知能工学特別講義 第７講

担当：和田山 正

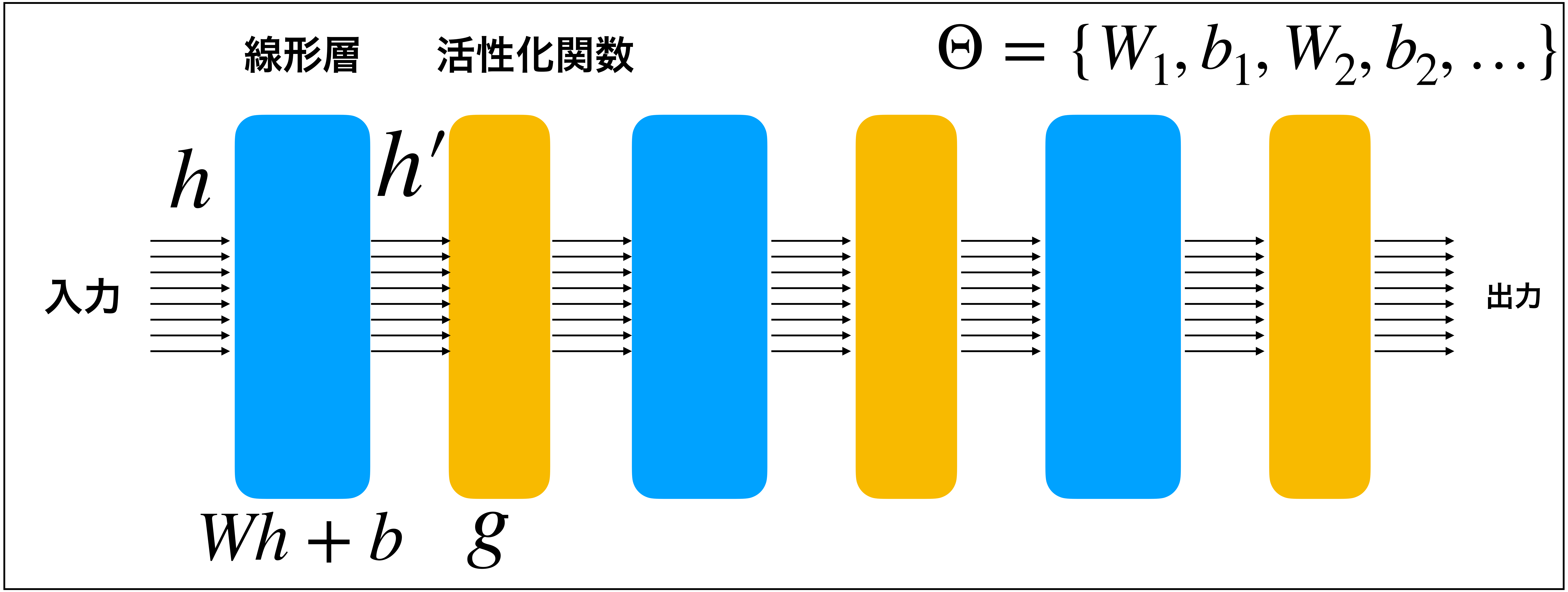
名古屋工業大学

本講義の内容

- 深層学習の解釈可能性問題
- 深層展開に基づく収束加速：ジグザグパラメータの謎
- 勾配法の収束
- チェビシェフステップに基づく収束加速

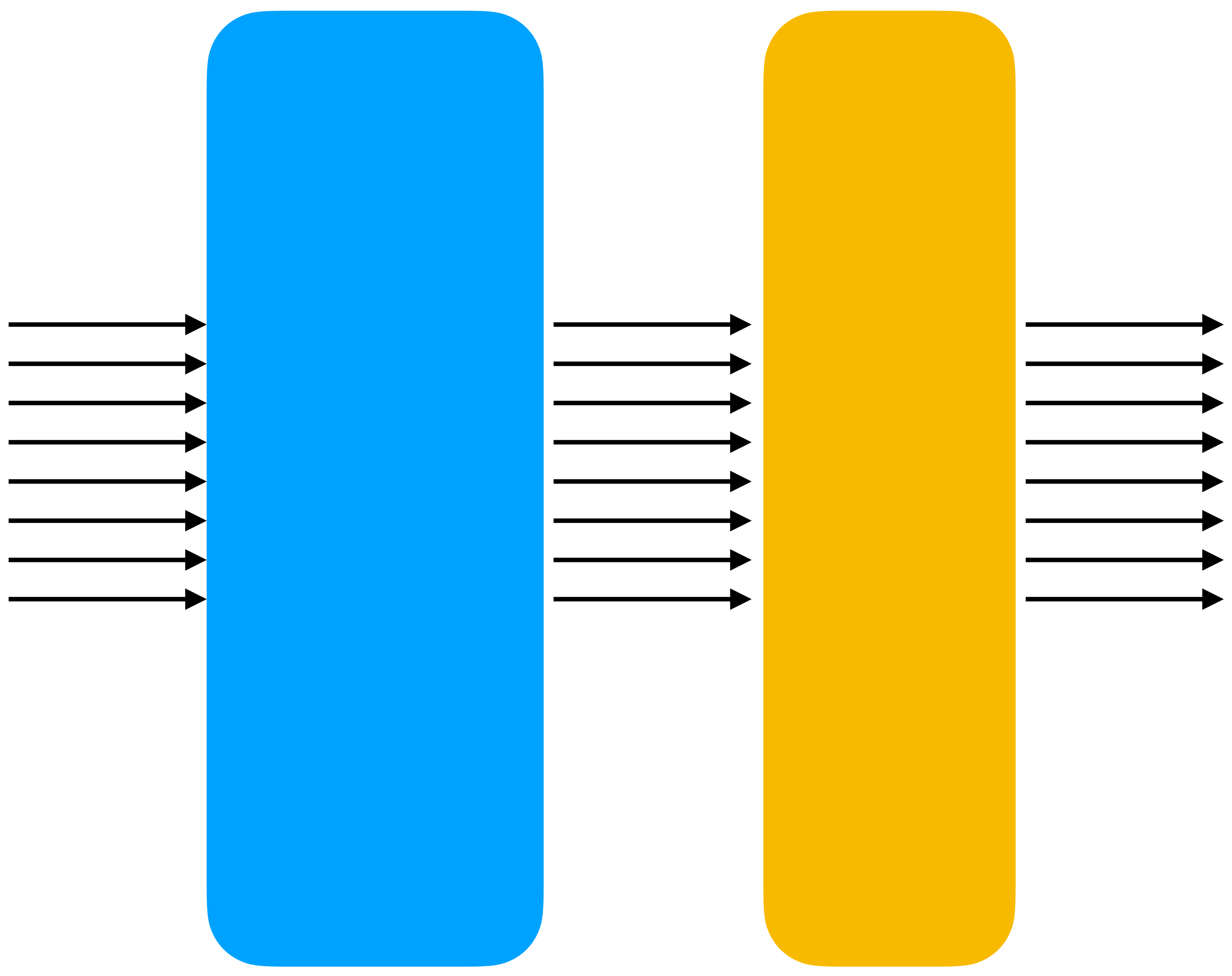
深層ニューラルネットワーク

$$f_\Theta : \mathbb{R}^n \to \mathbb{R}^m$$



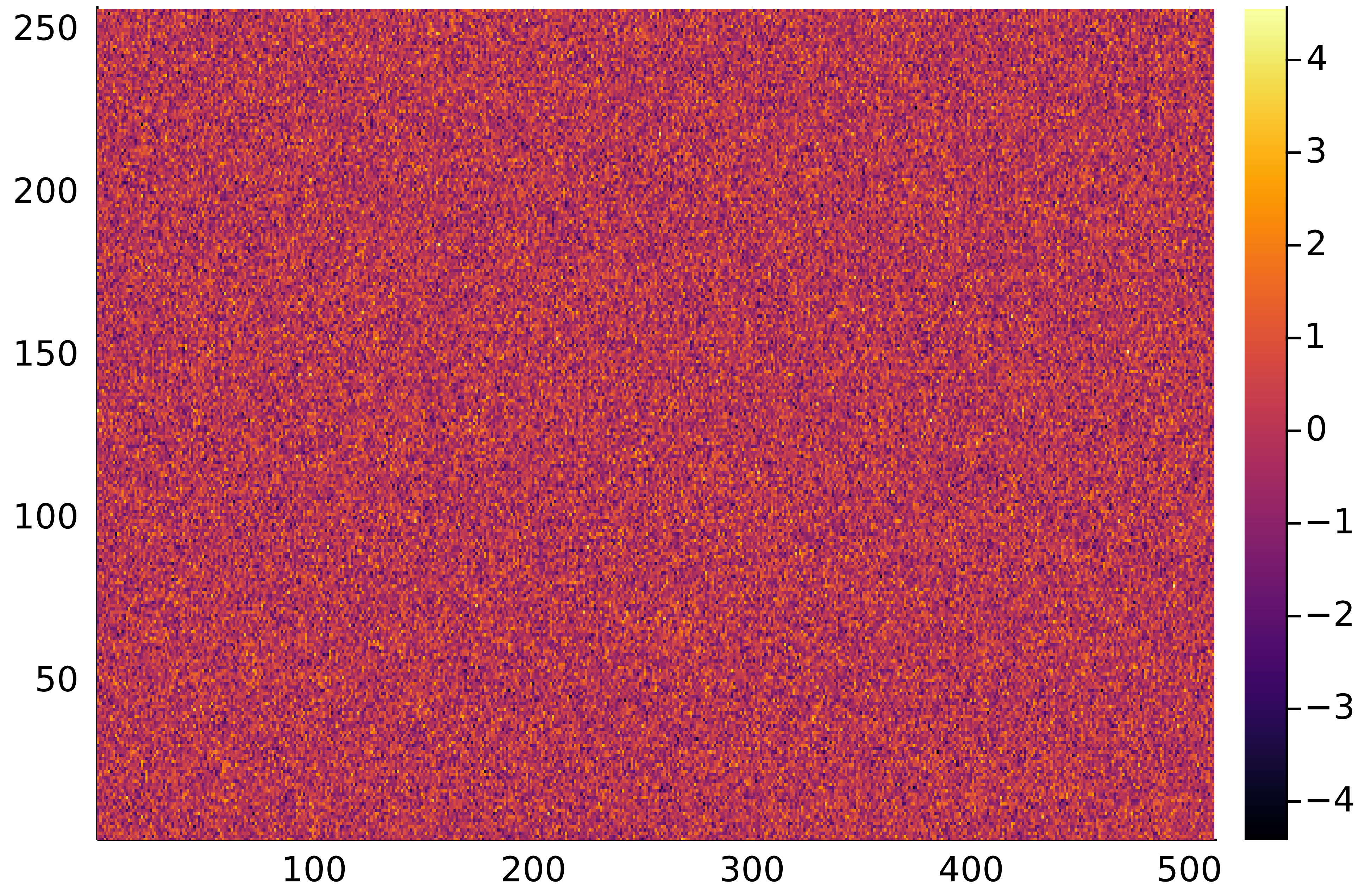
学習された結果を眺めてみる

線形層　活性化関数



$Wh + b$　g

例: Wの中身(512 x 512)



学習されたパラメータの「意味」を読み取ることは困難

深層学習における解釈可能性問題

- 深層学習によって得られた学習パラメータの意味を読み取ることが難しい
- 例えば、深層ネットワークを利用した製品が致命的な判別エラーを起こすケースがあることが分かった⇒どう修正すればよいのか分からない
- 意思決定に利用する場合、判断の根拠が見えない
- なぜ深層学習がうまく動いているのか、深いところで理解できない
- 深層学習における「解釈可能性問題」は現在のホットトピック

解釈可能性問題に対する一つの方向性

	モデルベース アプローチ(深層展開)	ブラックボックス アプローチ
メリット	ドメイン知識の有効活用 既存の優れたアルゴリズムの利用 適切なパラメータを学習可能に 解釈可能性に優れる	ドメイン知識不要 データ駆動型アルゴリズム設計 (例: NNモデル作成→訓練プロセス)
デメリット	ドメイン知識が必要	解釈可能性問題がある

深層展開の解釈可能性と「戦略の発見」

- 深層展開の学習結果を眺めてみることで、何か新しく学べることがあるのではないか？
- 例えば、収束を加速するための「パラメータ戦略」があるのではないか？
- 新しい工学的な知見が学習結果から得られると大変面白い

TISTAの復元性能(再掲)

学習パラメータの「ジグザグ形状」が見られる

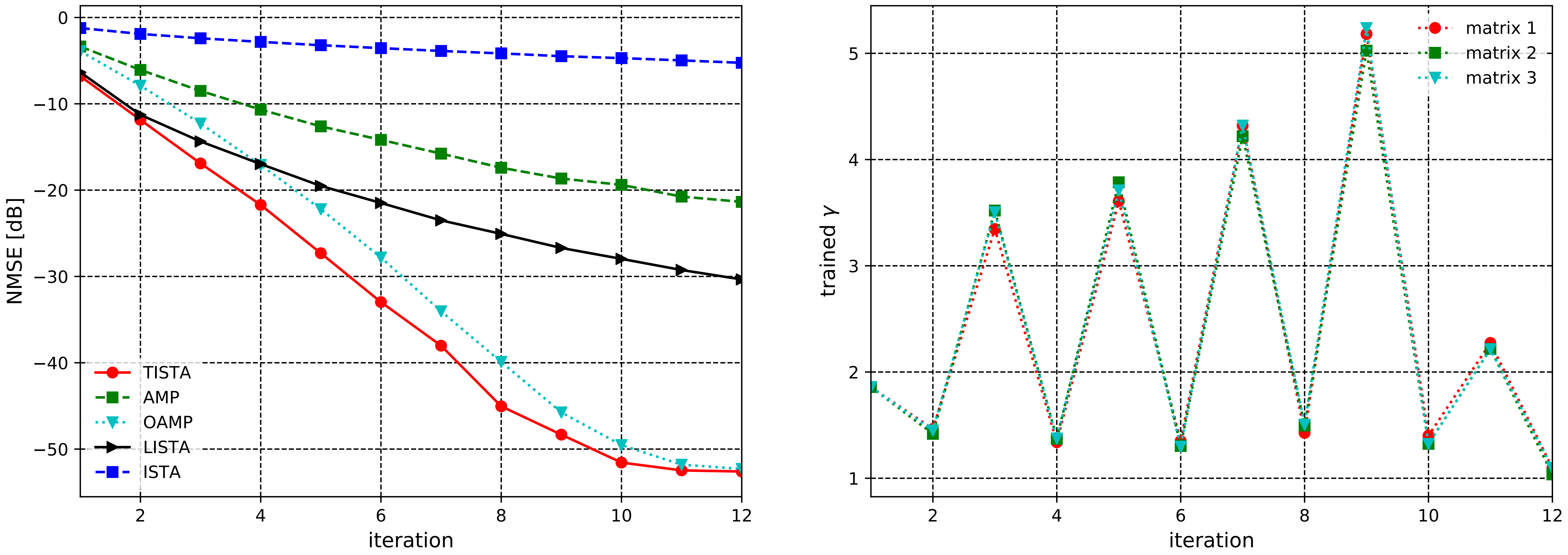
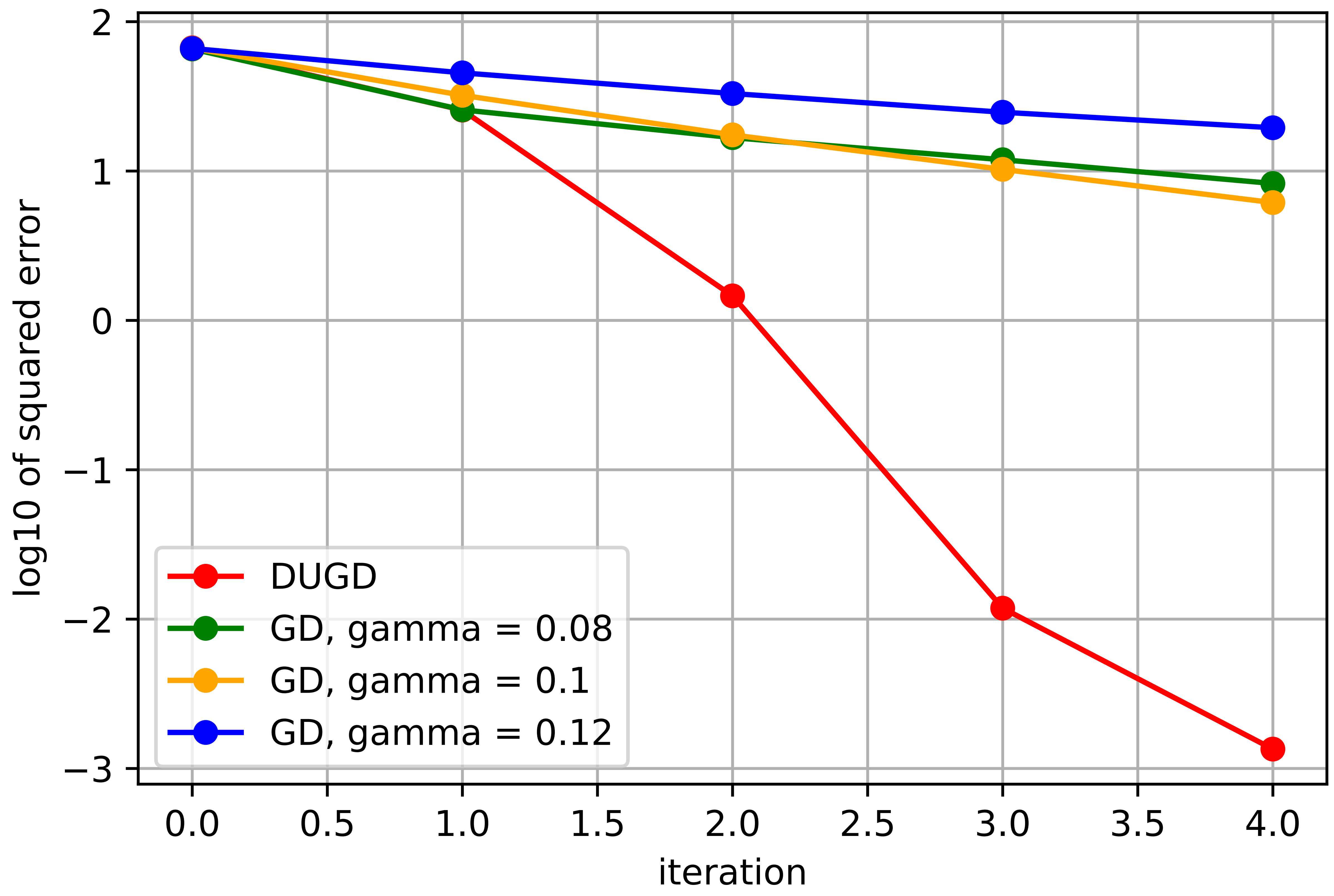
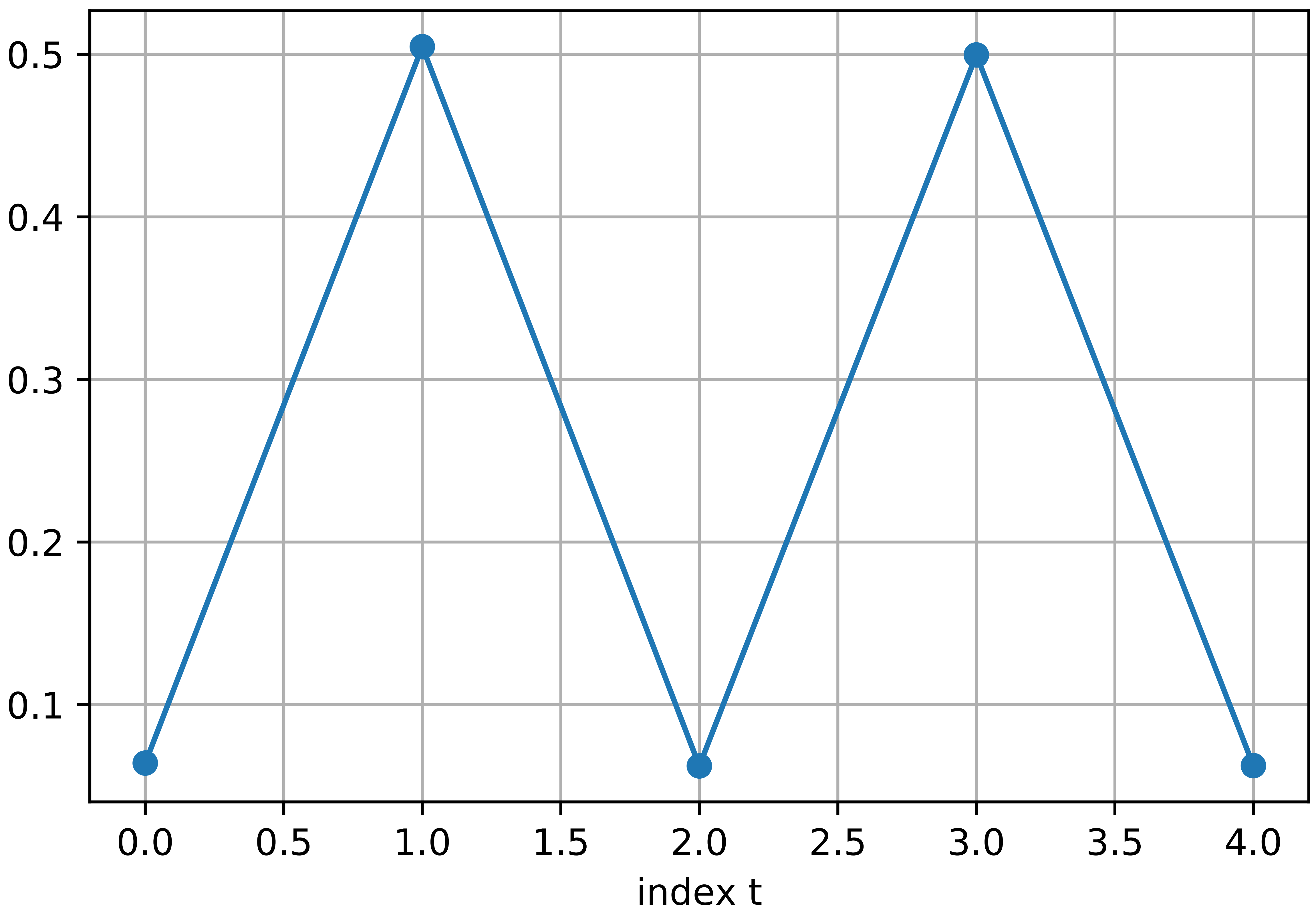


Figure 7: Sparse signal recovery for $(n, m) = (300, 150)$, $p = 0.1$ and SNR$= 50$ dB. (Left) MSE performances as a function of iteration steps. (Right) Trained values of γ_t under three different mesurement matrices. The initial value is set to 1.0.

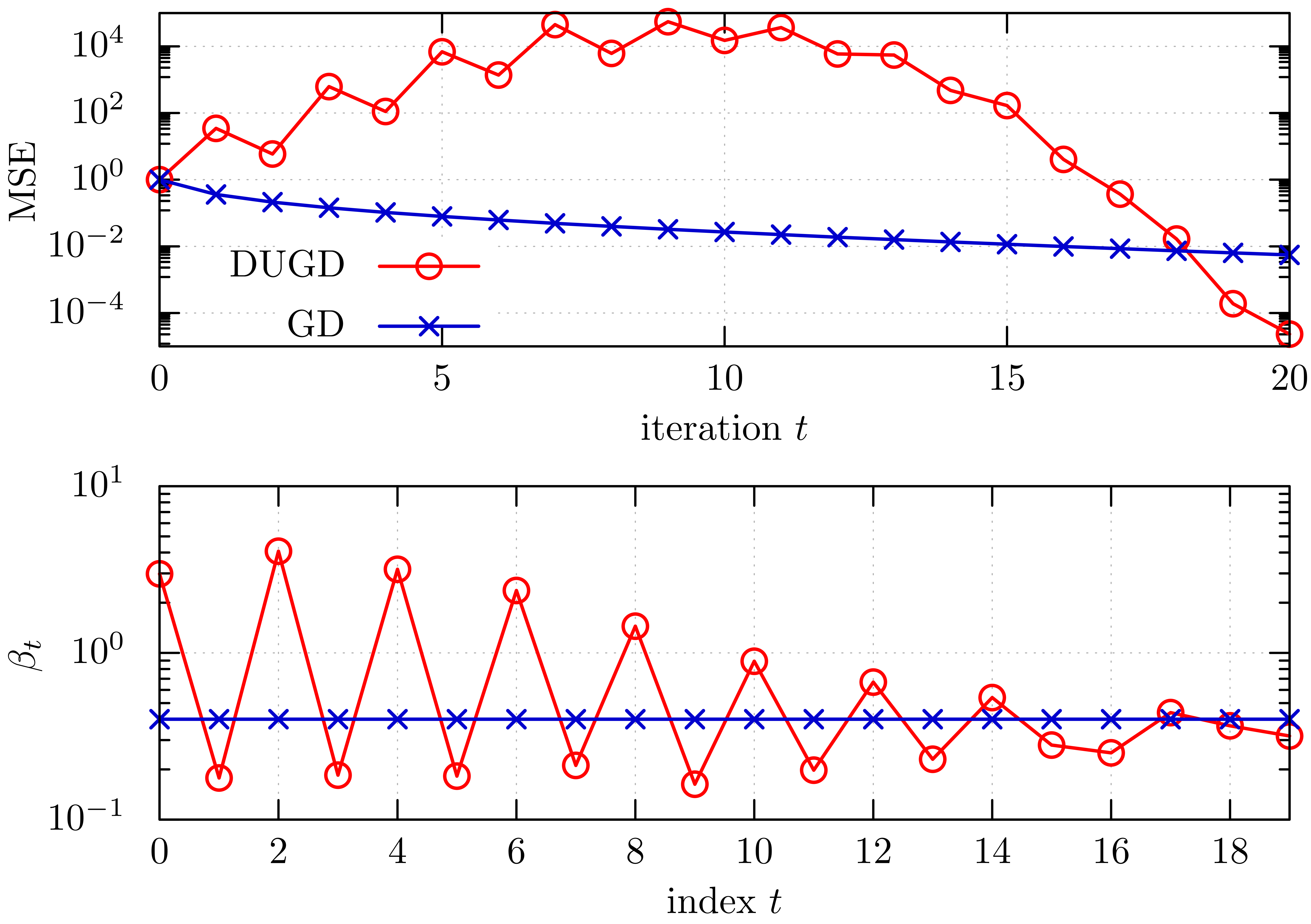
勾配法(演習課題)

学習パラメータの「ジグザグ形状」が見られる





勾配法



MSE
10^4
10^2
10^0
10^{-2}
10^{-4}
DUGD
GD
0
5
10
15
20
iteration t
β_t
10^1
10^0
10^{-1}
0
2
4
6
8
10
12
14
16
18
index t

ジグザグパラメータの謎

- 深層展開において、時変ステップサイズパラメータを学習可能とする場合には、ほとんどの場合に「ジグザグ形状」の
- 固定ステップサイズの最適な選び方はよく知られている
- ジグザグ的選択でなぜ収束が高速になるのか？

→ 研究の最前線

深層展開に基づく信号処理アルゴリズムの設計
—収束加速とその理論的解釈—, 和田山・高邉

https://www.jstage.jst.go.jp/article/essfr/14/1/14_60/_pdf/-char/ja

勾配法 再訪

目的関数 $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$ (Aは正定値対称行列とする)

勾配ベクトル $\nabla f(\mathbf{x}) = \mathbf{A}\mathbf{x}$

勾配法ステップ $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \gamma_k \nabla f(\mathbf{x}^{(k)})$

書き直すと

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \gamma_k \mathbf{A}\mathbf{x}^{(k)} = (\mathbf{I} - \gamma_k \mathbf{A})\mathbf{x}^{(k)}$$

固定ステップの場合の収束性(1)

$$\mathbf{x}^{(k+1)} = (\mathbf{I} - \gamma \mathbf{A})\mathbf{x}^{(k)}$$

$\mathbf{B} = \mathbf{I} - \gamma \mathbf{A}$ の**スペクトル半径が1未満の場合に収束**する！

固有値とスペクトル半径

n x n 正定値対称行列Qの固有値λは

$$\mathbf{Q}\mathbf{x} = \lambda \mathbf{x}$$

を満たす

Qの固有値を$\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$とする。

* 対称性からすべての固有値は実数
* 正定値性からすべての固有値は正

スペクトル半径

$$\rho(\mathbf{Q}) = \max\{|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|\}$$

固定ステップの場合の収束性(2)

$$\mathbf{x}^{(k+1)} = (\mathbf{I} - \gamma\mathbf{A})\mathbf{x}^{(k)}$$

$$\|\mathbf{x}^{(k+1)}\| \leq \|\mathbf{I} - \gamma\mathbf{A}\|\|\mathbf{x}^{(k)}\| = \rho(\mathbf{I} - \gamma\mathbf{A})\|\mathbf{x}^{(k)}\|$$

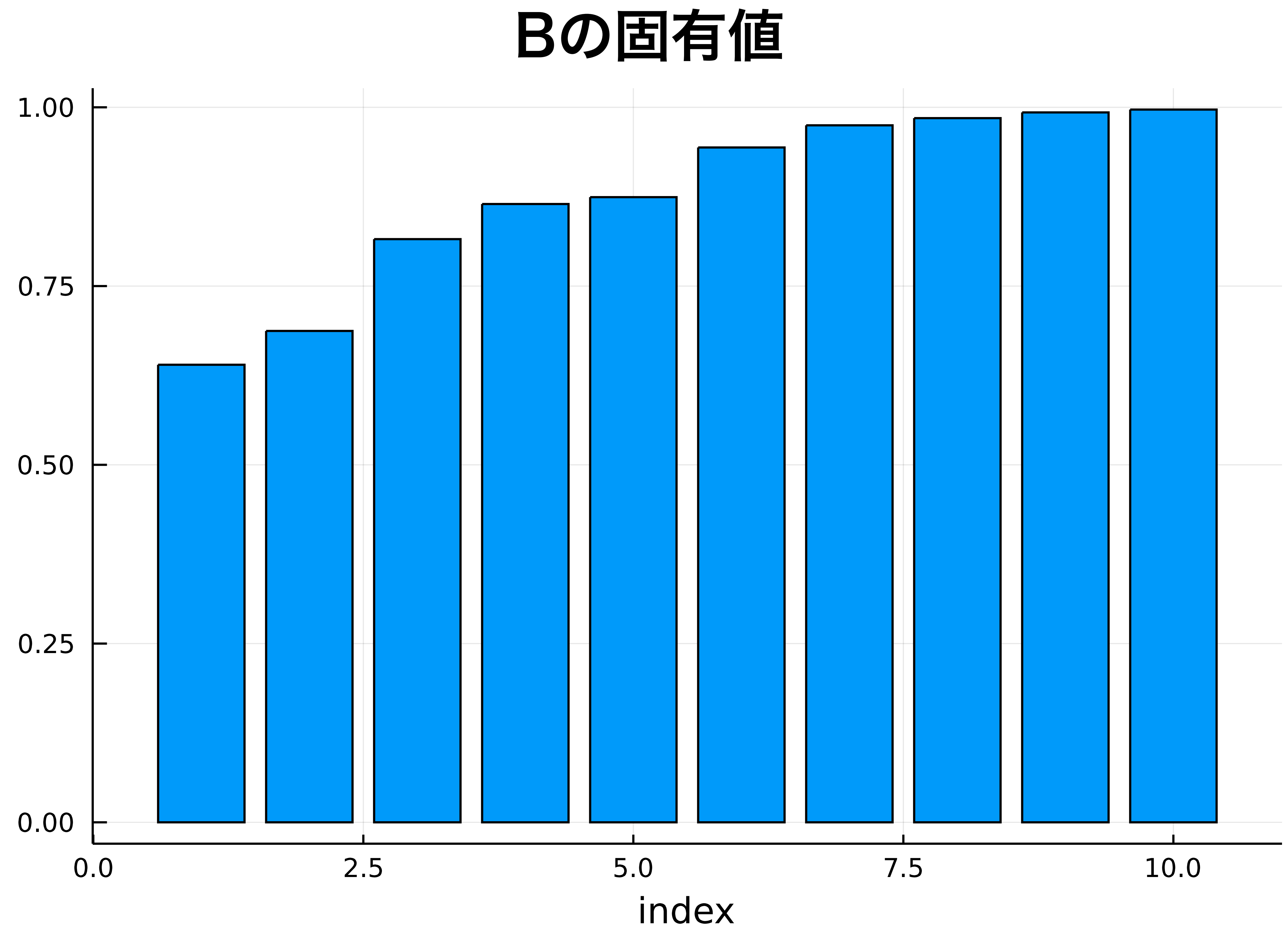
- スペクトル半径が1未満ならば、ステップ毎にx^{k}のノルムはどんどん小さくなる⇒原点（この例の最適解)に収束する
- スペクトル半径が小さいほど収束が高速になる
- γを決めるときには、スペクトル半径が小さくなるように選ぶ

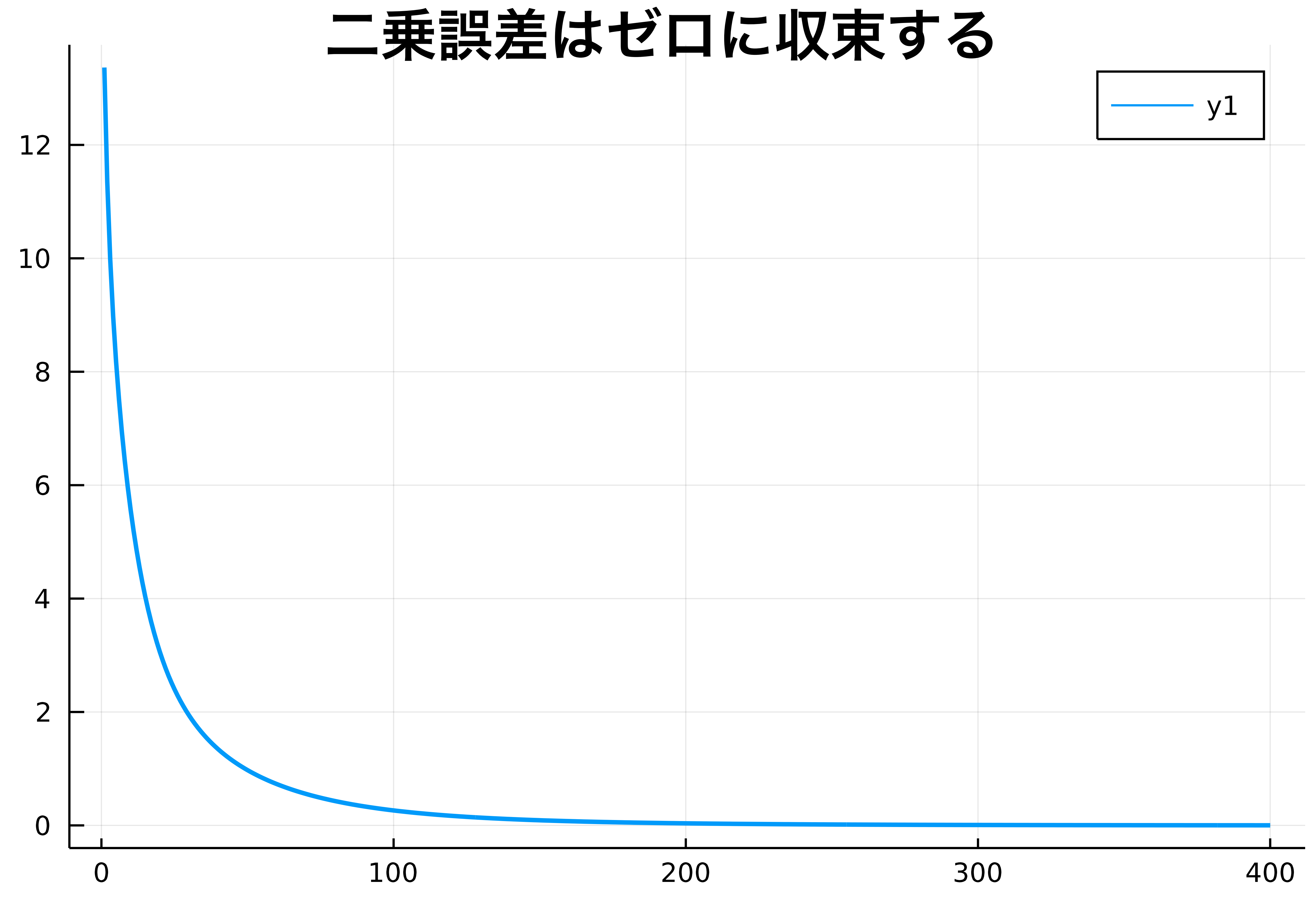
勾配法の振る舞い（1）

$$\mathbf{x}^{(k+1)} = (\mathbf{I} - \gamma \mathbf{A})\mathbf{x}^{(k)}$$

Aは10x10ランダム行列

$$\gamma = 0.01$$





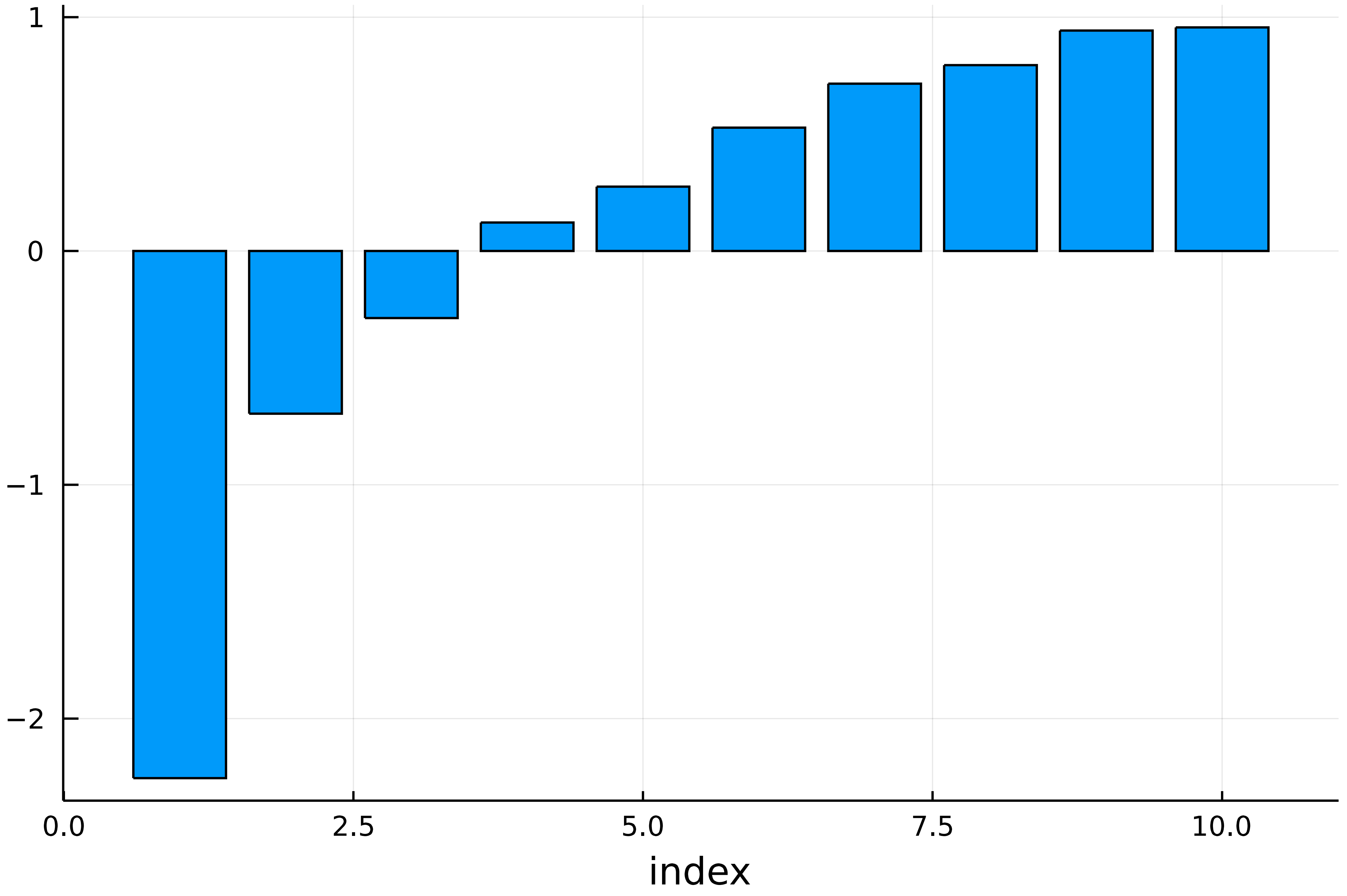
勾配法の振る舞い（2）

$$\mathbf{x}^{(k+1)} = (\mathbf{I} - \gamma\mathbf{A})\mathbf{x}^{(k)}$$

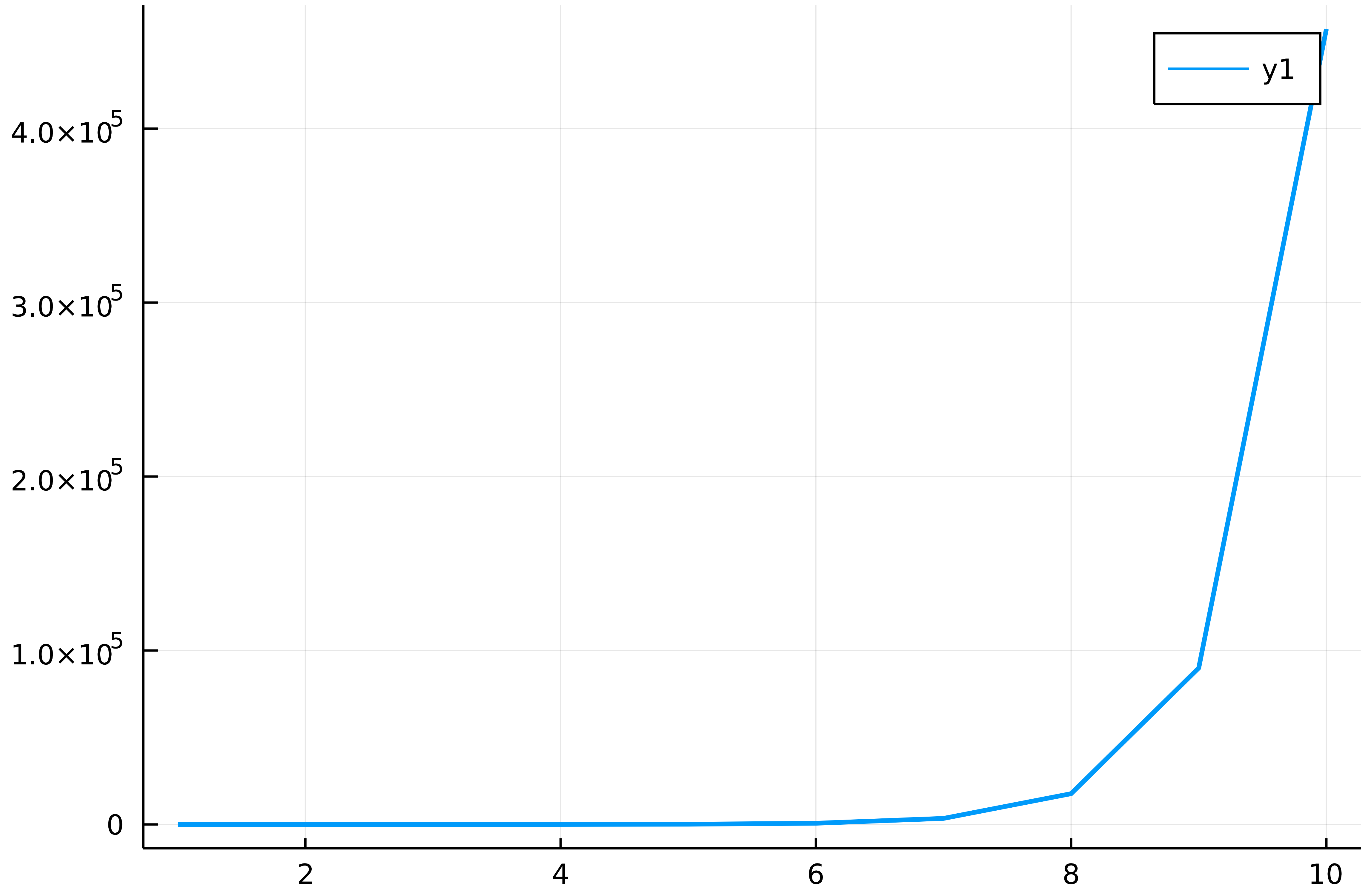
Aは10x10ランダム行列

$$\gamma = 0.1$$

Bの固有値



二乗誤差は発散する



周期的時変ステップを考える(1)

$$\mathbf{x}^{(k+1)} = (\mathbf{I} - \gamma_k \mathbf{A})\mathbf{x}^{(k)}$$

ただし、以下では周期的条件(周期T)

$$\gamma_0, \gamma_1, \ldots, \gamma_{T-1}, \gamma_T = \gamma_0, \gamma_{T+1} = \gamma_1, \cdots$$

を仮定する

例(T=2) $\gamma_0 = 1, \gamma_1 = 2, \gamma_2 = 1, \gamma_3 = 2, \gamma_4 = 1, \gamma_5 = 2, \ldots$

周期的時変ステップを考える(2)

$$\mathbf{x}^{(\ell T+1)} = (\mathbf{I} - \gamma_0 \mathbf{A})\mathbf{x}^{(\ell T)}$$

$$\mathbf{x}^{(\ell T+2)} = (\mathbf{I} - \gamma_1 \mathbf{A})(\mathbf{I} - \gamma_0 \mathbf{A})\mathbf{x}^{(\ell T)}$$

$$\mathbf{x}^{(\ell T+3)} = (\mathbf{I} - \gamma_2 \mathbf{A})(\mathbf{I} - \gamma_1 \mathbf{A})(\mathbf{I} - \gamma_0 \mathbf{A})\mathbf{x}^{(\ell T)}$$

$$\vdots$$

$$\mathbf{x}^{((\ell+1)T)} = \left(\prod_{j=0}^{T-1} (\mathbf{I} - \gamma_j \mathbf{A}) \right) \mathbf{x}^{(\ell T)}$$

Tステップ遷移行列Q

$$\mathbf{x}^{((\ell+1)T)} = \left(\prod_{j=0}^{T-1} (\mathbf{I} - \gamma_j \mathbf{A}) \right) \mathbf{x}^{(\ell T)}$$

$$\mathbf{Q} = \left(\prod_{j=0}^{T-1} (\mathbf{I} - \gamma_j \mathbf{A}) \right)$$

のスペクトル半径が収束のスピードを決める

γ_jをどう定めればスペクトル半径を小さくできるか？

スペクトル写像定理

多項式　$f(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 + \cdots + a_m\lambda^m$

行列Qの固有値を$\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$とする

行列多項式

$$f(\mathbf{Q}) = a_0\mathbf{I} + a_1\mathbf{Q} + a_2\mathbf{Q}^2 + \cdots + a_m\mathbf{Q}^m$$

の固有値は$\{f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n)\}$となる

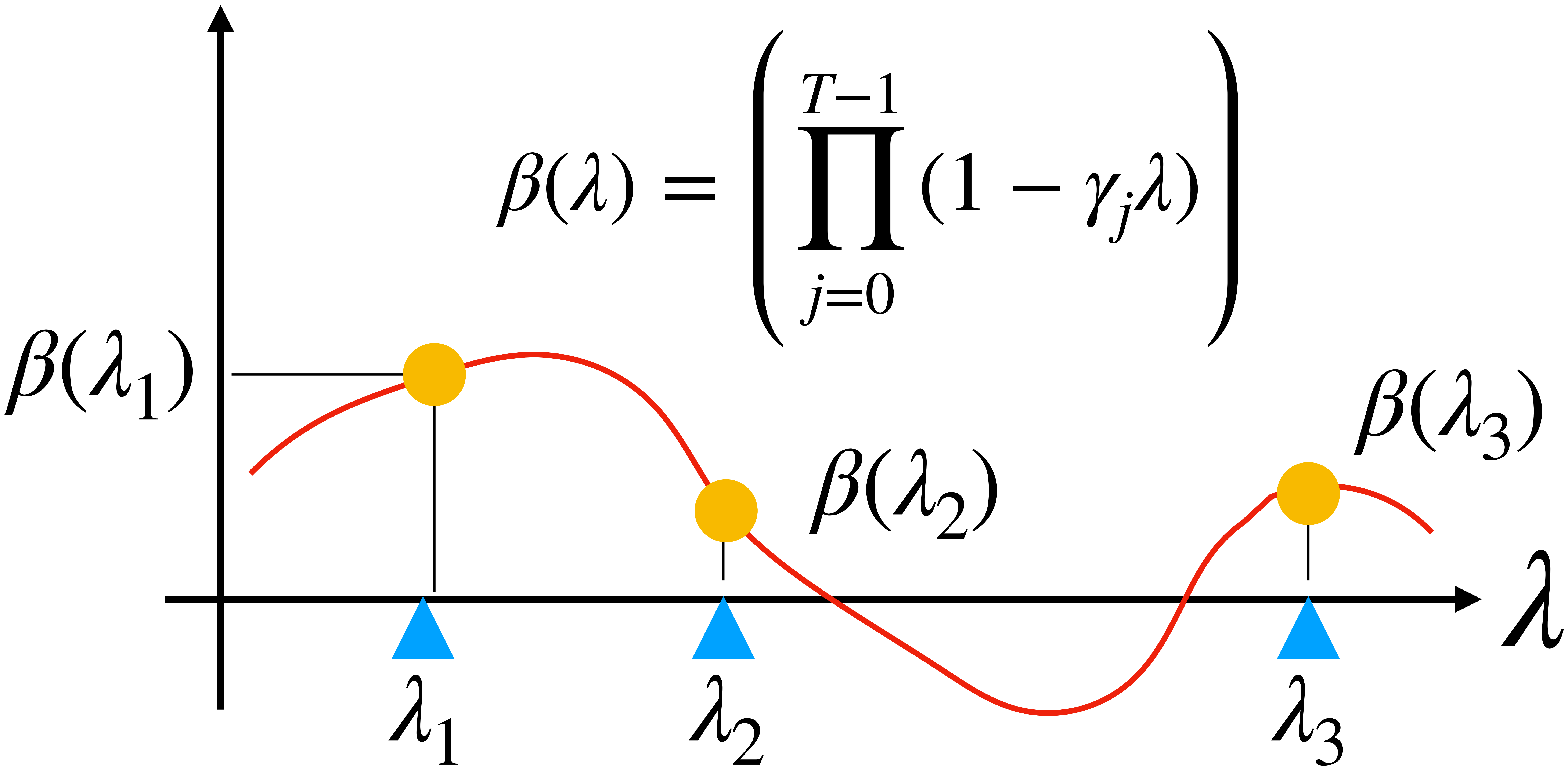
よいステップサイズ列の決定方針（1）

$$\mathbf{Q} = \left(\prod_{j=0}^{T-1} (\mathbf{I} - \gamma_j \mathbf{A}) \right)$$

⇒対応する多項式を作る

$$\beta(\lambda) = \left(\prod_{j=0}^{T-1} (1 - \gamma_j \lambda) \right)$$

よいステップサイズ列の決定方針（2）



$\beta(\lambda)$として、「一定範囲内で収まりの良い」多項式を選ぶの有利

チェビシェフ多項式

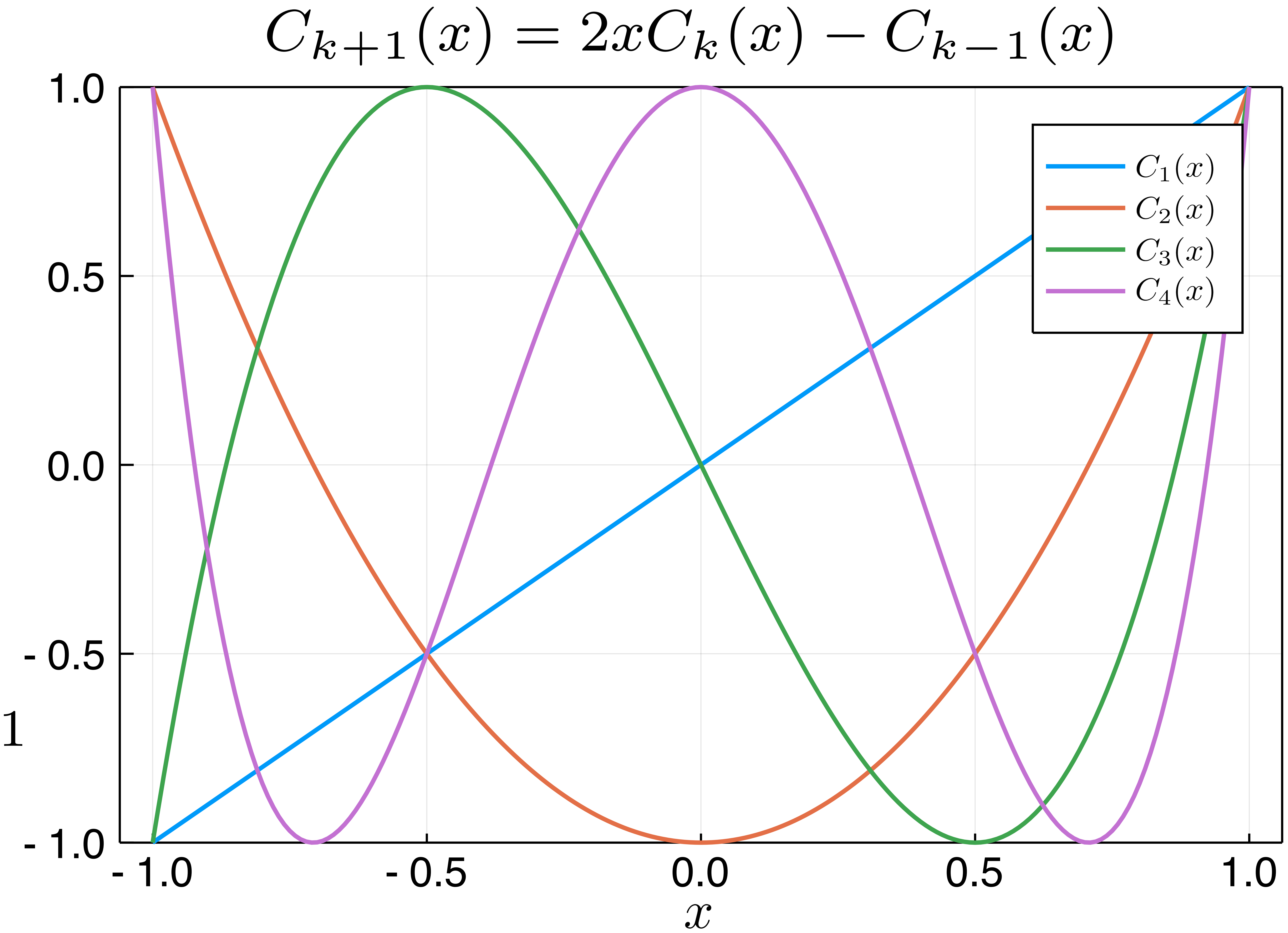
$$C_0(x) = 1$$

$$C_1(x) = x$$

$$C_2(x) = 2x^2 - 1$$

$$C_3(x) = 4x^3 - 3x$$

$$C_4(x) = 8x^4 - 8x^2 + 1$$



チェビシェフステップ系列

$$\gamma_k = \left[\frac{\lambda_n + \lambda_1}{2} + \frac{\lambda_n - \lambda_1}{2}\cos\left(\frac{2k+1}{2T}\pi\right)\right]^{-1},\ k = 0,1,\ldots,T-1$$

λ_1, λ_nはそれぞれAの最大・最小固有値

$[\lambda_1, \lambda_n]$にチェビシェフ多項式を写した

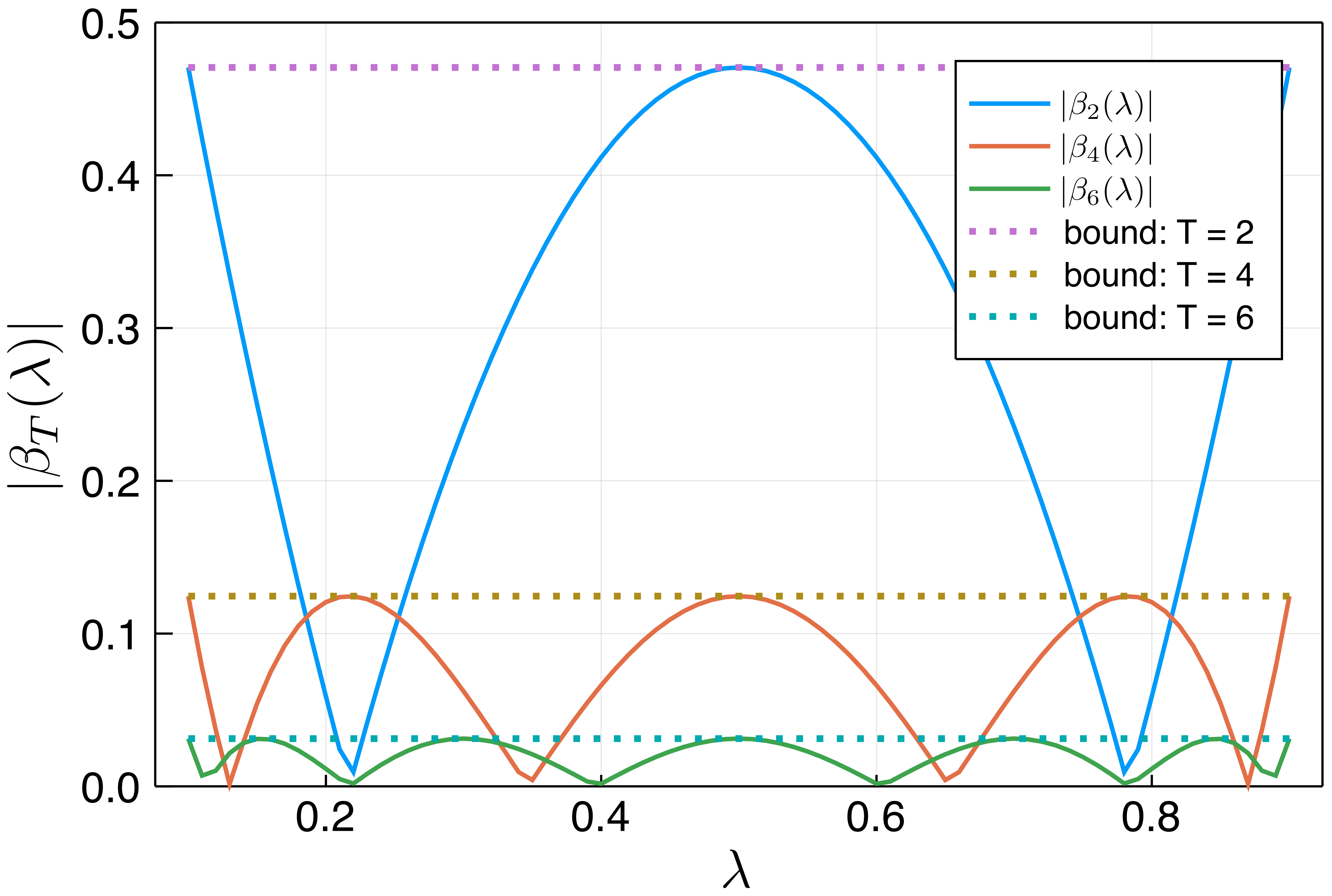
ときの根(チェビシェフ根)の逆数

Qのスペクトル半径の上界

チェビシェフステップ系列を使った場合

$$\rho(\boldsymbol{Q}) \leq \operatorname{sech}\left(T \cosh^{-1}\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)\right)$$

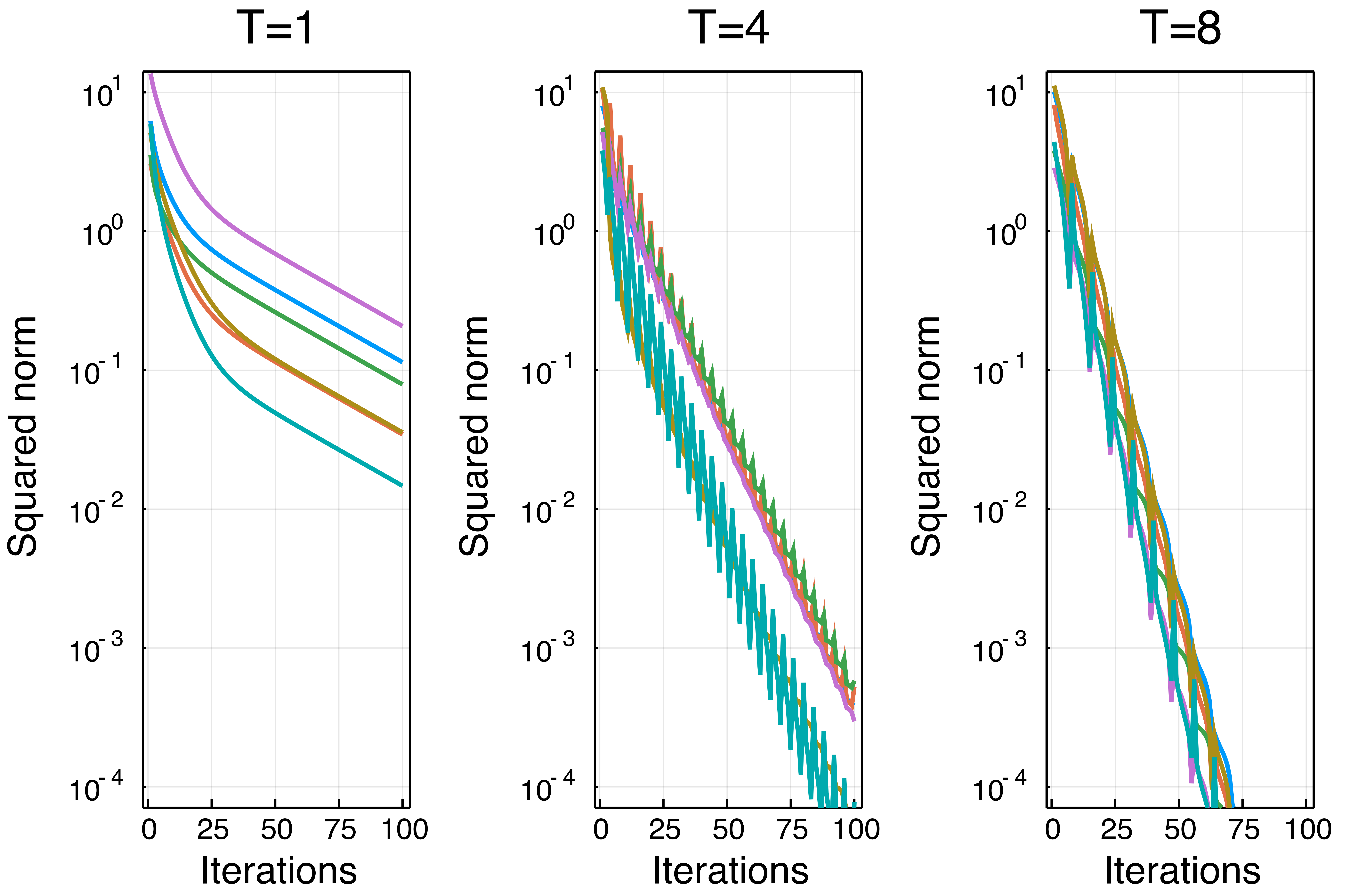
多項式 β の形状



0.5
0.4
0.3
0.2
0.1
0.0
0.2
0.4
0.6
0.8
λ
$|\beta_T(\lambda)|$
$|\beta_2(\lambda)|$
$|\beta_4(\lambda)|$
$|\beta_6(\lambda)|$
bound: T = 2
bound: T = 4
bound: T = 6

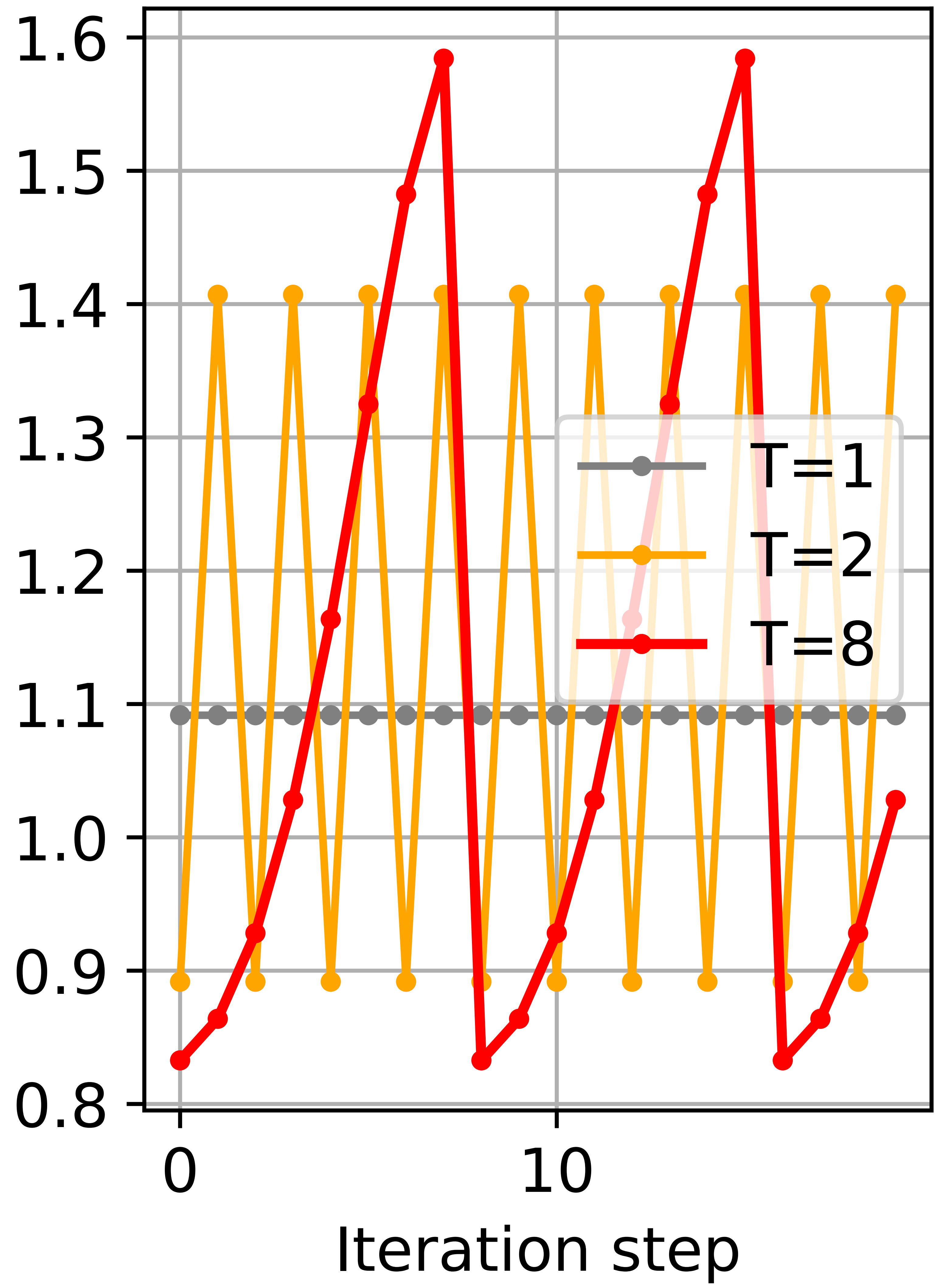
チェビシェフステップ系列による収束加速

n=10



チェビシェフステップ系列の形状

$$\gamma_k = \left[\frac{\lambda_n + \lambda_1}{2} + \frac{\lambda_n - \lambda_1}{2} \cos\left(\frac{2k+1}{2T}\pi \right) \right]^{-1}$$



深層展開による学習パラメータとチェビシェフステップ

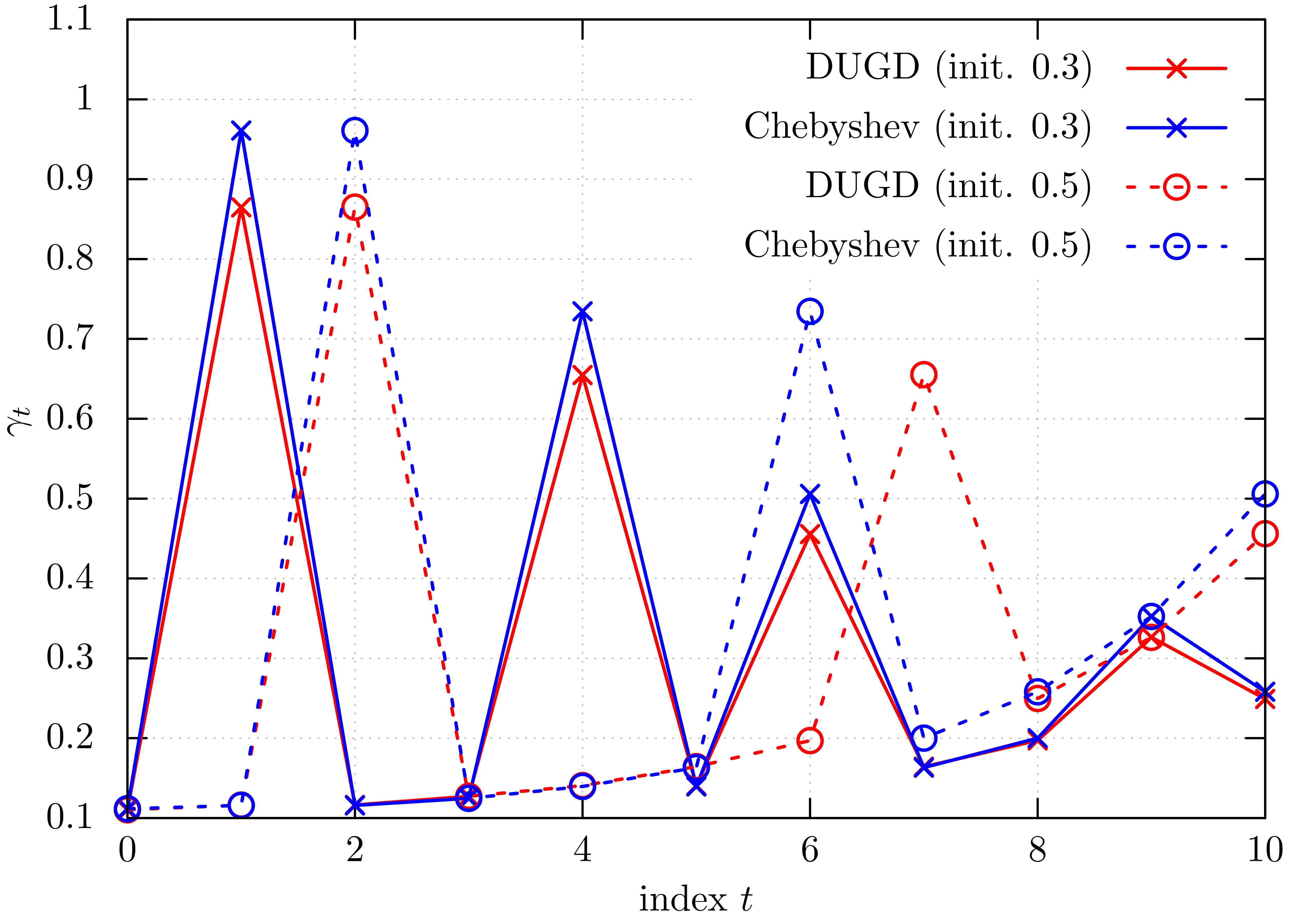


図 14　DUGD 法による学習後ステップサイズパラメータ (赤) とチェビシェフステップ（青）: $(n, m) = (250, 1000)$, $T = 11$.

平均二乗誤差特性の比較

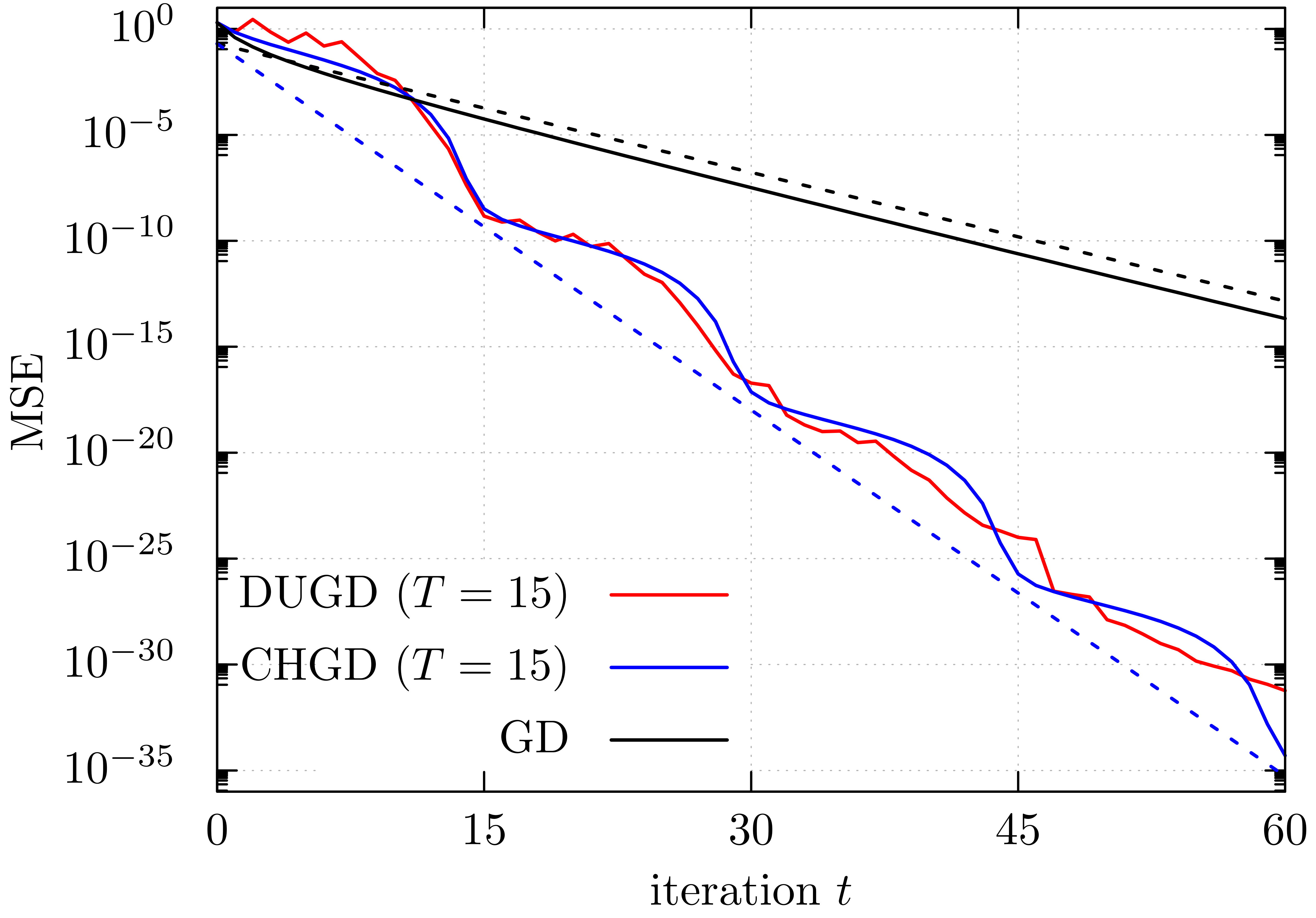


図 15　平均二乗誤差の比較 (DUGD 法，CHGD 法，GD 法): $(n, m) = (250, 1000)$ ($\kappa = 8.62$)．破線は理論値.

観察とまとめ

- チェビシェフステップ系列と深層展開による学習結果は非常に近い
- 言い換えると深層展開は、チェビシェフステップに類似したステップサイズを「自動的に見つけてきている」
- 上で見たチェビシェフステップサイズの理論は深層展開の学習結果の観察に基づき見いだされたものである
- 深層展開の学習結果をよく観察・考察することで、新しいアルゴリズム性能向上のための指針・戦略が得られることがある→データからの「法則の帰納的発見」

学習結果の解釈に基づく法則・戦略の発見

深層展開に基づくパラメータ調整 → 学習結果のパラメータの観察・考察 → 新しい法則・戦略の発見

- 機械学習技術は単なる「精度のよい推定技法」ではなく、使い方よると科学的発見の力強い道具になる
- モデルベースアプローチが有望

全体のまとめ

- 深層学習は非常に強力な機械学習技術であり、今後もさまざまな分野に応用されていくだろう
- スパース信号処理は無線通信ネットワーク６Gを含め、今後も重要であり続けるだろう
- 微分可能プログラミングや深層展開の考え方は、従来のアルゴリズム設計の方法を変えていく可能性がある
- 学習結果を緻密に観察することで新しい発見に繋がる場合がある