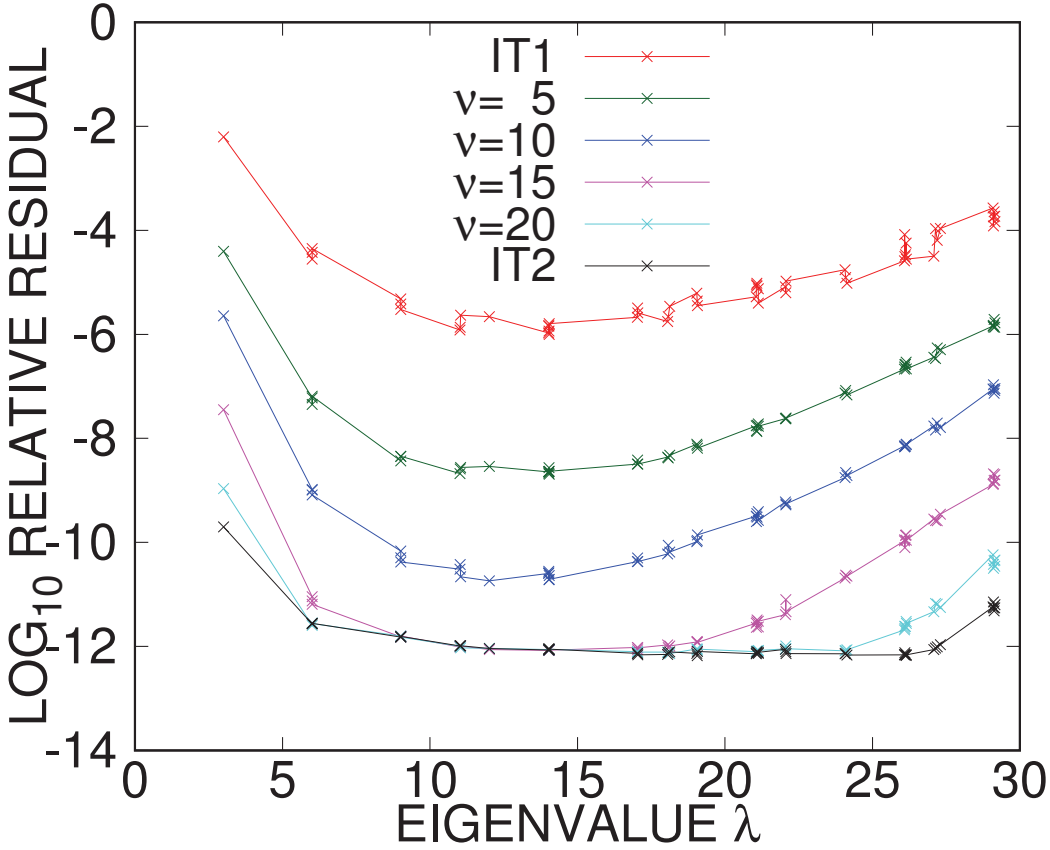
相対残差の大きさ Θ の対数

実数シフトのレゾルベントを2つ用いたフィルタによる例

フィルタ F2-1　（$n = 23$，$\mu = 1.5$，g_s =1E-10，g_p =1E-3 ）

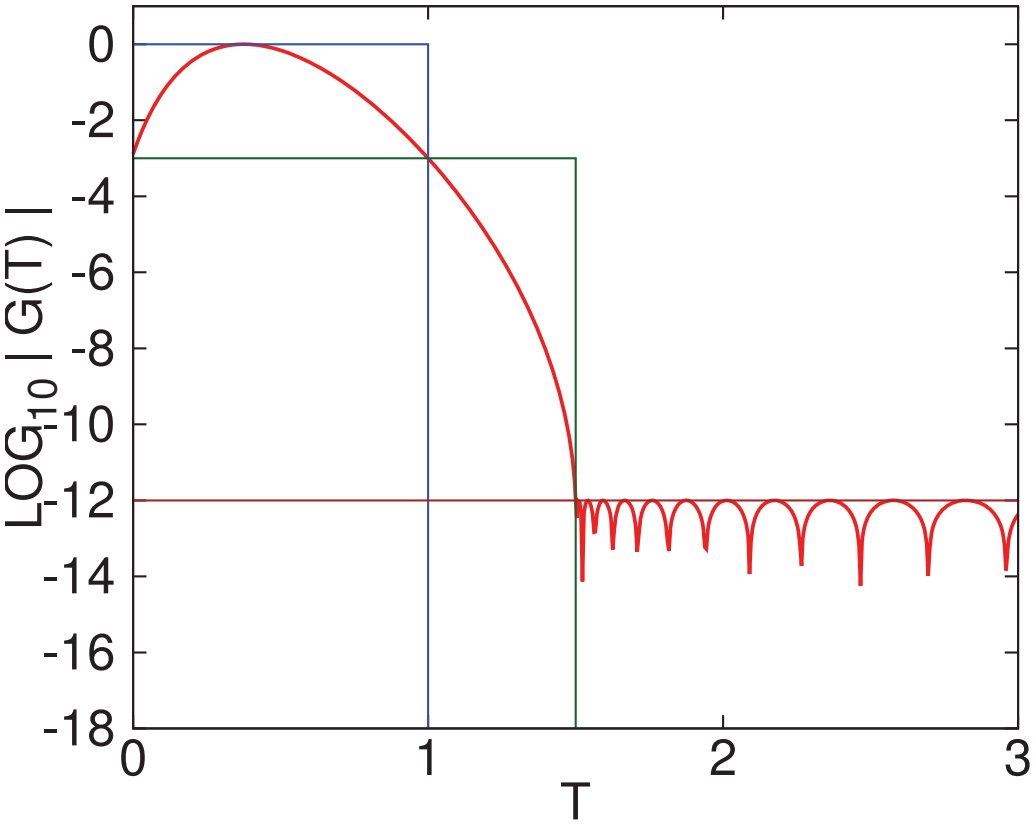


伝達関数の大きさ $|g(t)|$ の対数

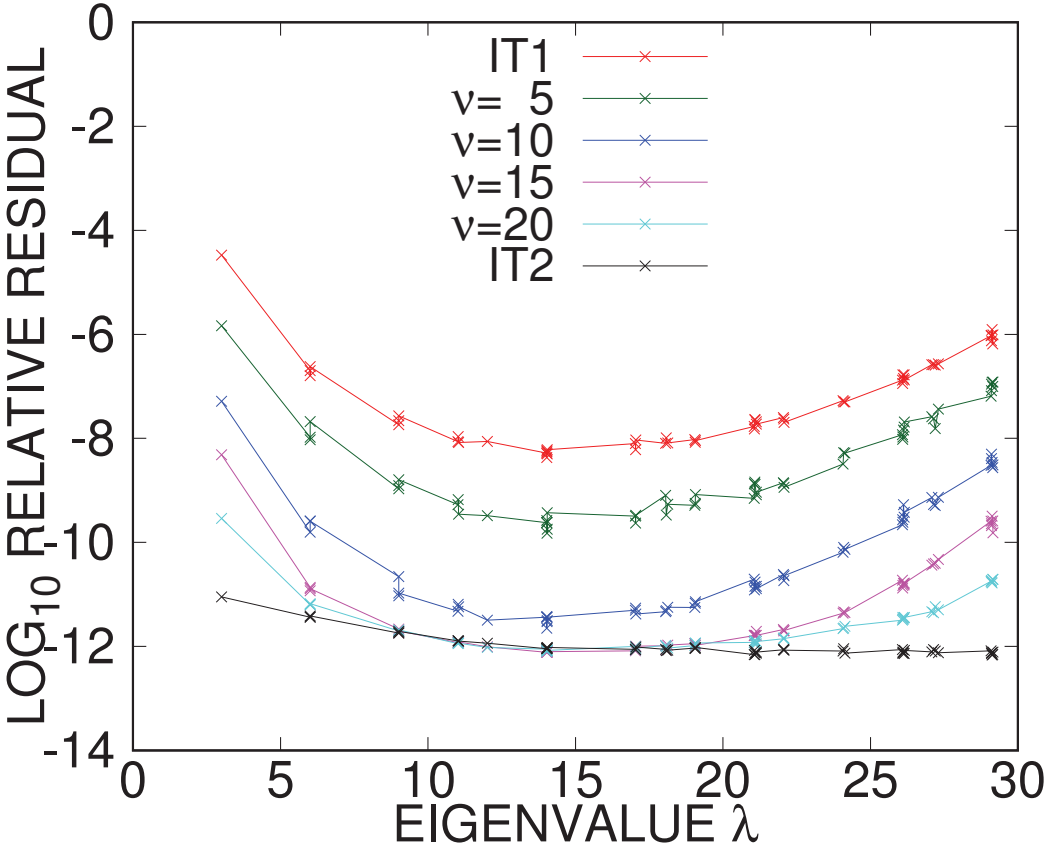


相対残差の大きさ Θ の対数

フィルタ F2-2　($n = 38$, $\mu = 1.5$, g_s =1E-12, g_p =1E-3)

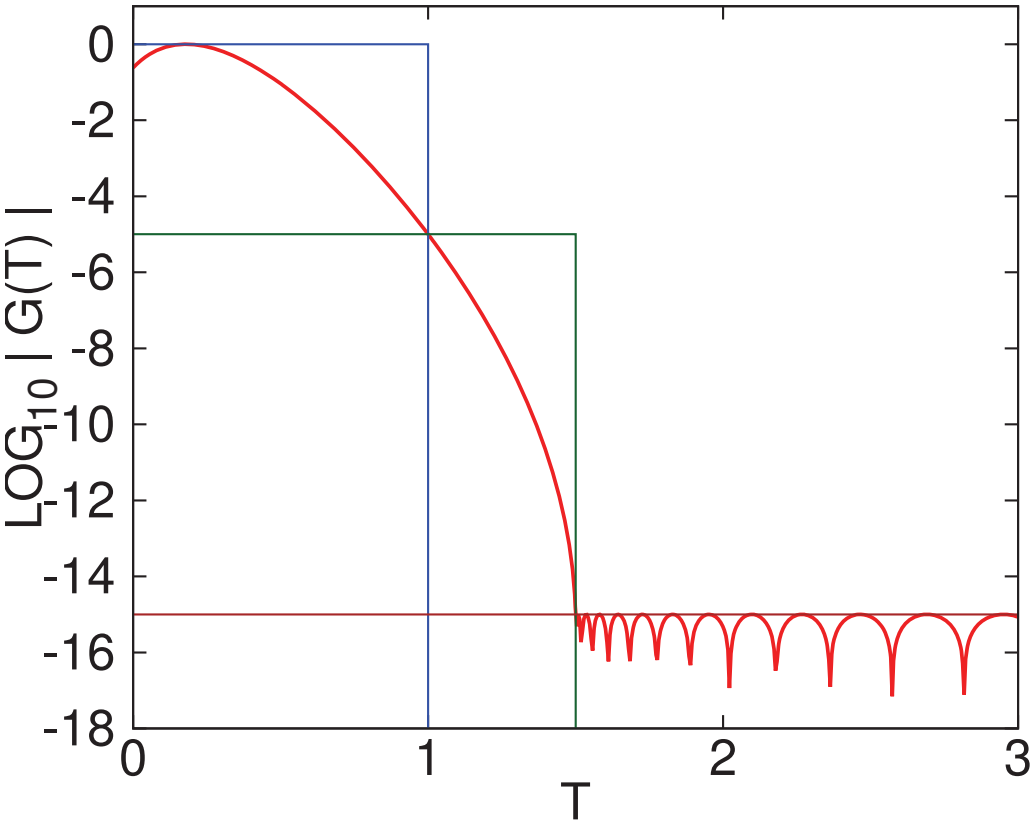


伝達関数の大きさ $|g(t)|$ の対数

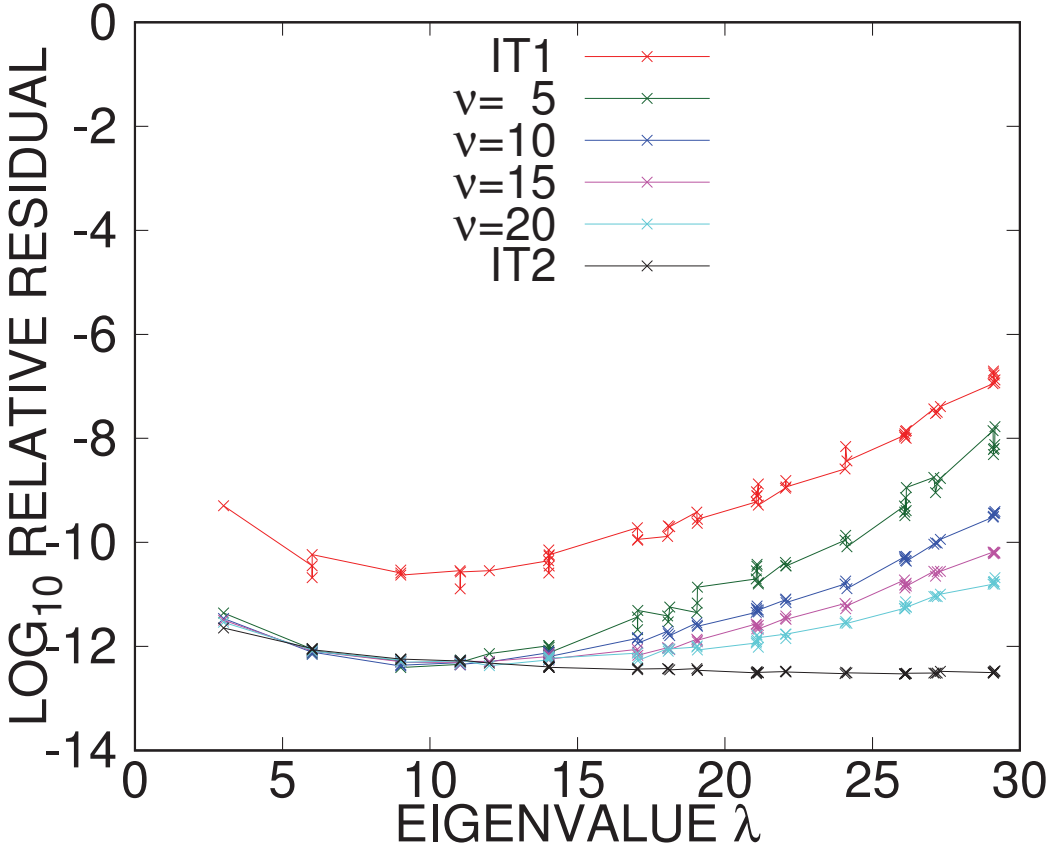


相対残差の大きさ Θ の対数

フィルタ F2-3　（$n = 38$，$\mu = 1.5$，g_s =1E-15，g_p =1E-5 ）

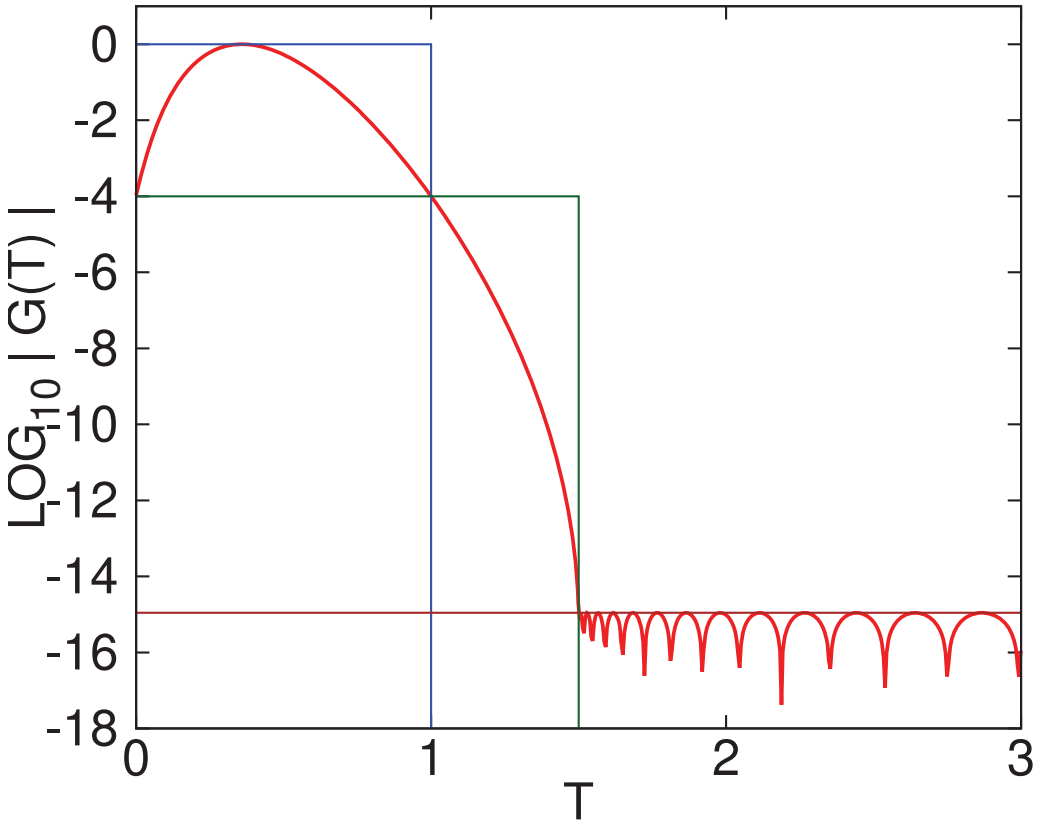


伝達関数の大きさ $|g(t)|$ の対数

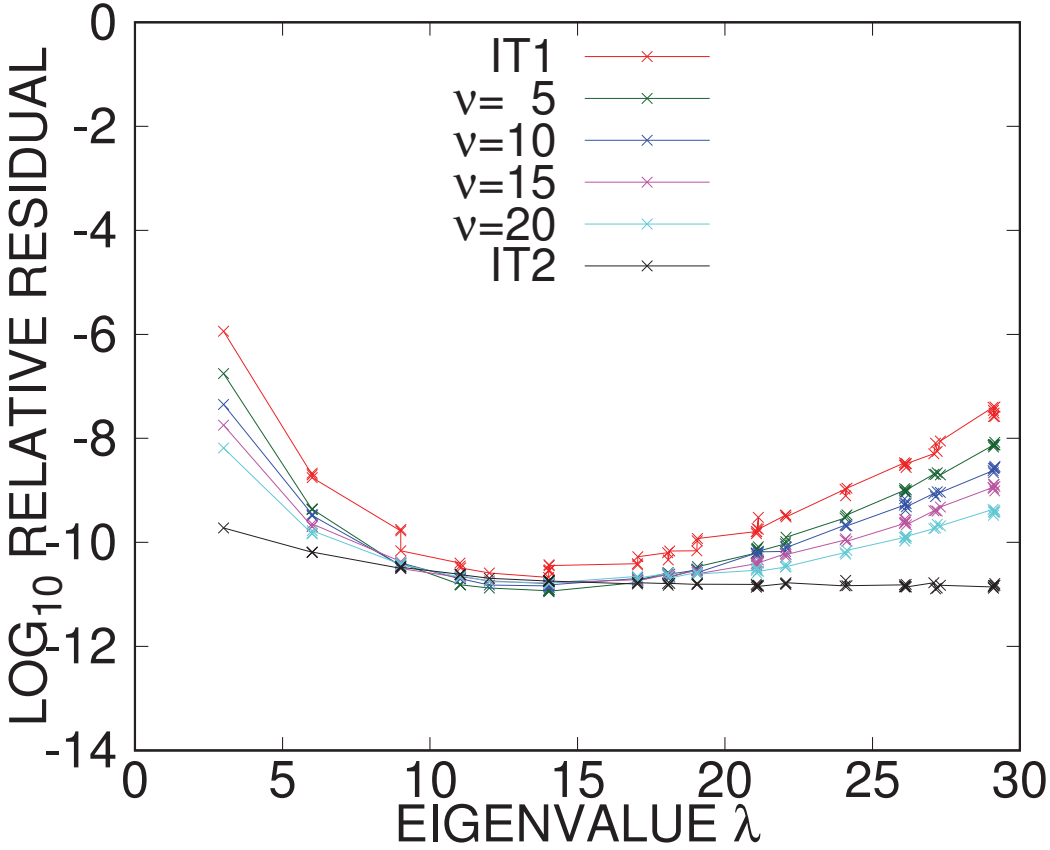


相対残差の大きさ Θ の対数

フィルタ F2-4　($n = 40$, $\mu = 1.5$, g_s =1.1E-15, g_p =1E-4)



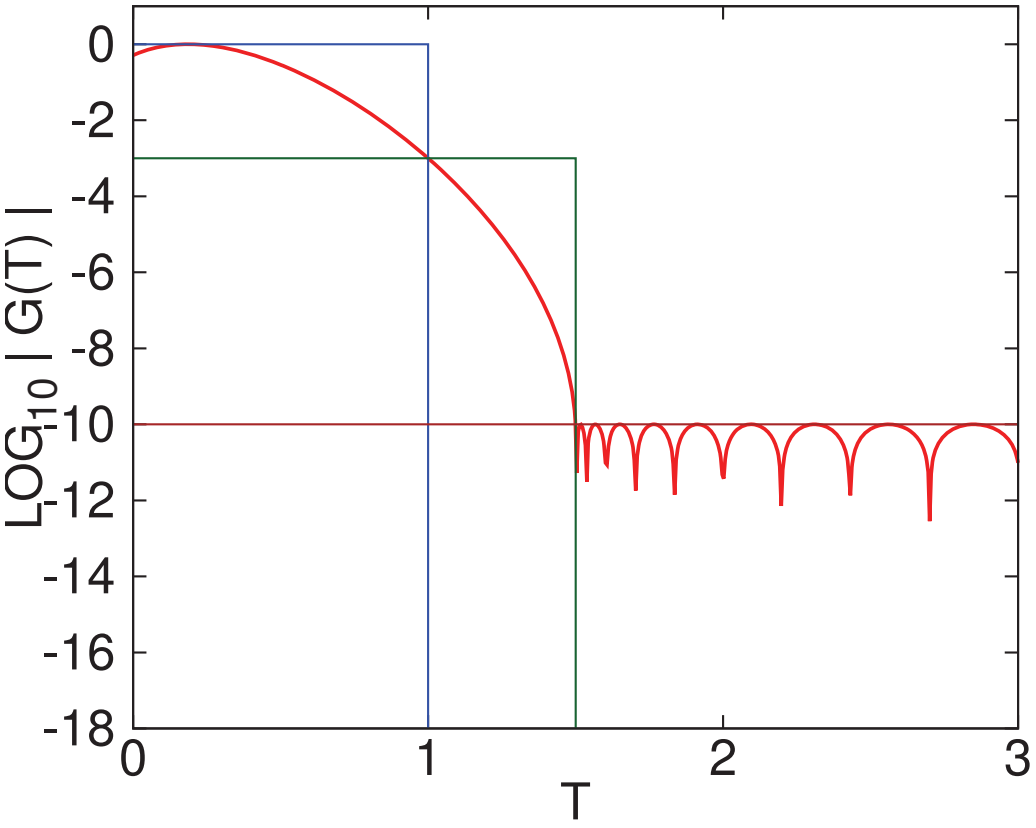
伝達関数の大きさ $|g(t)|$ の対数



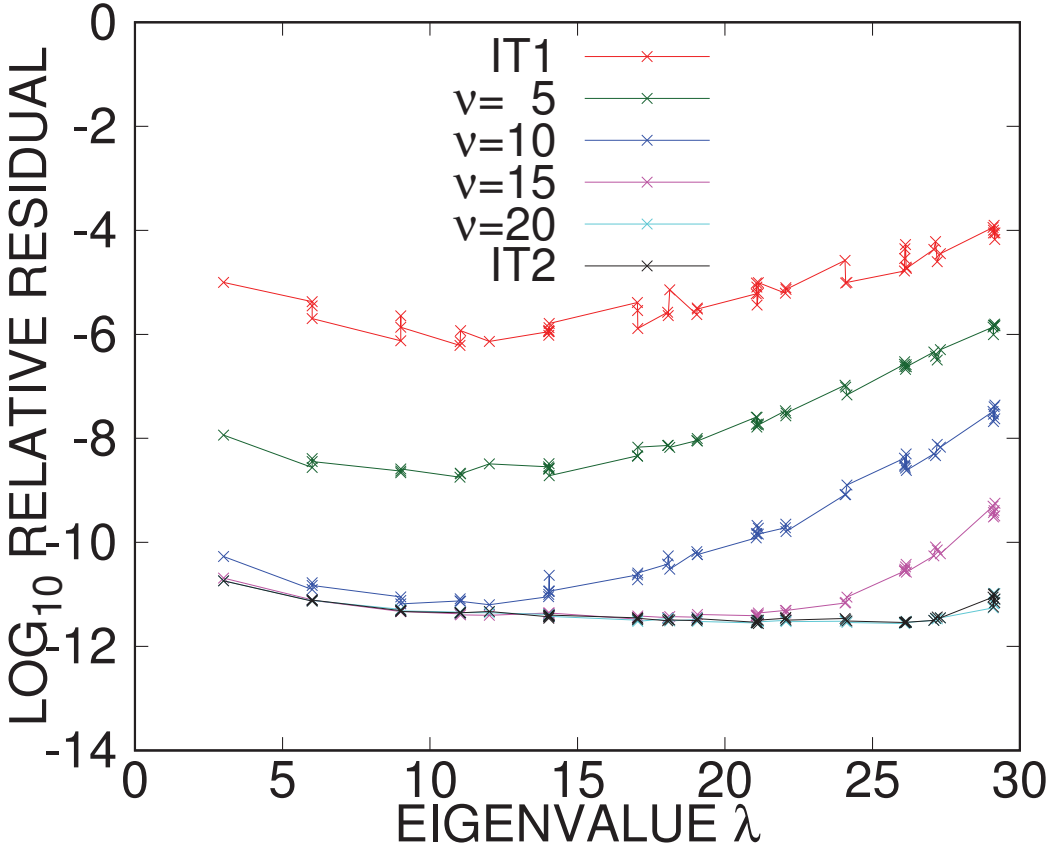
相対残差の大きさ Θ の対数

実数シフトのレゾルベントを3つ用いたフィルタによる例

フィルタ F3-I-1　($n = 20$, $\mu = 1.5$, g_s =1.1E-10, g_p =1E-3)

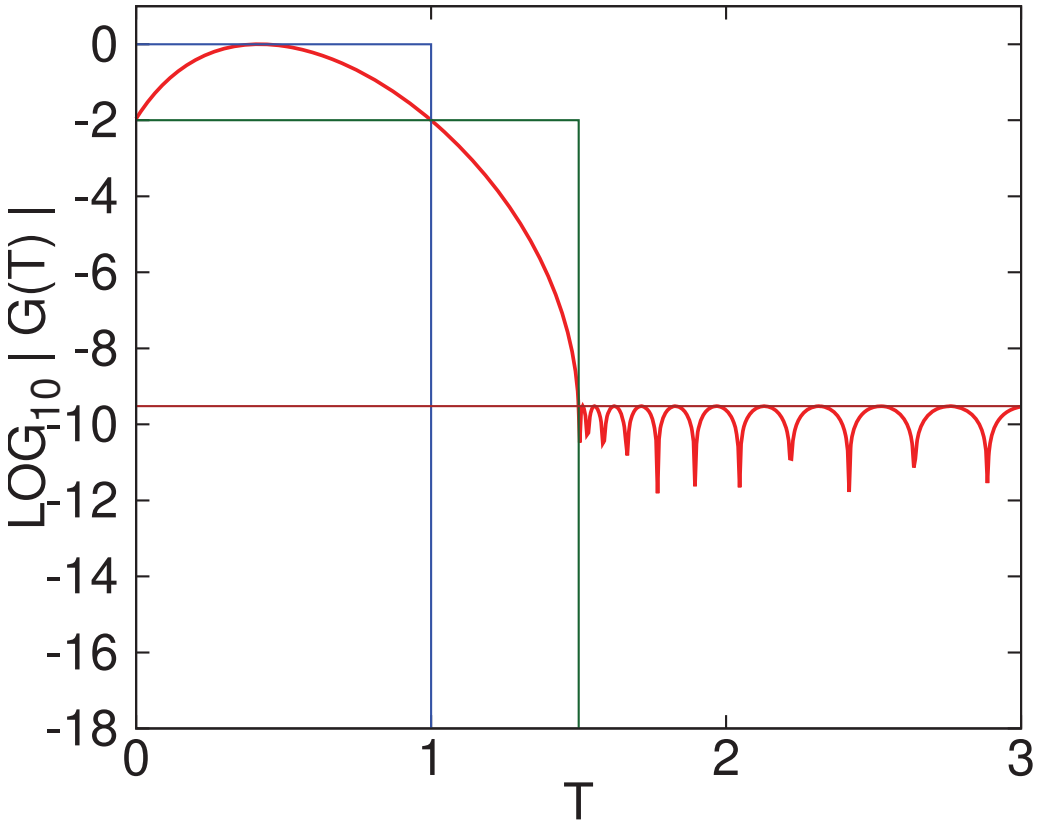


伝達関数の大きさ $|g(t)|$ の対数

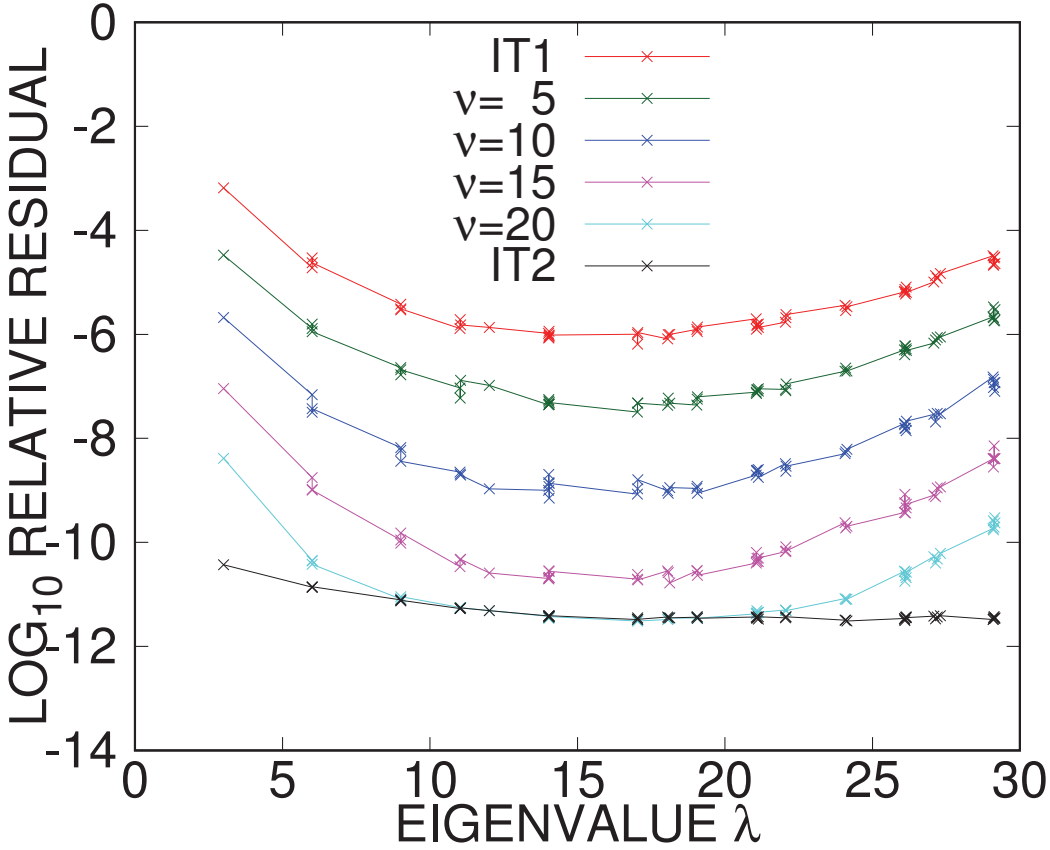


相対残差の大きさ Θ の対数

フィルタ F3-I-2　（$n = 30$，$\mu = 1.5$，g_s =3E-10，g_p =1E-2 ）

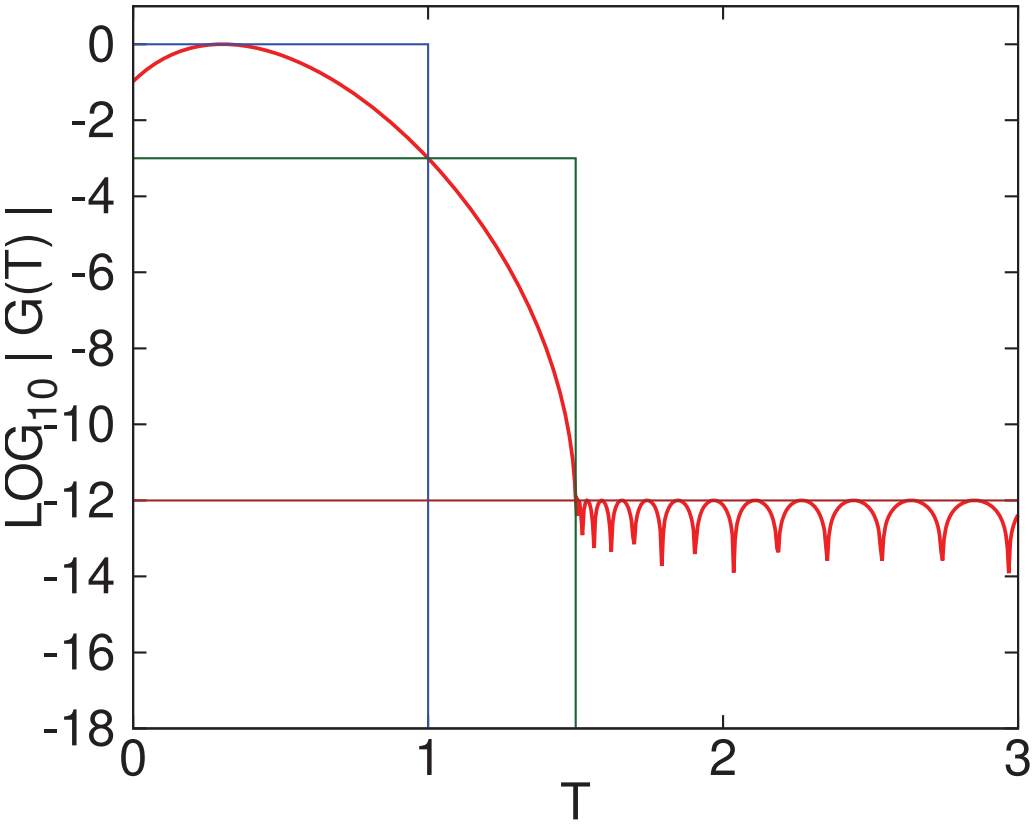


伝達関数の大きさ $|g(t)|$ の対数

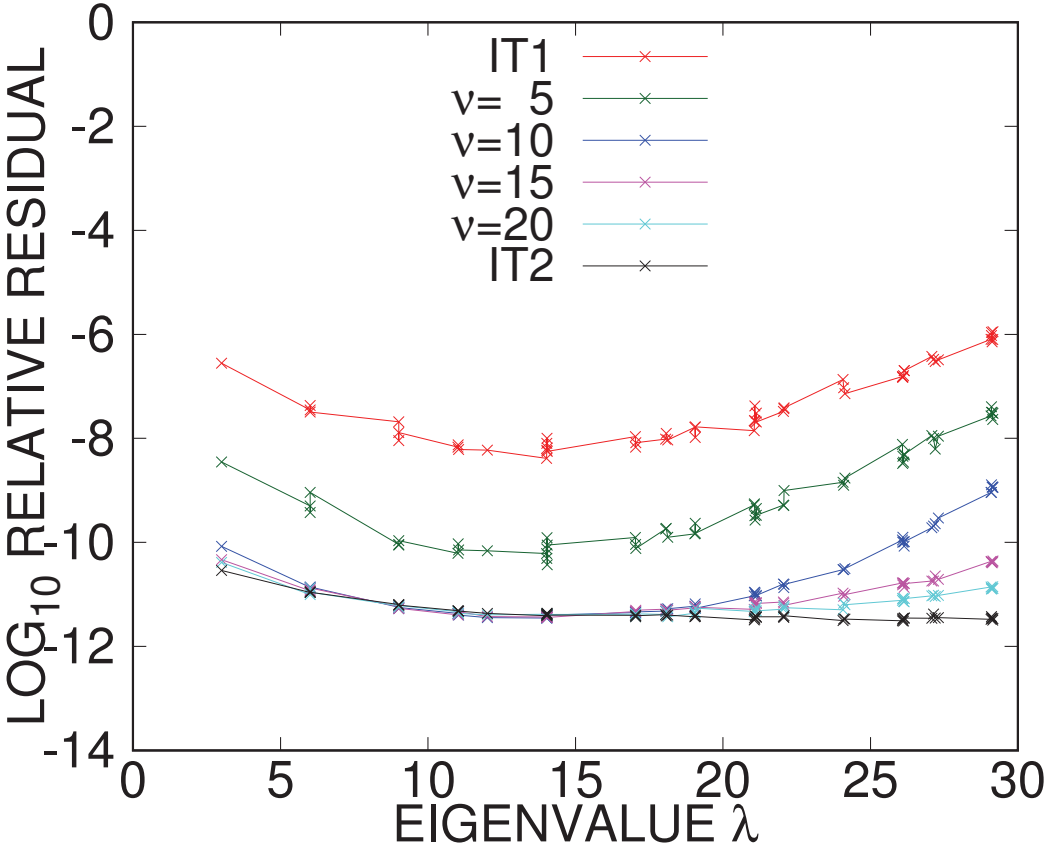


相対残差の大きさ Θ の対数

フィルタ F3-I-3　（$n = 30$，$\mu = 1.5$，g_s =3E-12，g_p =1E-3 ）

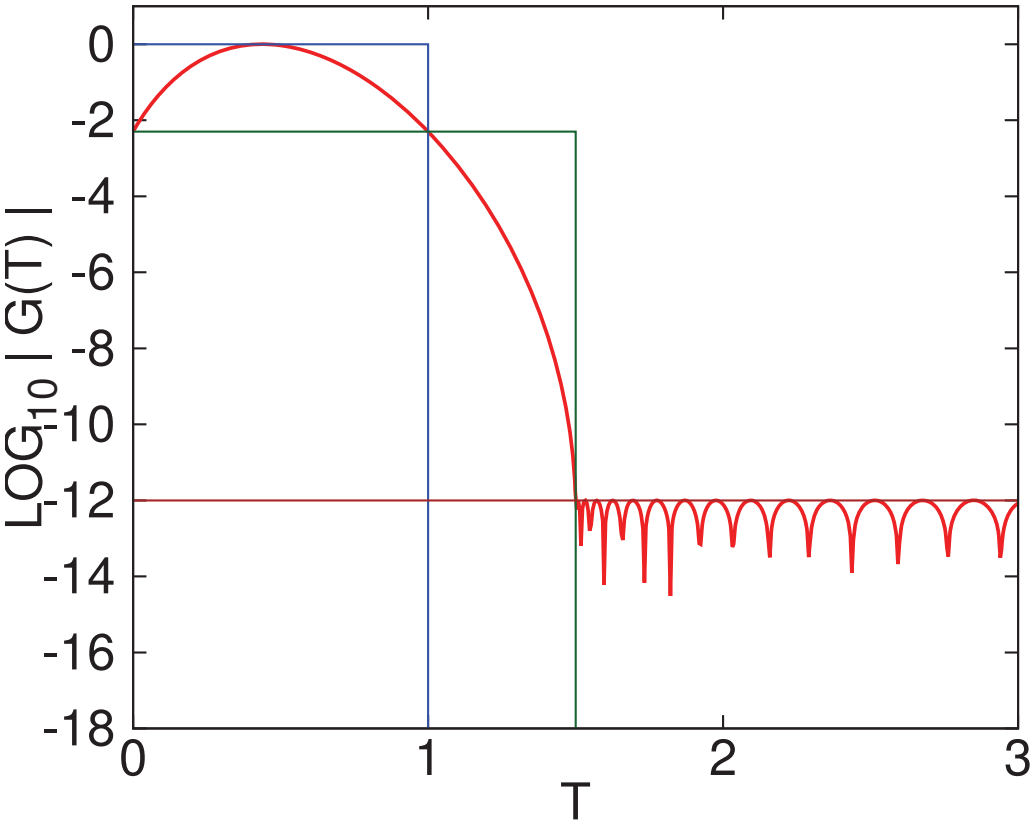


伝達関数の大きさ $|g(t)|$ の対数

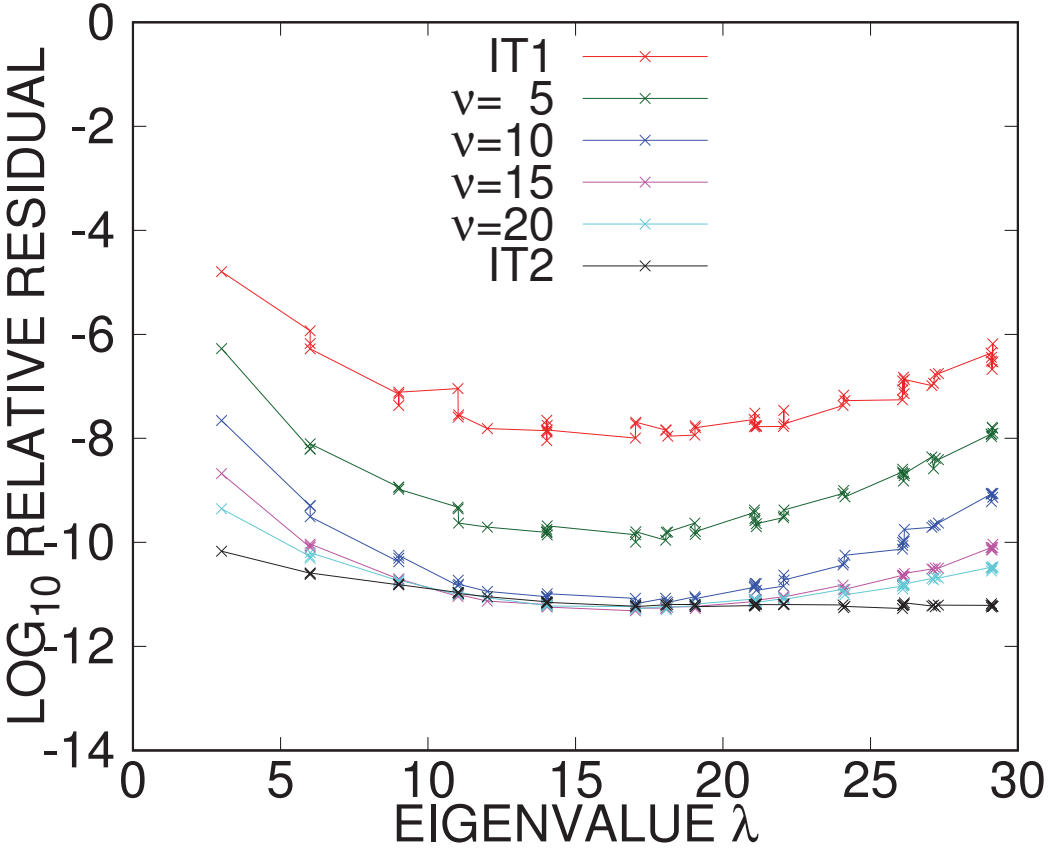


相対残差の大きさ Θ の対数

フィルタ F3-I-4　（$n = 40$，$\mu = 1.5$，g_s =1E-12，g_p =5E-3 ）

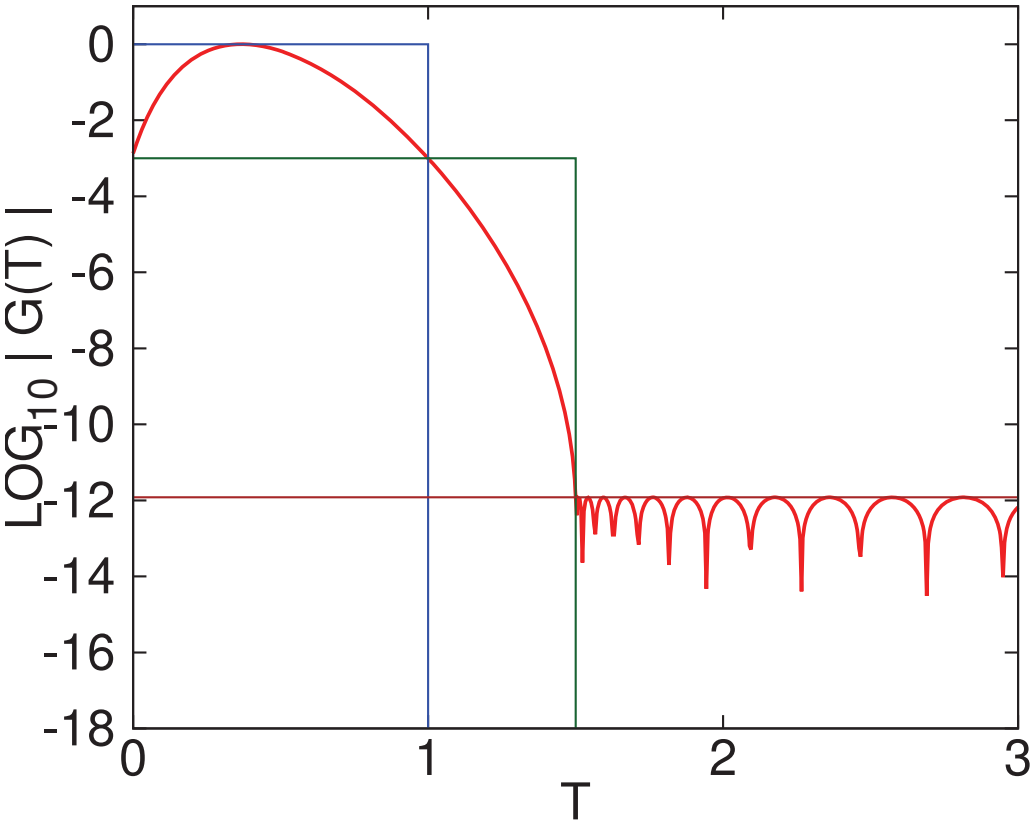


伝達関数の大きさ $|g(t)|$ の対数

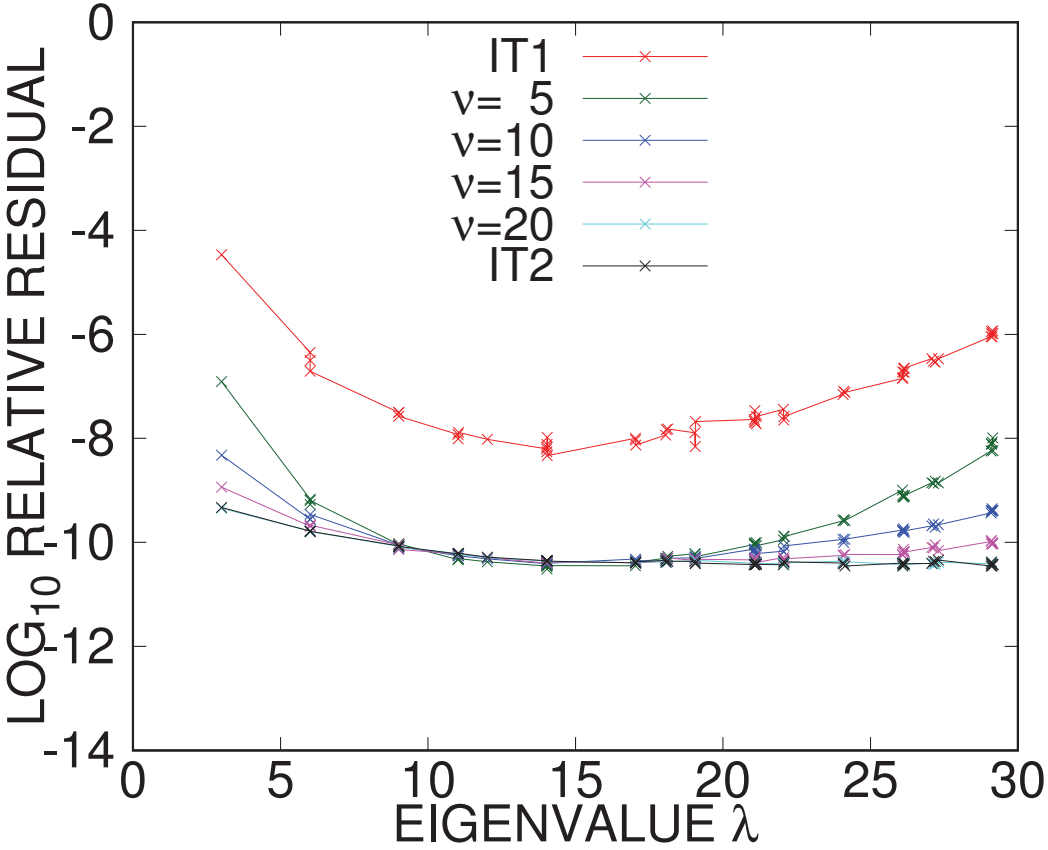


相対残差の大きさ Θ の対数

フィルタ F3-I-5　($n = 20$, $\mu = 1.5$, g_s =1.2E-12, g_p =1E-3)

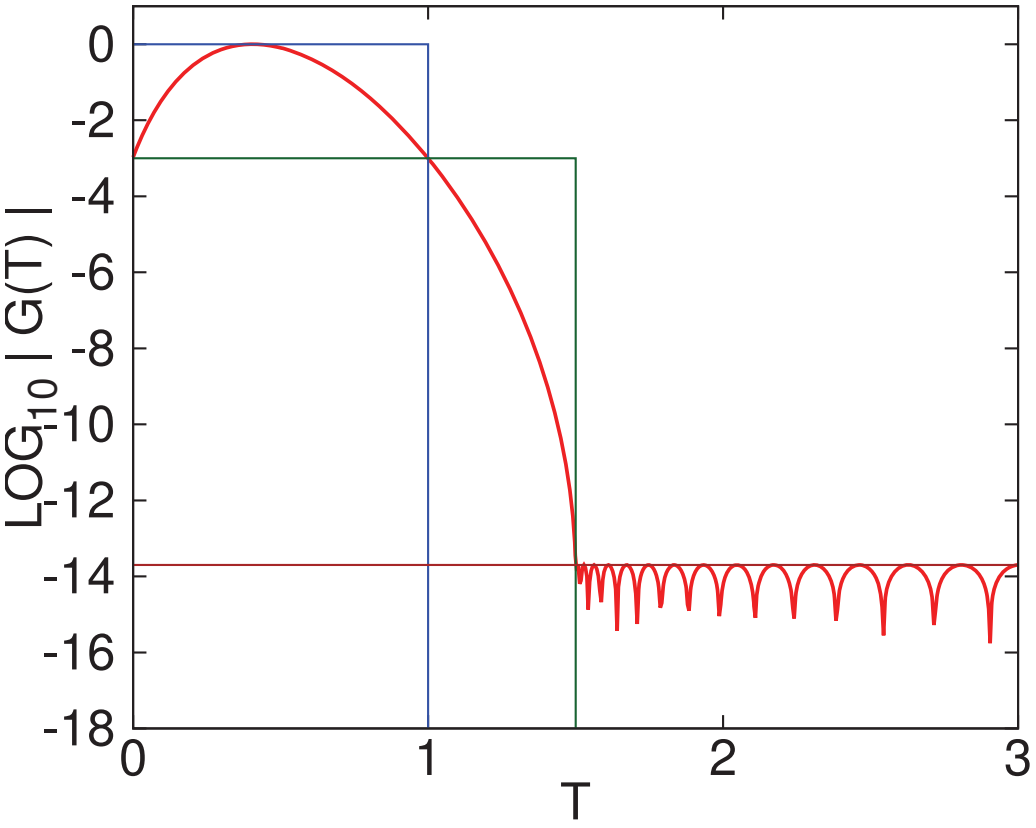


伝達関数の大きさ $|g(t)|$ の対数

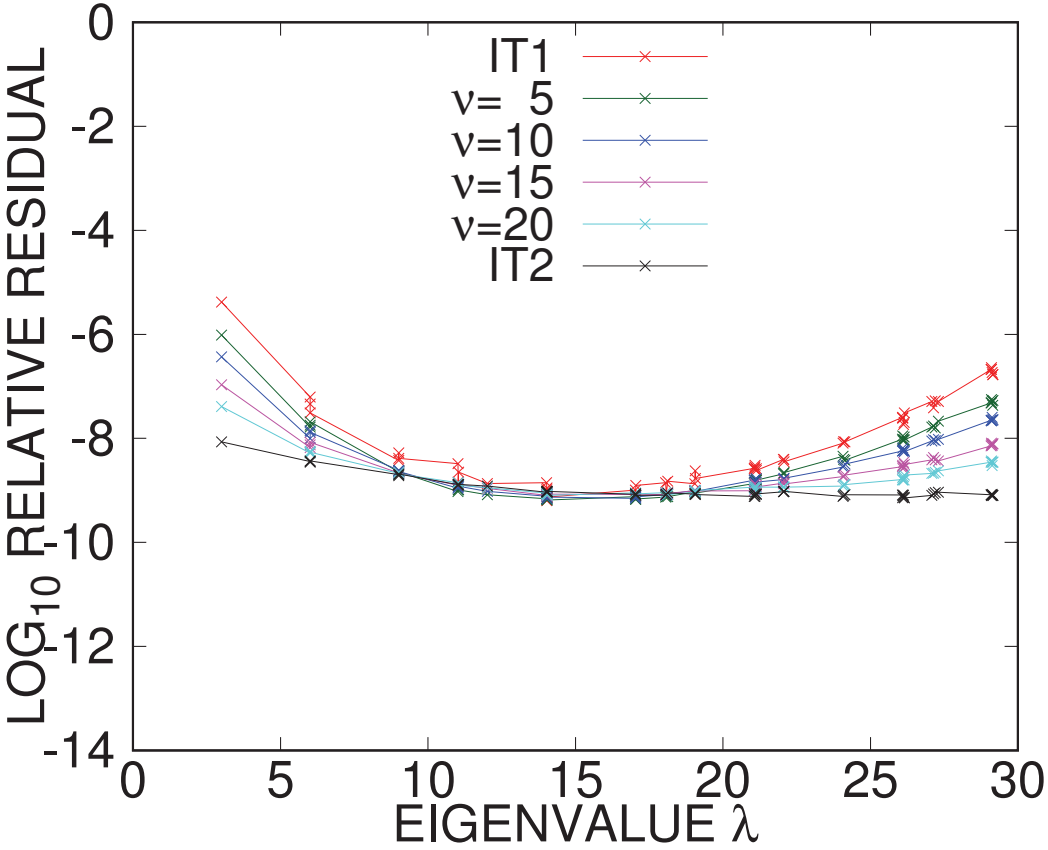


相対残差の大きさ Θ の対数

フィルタ F3-I-6　（$n = 30$，$\mu = 1.5$，g_s =2E-14，g_p =1E-3 ）

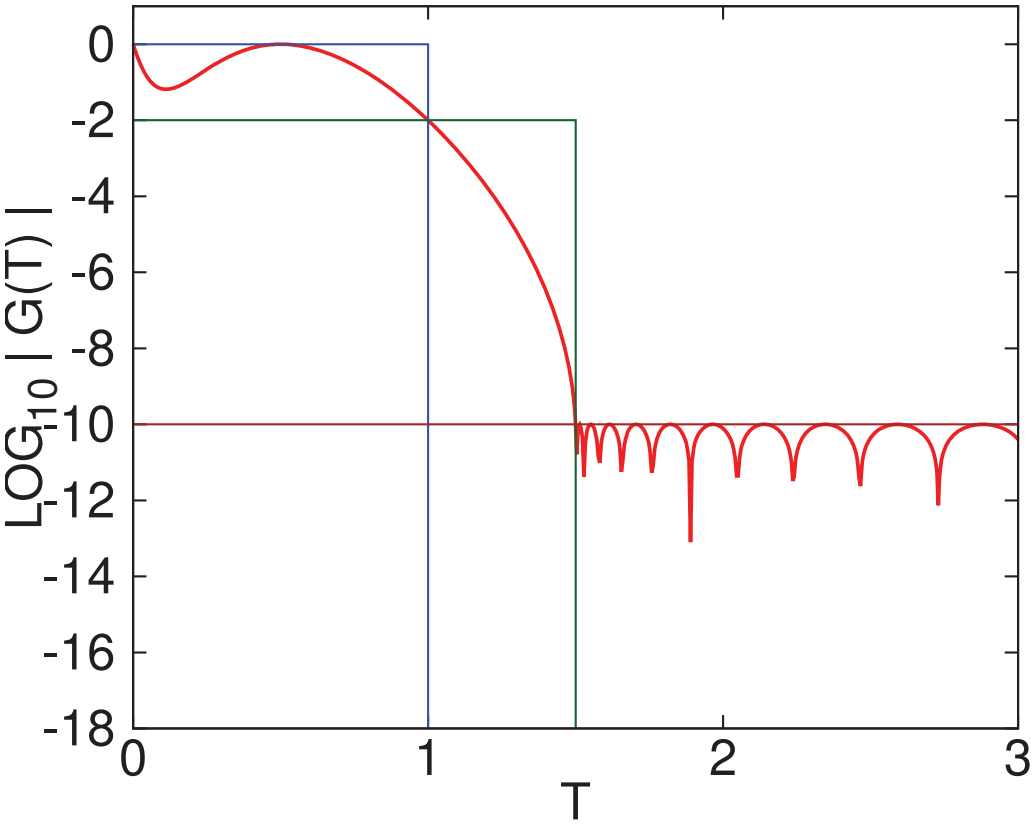


伝達関数の大きさ $|g(t)|$ の対数

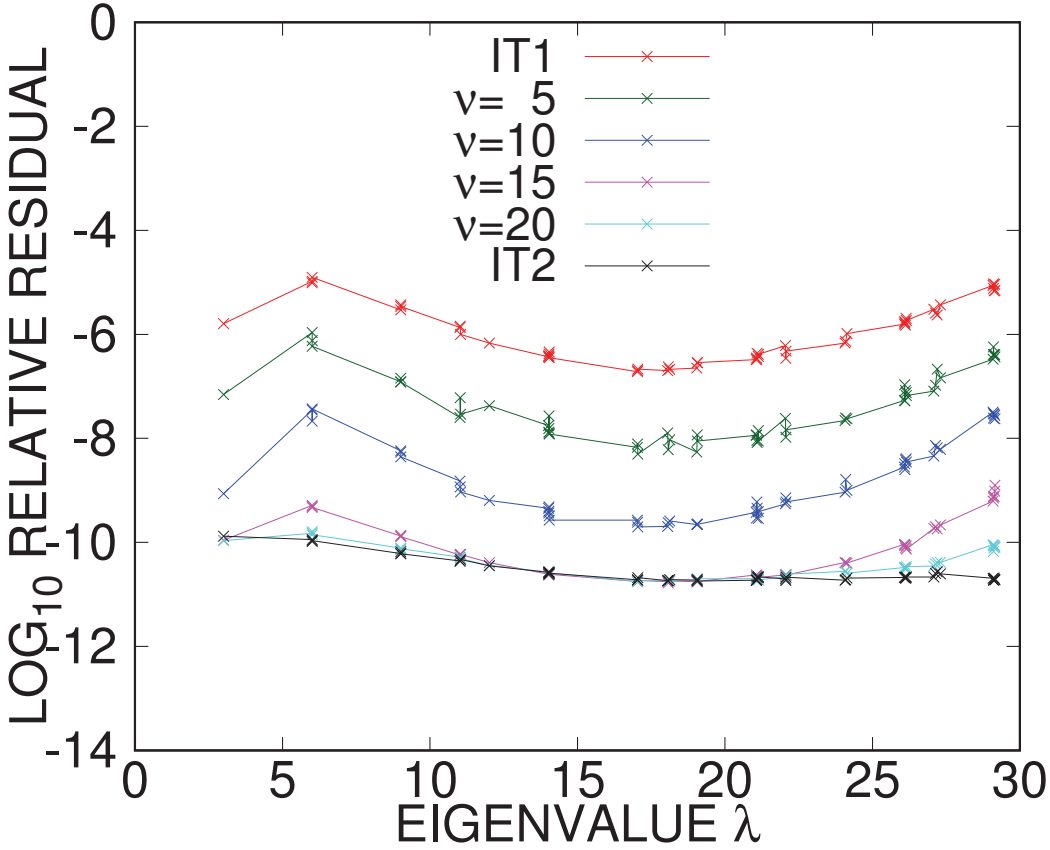


相対残差の大きさ Θ の対数

フィルタ F3-II-1　($n = 30$, $\mu = 1.5$, g_s =1E-10, g_p =1E-2)

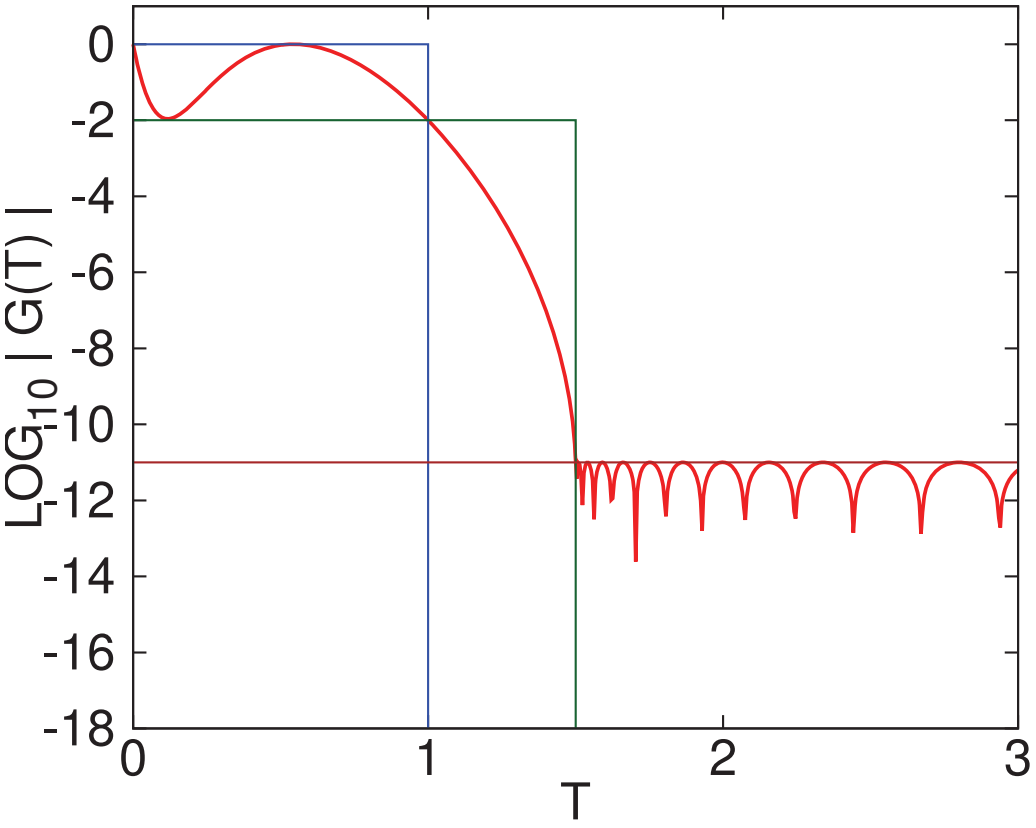


伝達関数の大きさ $|g(t)|$ の対数

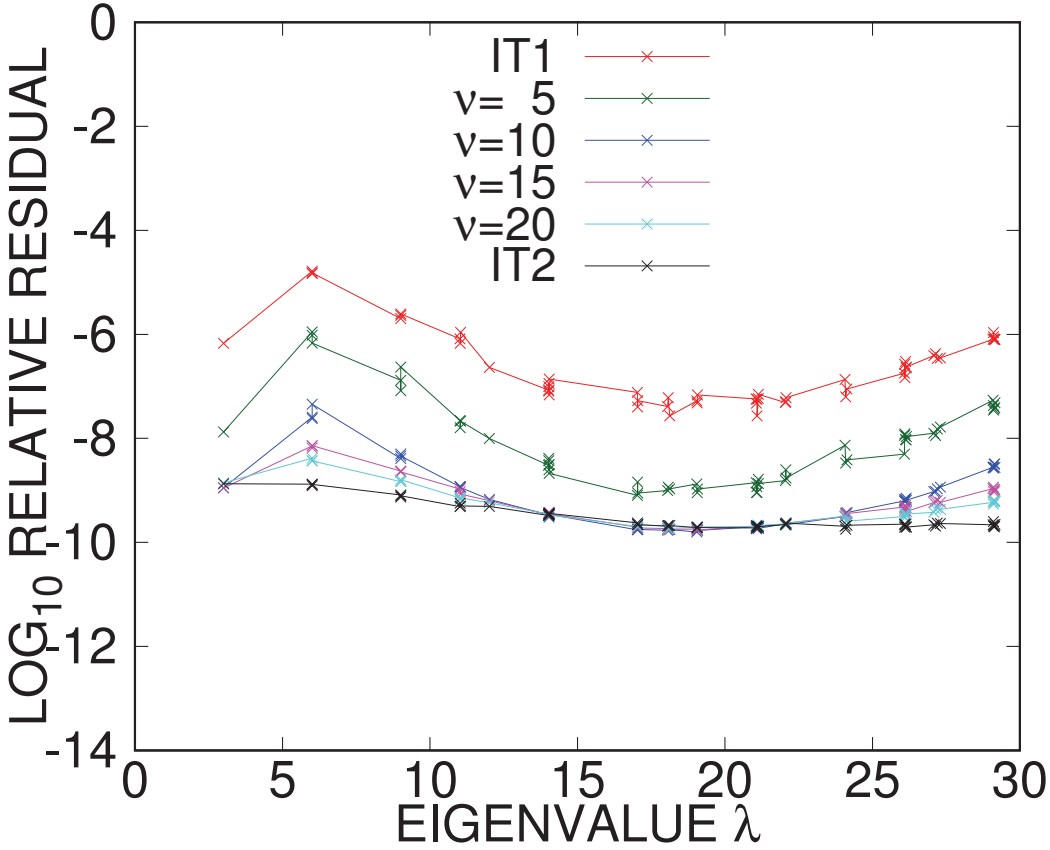


相対残差の大きさ Θ の対数

フィルタ F3-II-2　($n = 34$, $\mu = 1.5$, g_s =1E-11, g_p =1E-2)

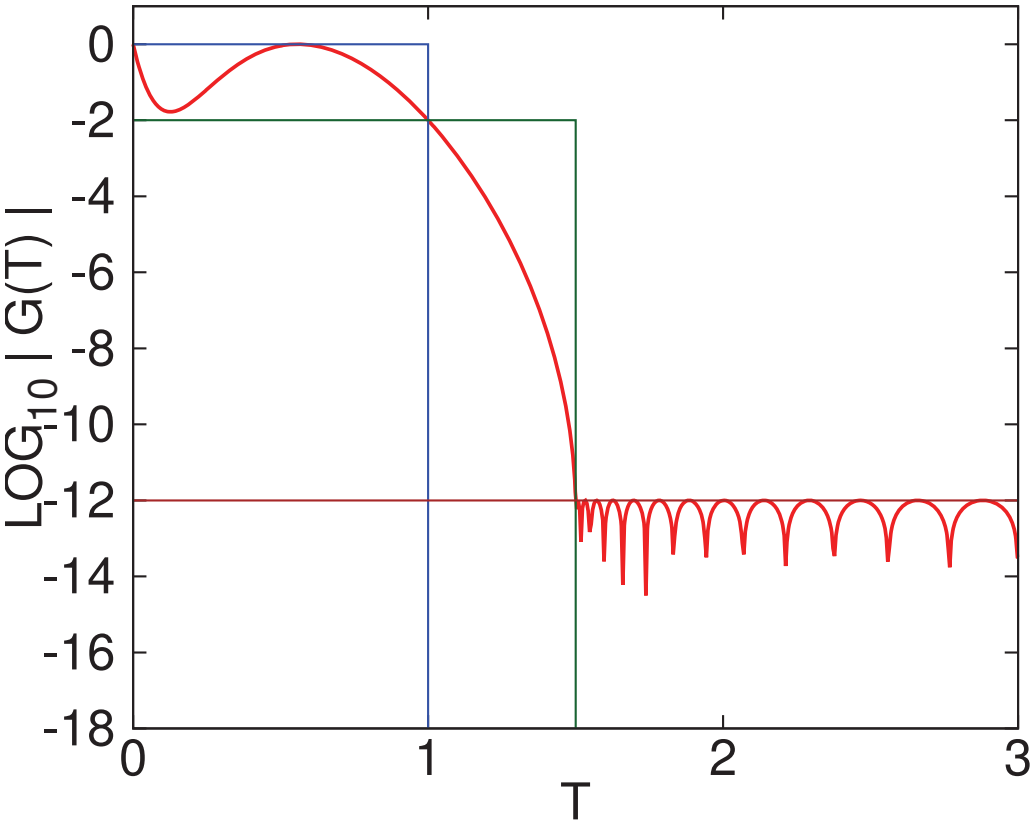


伝達関数の大きさ $|g(t)|$ の対数

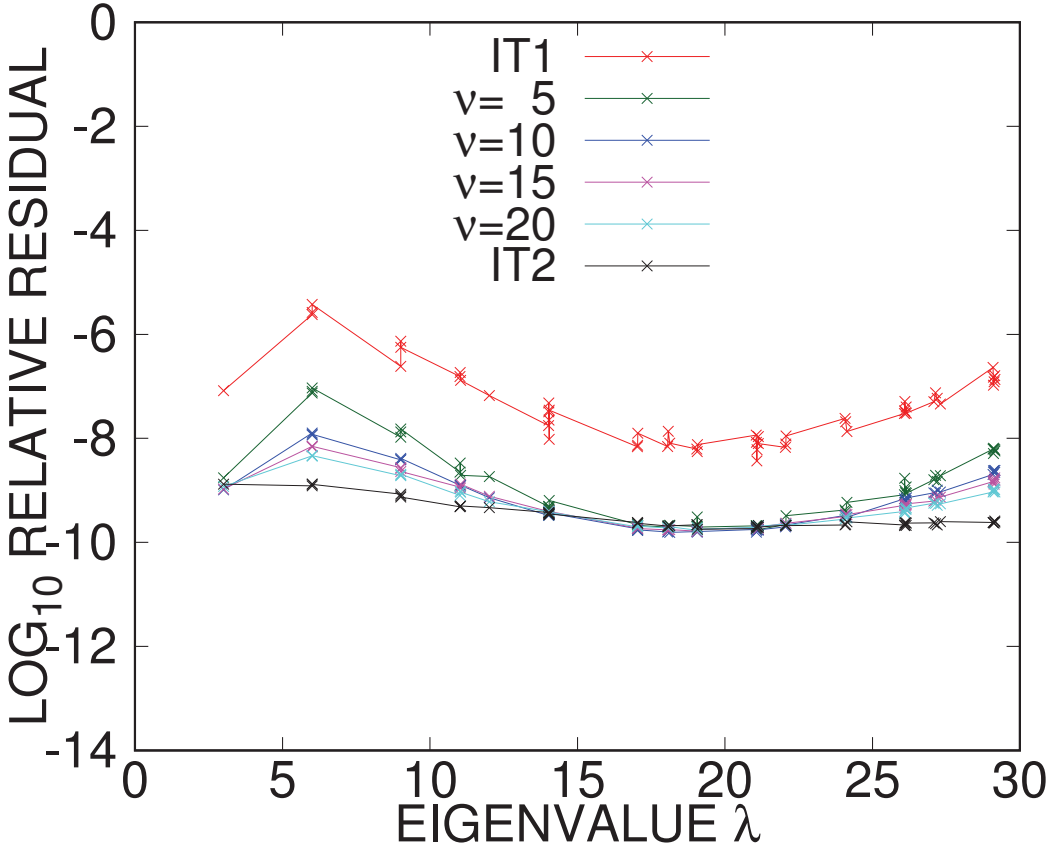


相対残差の大きさ Θ の対数

フィルタ F3-II-3　($n = 41$, $\mu = 1.5$, g_s =1E-12, g_p =1E-2)



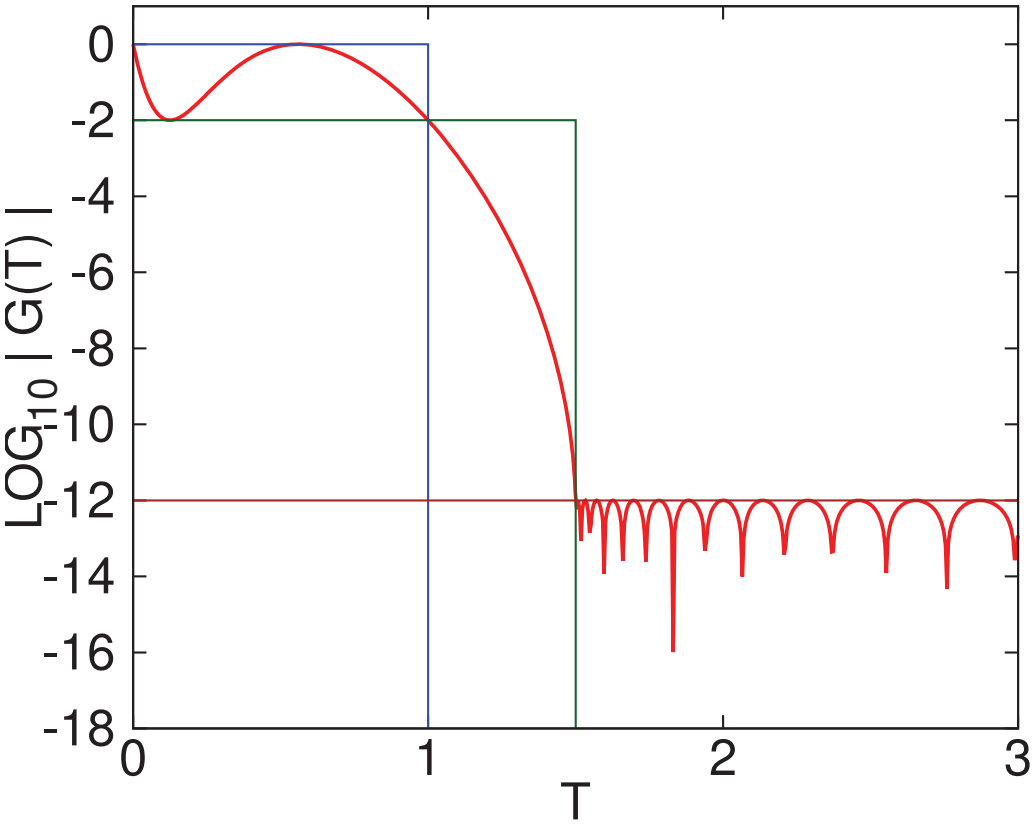
伝達関数の大きさ $|g(t)|$ の対数



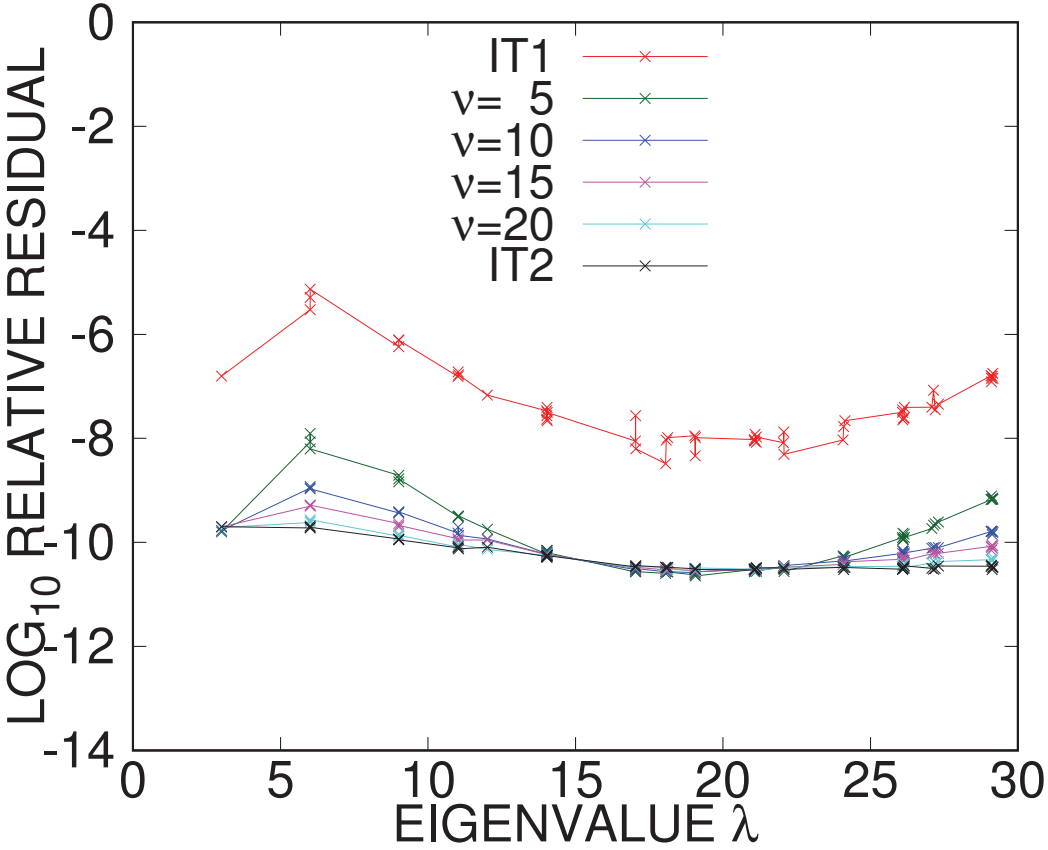
相対残差の大きさ Θ の対数

実数シフトのレゾルベントを4つ用いたフィルタによる例

フィルタ F4-I-1　($n = 23$, $\mu = 1.5$, g_s =1E-12, g_p =1E-2)

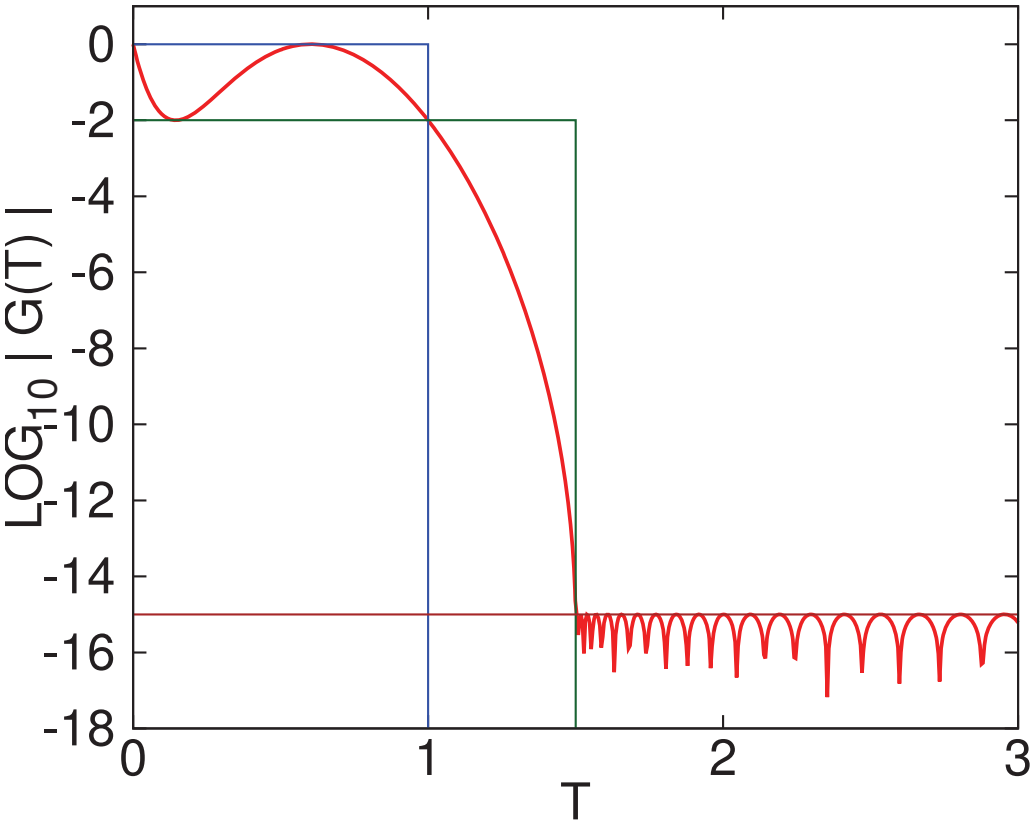


伝達関数の大きさ $|g(t)|$ の対数

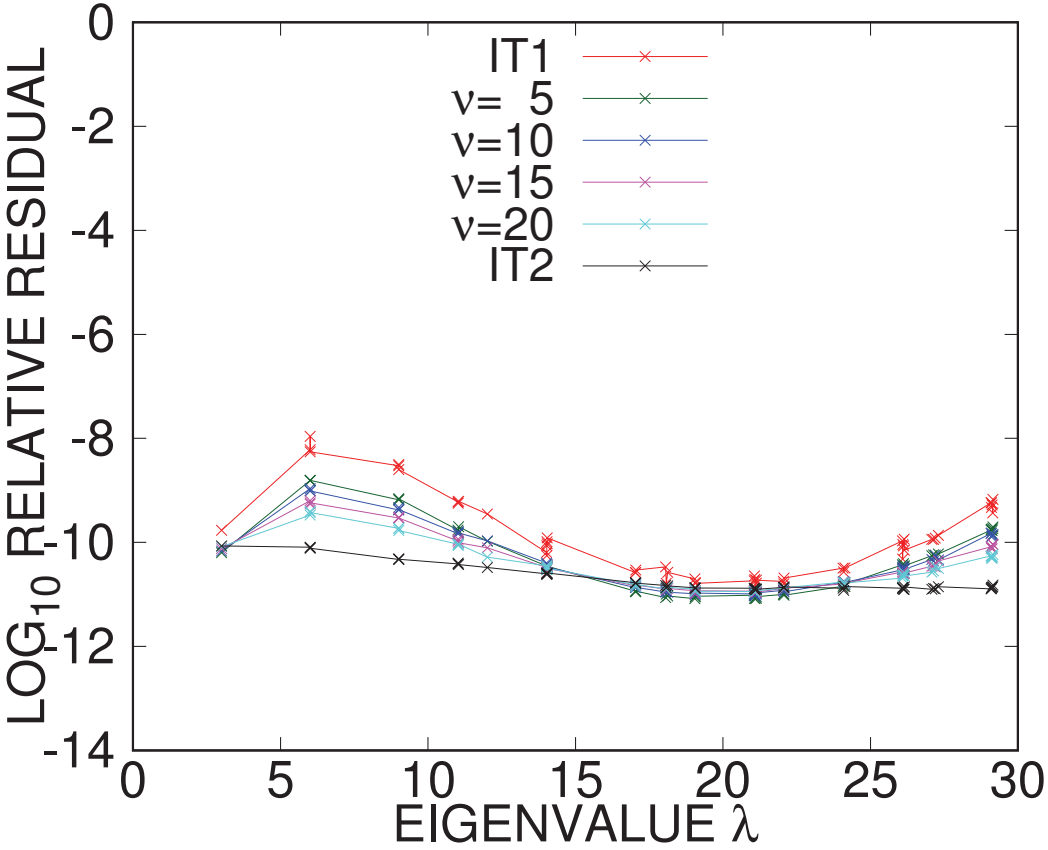


相対残差の大きさ Θ の対数

フィルタ F4-I-2　($n = 40$, $\mu = 1.5$, g_s =1E-15, g_p =1E-2)

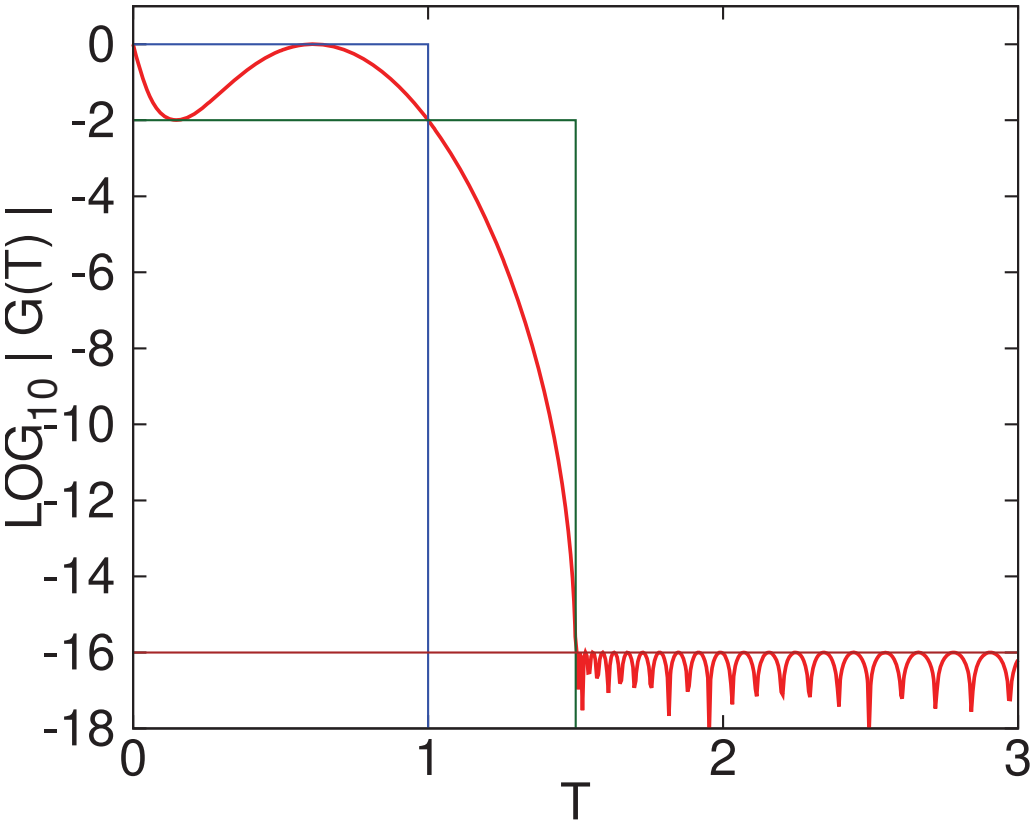


伝達関数の大きさ $|g(t)|$ の対数

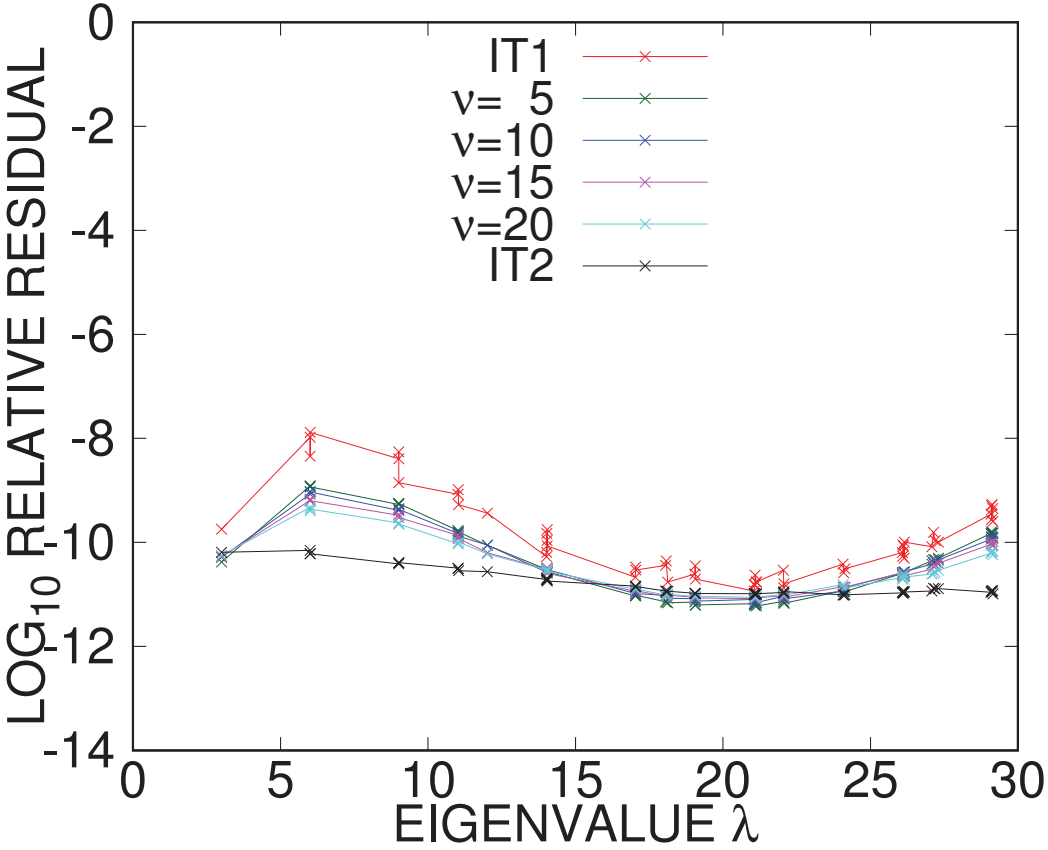


相対残差の大きさ Θ の対数

フィルタ F4-I-3　($n = 50$, $\mu = 1.5$, g_s =1E-16, g_p =1E-2)

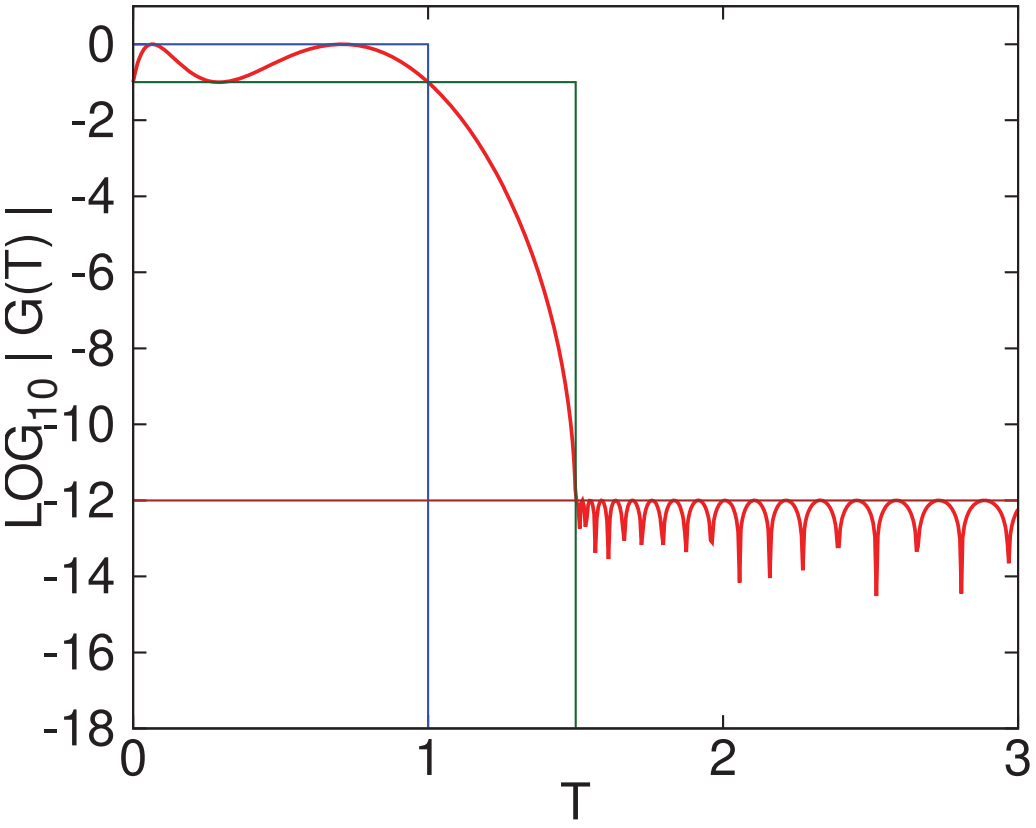


伝達関数の大きさ $|g(t)|$ の対数

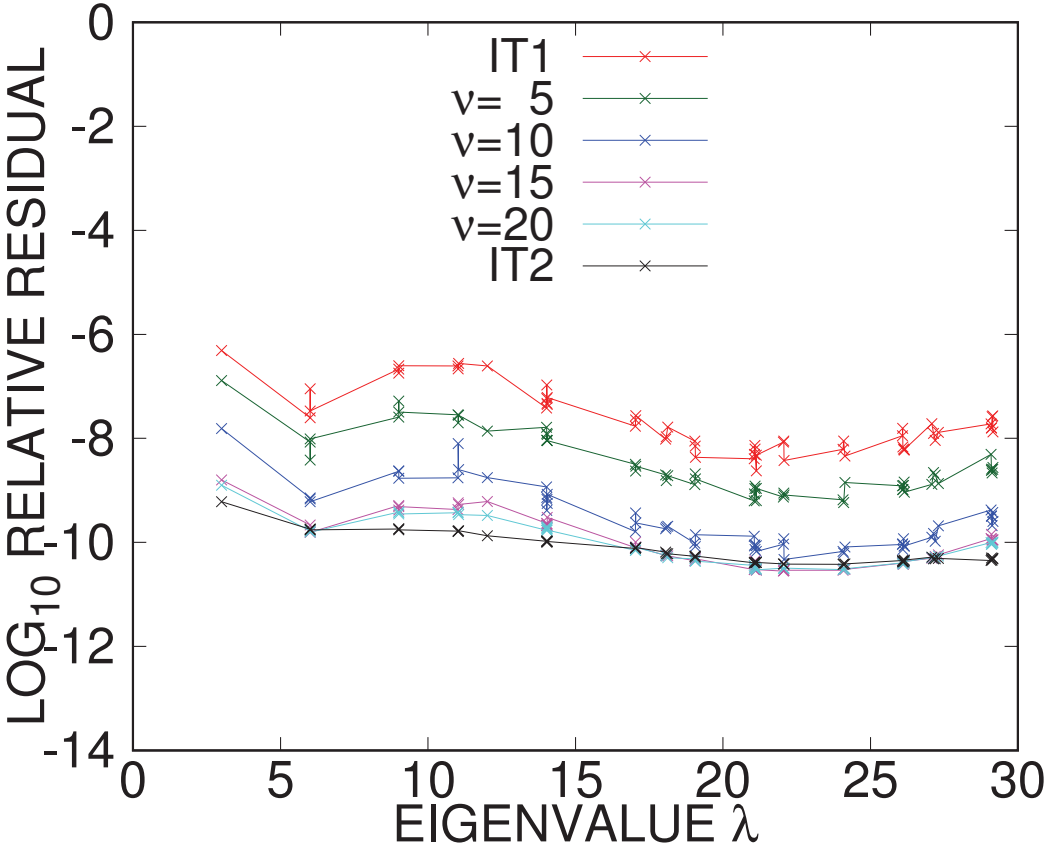


相対残差の大きさ Θ の対数

フィルタ F4-II-1　($n = 63$, $\mu = 1.5$, g_s =1E-12, g_p =1E-1)

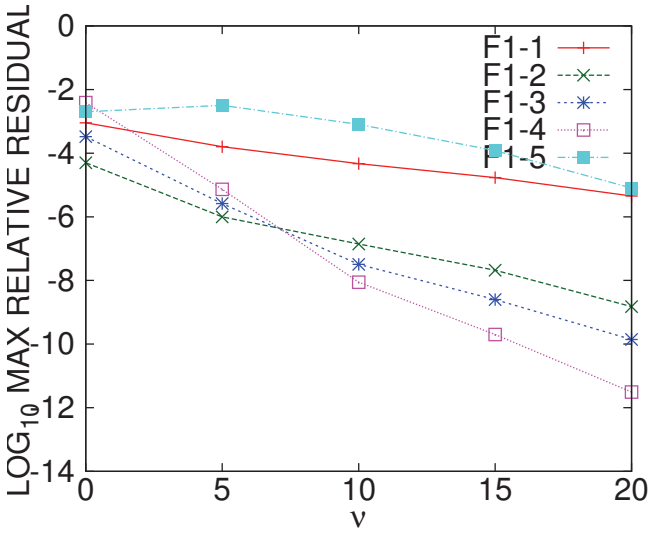


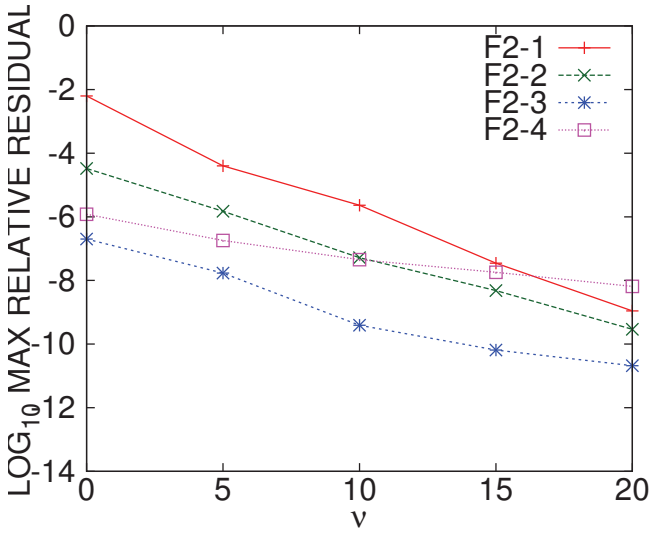
伝達関数の大きさ $|g(t)|$ の対数

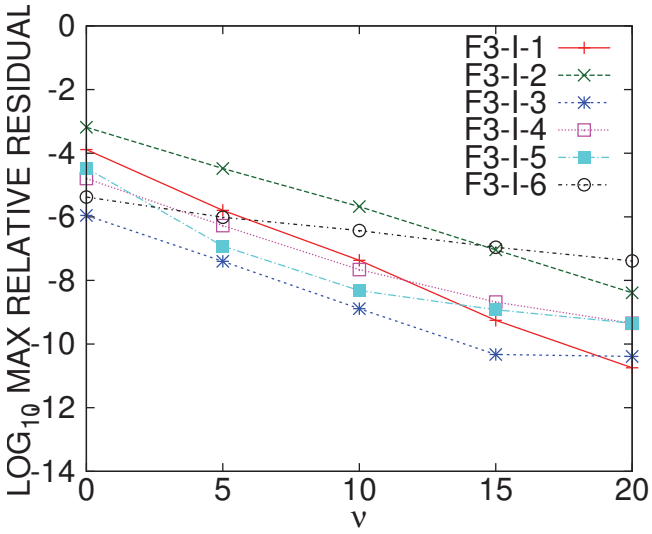


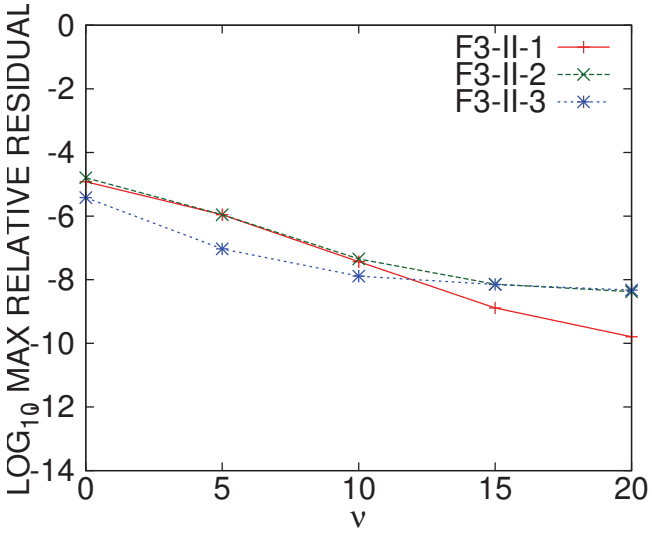
相対残差の大きさ Θ の対数

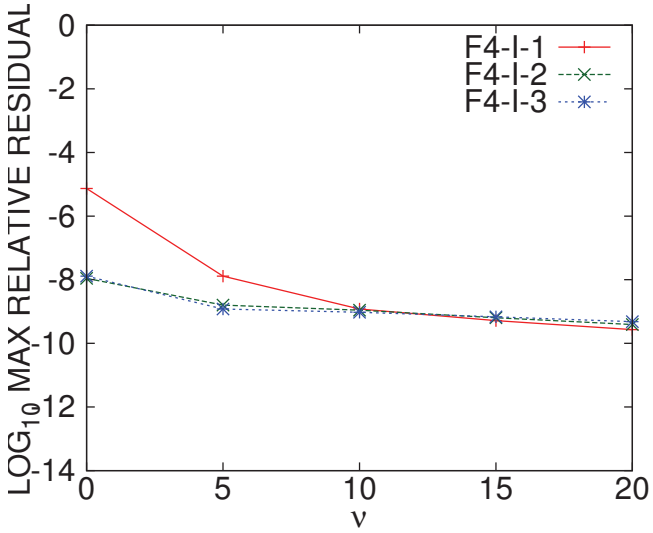
前処理用のフィルタの次数 ν と 相対残差の大きさの最大値

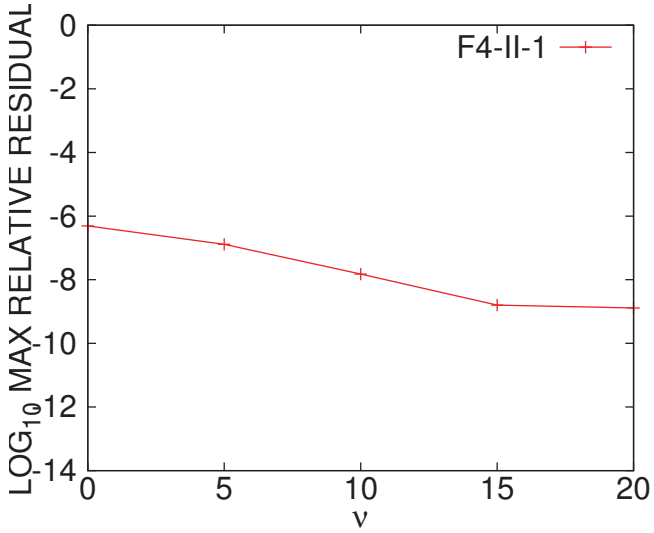




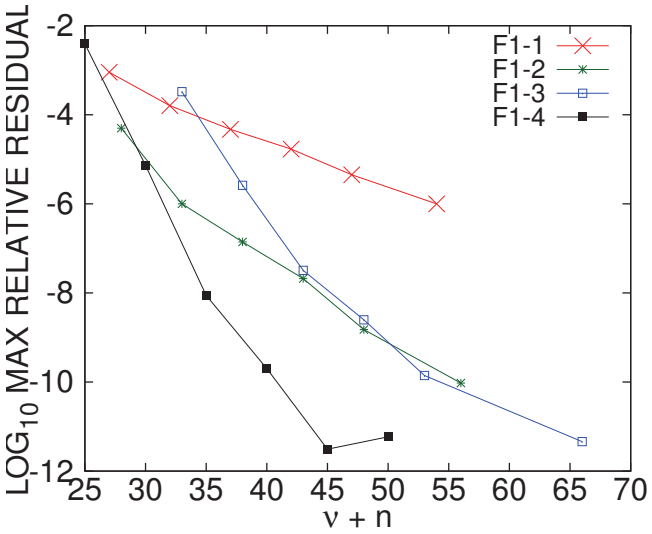


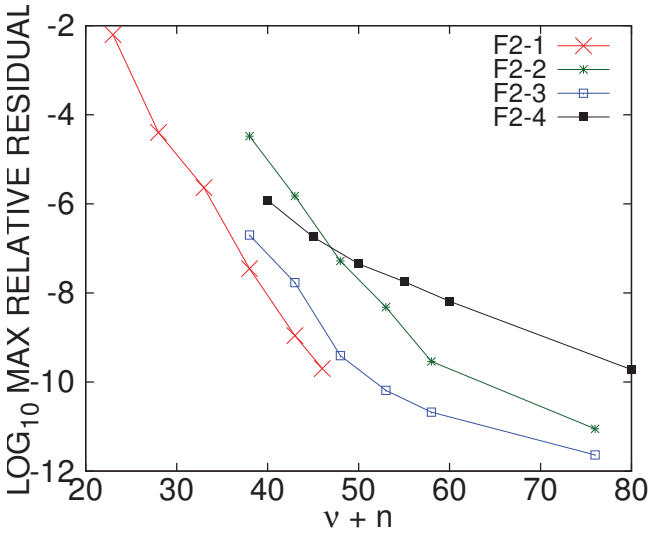


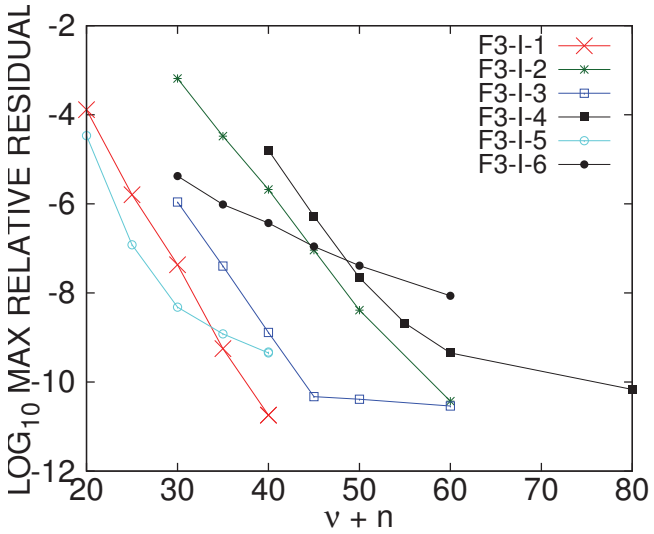


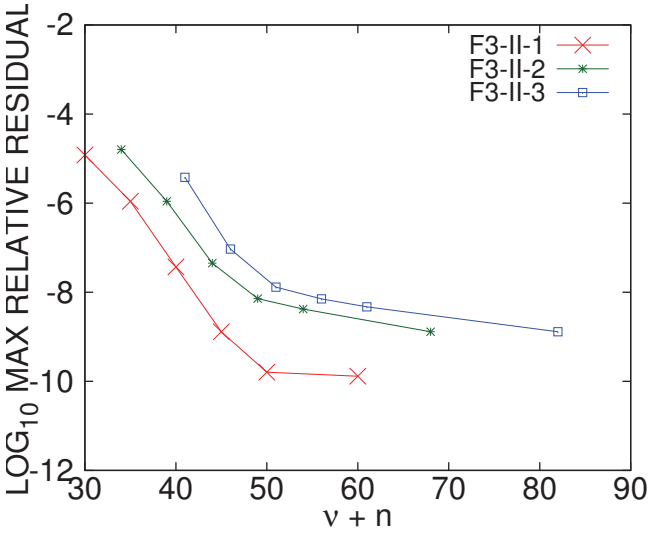


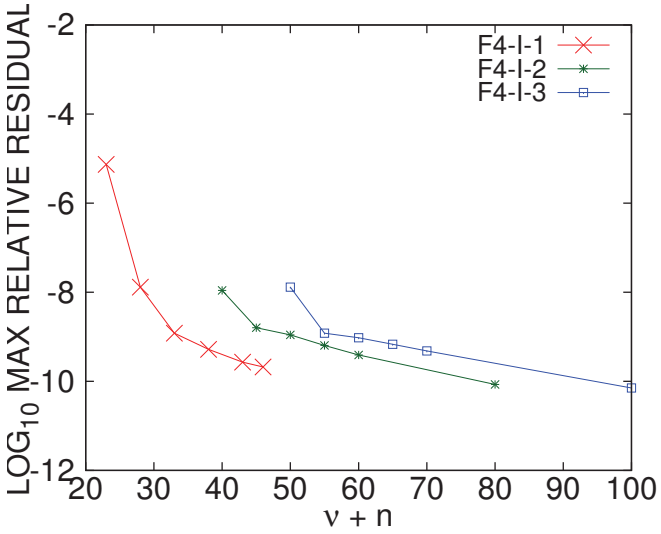
フィルタの次数の合計 $\nu + n$ と 相対残差の大きさの最大値

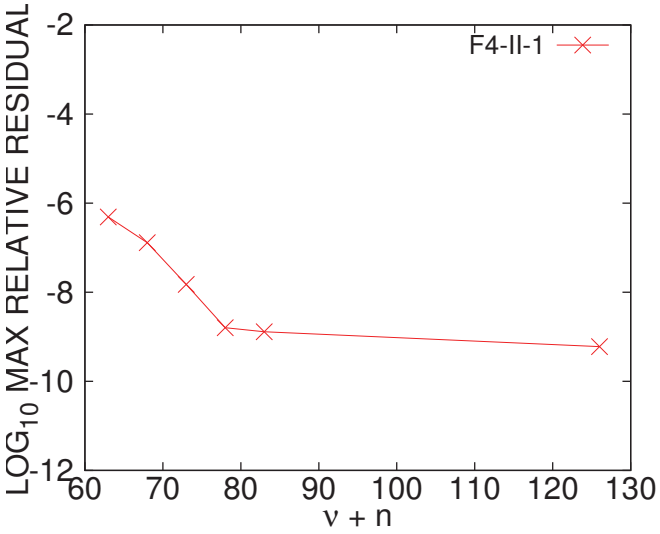




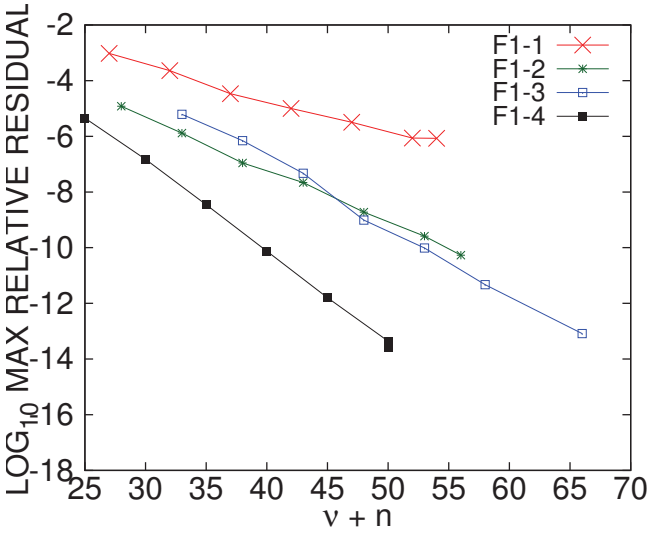


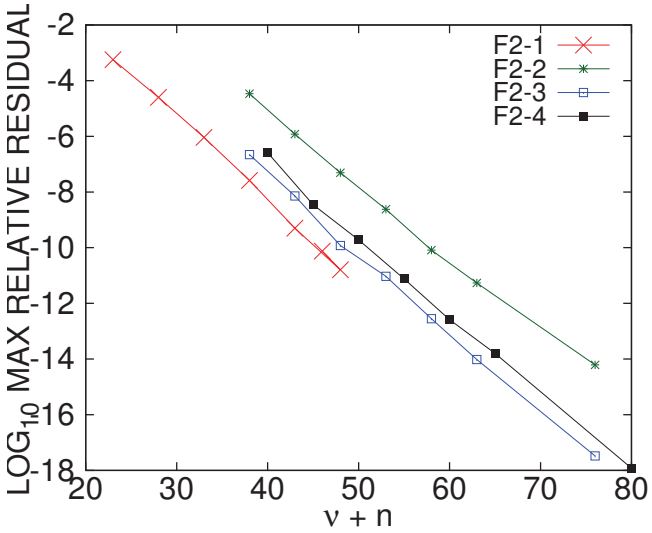


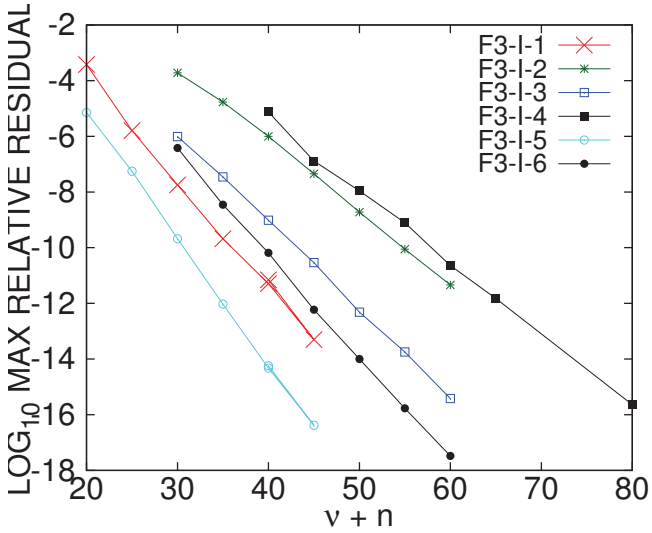


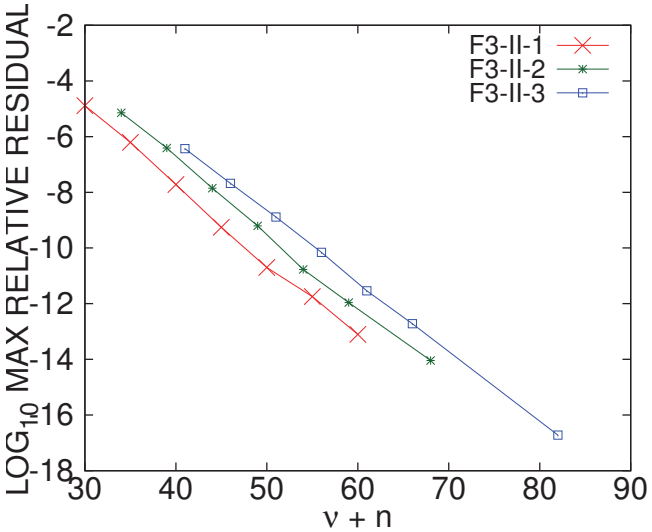


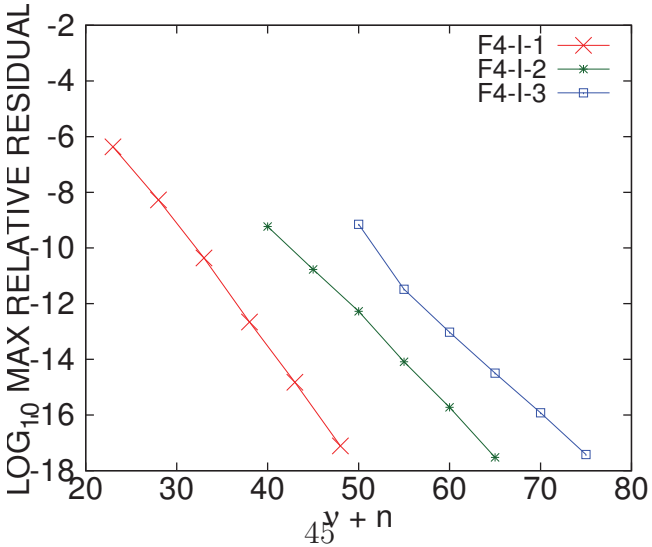
フィルタの次数の合計 $\nu + n$ と 相対残差の大きさの最大値
（四倍精度計算，小規模問題 $(N_1, N_2, N_3) = (30, 40, 50)$）

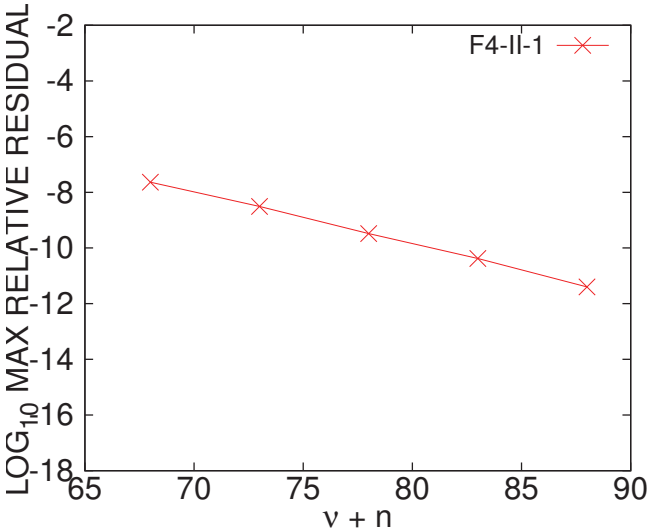












実験結果と現状のまとめ

- 次数の低いフィルタとB-正規直交化を用いた前処理で，近似固有対の精度を改良できる.
- 改良は前処理に費やす計算量に応じたものになる.
- 本来のフィルタの次数 n を大きくとり前処理をしない $\nu = 0$ よりも，n を少なめにとった分を前処理のフィルタの次数 ν にあてる方が良さそう.
- 低次のフィルタでベクトルの組の前処理を行えば，近似固有対の相対残差の最大値を減らせるので，単一のレゾルベントから構成されたフィルタでもうまく利用できて計算量を少なくできる.