Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Modified iterations for data-sparse solution of linear systems

(revised version: April 2021)

by

Wolfgang Hackbusch and André Uschmajew

Preprint no.: 58 2020

Modified iterations for data-sparse solution of linear systems

Wolfgang Hackbusch[*] André Uschmajew[*]

Abstract

A modification of standard linear iterative methods for the solution of linear equations is investigated aiming at improved data-sparsity with respect to a rank function. The convergence speed of the modified method is compared to the rank growth of its iterates for certain model cases. The considered general setup is common in the data-sparse treatment of high dimensional problems such as sparse approximation and low rank tensor calculus.

1 Introduction

In this work we investigate modifications of linear fixed-point iterations for computing approximate solutions of a linear equation

$$Au = f \tag{1.1}$$

in a Banach space $\mathbf{V}$. A standard linear iterative method for solving (1.1) takes the form

$$u_{m+1} = Mu_m + Nf \tag{1.2}$$

and corresponds to a linear fixed-point equation

$$u = Mu + Nf. \tag{1.3}$$

The matrix M is called the iteration matrix of the method. When (1.3) is related to (1.1) via $M = I - NA$, then N or its inverse is called the preconditioner of A. If the choice of N is a concrete function of A, then this function defines a class of iterative solvers for all (invertible) A. For instance, for the Jacobi method N is the inverse of the diagonal part of A. In other cases it may be a polynomial of A, and so on. However, in this work we will make direct assumptions on the properties of M and Nf, which may or may not be realizable for a given A. Hence we take the fixed-point problem (1.3) as the starting point of our considerations.

In a practical implementation, if the dimension of $\mathbf{V}$ is very large or (in principle) infinite, the use of data-sparse representations of its elements is required for storing the iterates and performing the matrix-vector products. An example that motivates our work is low-rank matrix and tensor formats which can be used for the numerical treatment of high-dimensional problems, and have found many applications in scientific computing [5, 13, 17, 16, 19]. In these formats the numerical

[*] Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany
Email: wh@mis.mpg.de, uschmajew@mis.mpg.de

complexity for storing vectors and performing basic linear algebra operations is typically captured by one or several *rank functions*, $\operatorname{rank}\colon \mathbf{V} \to \mathbb{N} \cup \{+\infty\}$. These rank functions are usually sub-additive

$$\operatorname{rank}(u+v) \leq \operatorname{rank}(u) + \operatorname{rank}(v),$$

and satisfy $\operatorname{rank}(0) = 0$. More generally, such a sub-additive rank function may arise from a generating set $\mathcal{D} \subseteq \mathbf{V}$, typically called a dictionary, by defining $\operatorname{rank}(u)$ for $u \neq 0$ as the minimal number of elements from $\mathcal{D}$ needed to write u as a linear combination, or $+\infty$ if u is not in the finite span. The goal is then to find possibly sparse, that is low-rank representations of a sought solution in the dictionary. This very general concept of expansion in dictionaries occurs frequently in nonlinear approximation, and covers classical sparsity (then the dictionary consists of unit vectors) or more general best M-term approximation problems. For low-rank matrices or tensors the dictionary consists of all elementary tensors. Of course, when using data-sparse representations with respect to a dictionary, it is implicitly assumed that the true solution of the problem admits accurate 'low-rank' approximations, but verifying this analytically in advance can be difficult depending on the application. Also note that in many applications the choice of the dictionary is not only motivated by reducing the numerical complexity, but has some well defined and problem dependent purpose for revealing the structure, patterns, principal subspaces etc. of some measured data.

In this paper we investigate the rank accumulation in iterative methods like the linear fixed-point iteration (1.2) in relation to its convergence speed. When looking at (1.2) we see that rank increase occurs due to two operations: the application of the operator M and the addition of Nf. To deal with the first we assume a multiplicative model, in which the rank of Mu_m (and typically also the cost of forming Mu_m) is proportional to the rank of u_m, that is, $\operatorname{rank}(Mu_m) \leq \mu_1 \operatorname{rank}(u_m)$. Then

$$\operatorname{rank}(u_{m+1}) \leq \mu_1 \operatorname{rank}(u_m) + \operatorname{rank}(Nf). \tag{1.4}$$

For several steps one can either turn this into an exponentially growing bound, or draw upon refined estimates on how powers M^ℓ or polynomials $p(M)$ increase the rank.

Here we can mention two examples. In sparse approximation in $\mathbf{V} = \mathbb{R}^n$, when the rank of a vector is defined as the number of its nonzero elements, a banded matrix M can be efficiently applied to a sparse vector, but will increase the number of nonzero entries by a multiple of the bandwidth. The bandwidth of M^ℓ, however, does not grow exponentially but only linearly in ℓ. In fact, for the same reason the bandwith of any polynomial $p_\ell(M)$ of degree at most ℓ grows only linearly. Note that if in the initial linear equation (1.1) the matrix A has a banded structure and $N = D$ is a diagonal preconditioner (e.g. in the Jacobi method), then $M = I - DA$ has the same band structure.

As a second example assume that (1.3) is a fixed point matrix equation in $\mathbf{V} = \mathbb{R}^{n \times n}$ and M is a Kronecker product operator, $M = \tilde{M}_1 \otimes \tilde{M}_2$. It is well known that a matrix Sylvester equation, that is, (1.1) with $\mathbf{A} = \tilde{A}_1 \otimes I + I \otimes \tilde{A}_2$, can be transformed into such a problem with M having spectral radius less than one, if both $\tilde{A}_1$ and $\tilde{A}_2$ have eigenvalues with negative real part [25]. The Kronecker product operator M can then be efficiently applied to a low-rank matrix (in the usual sense) and does not increase the rank at all. Therefore, applying a polynomial $p_\ell(M)$ of degree ℓ to a matrix increases the rank at most by a factor of $\ell + 1$; see section 3.1 for a more general example.

Our main attention in this work is on the second step in the iteration (1.2) which is the addition of Nf. In standard applications, Nf is a fully populated vector and the inequality (1.4) indicates that even one such step is infeasible if $\operatorname{rank}(Nf)$ is very large or infinite. The standard approach would be replacing Nf by an approximation $\tilde{Nf}$ of acceptable rank, and will be discussed in section 3.2. As an alternative, we propose using a sequence $g_m \to Nf$ of approximations with (usually) growing rank. This leads to the modified fixed-point iteration

$$\hat{u}_{m+1} = M\hat{u}_m + g_{m+1} \tag{1.5}$$

considered in this paper. It turns out that the $\hat{u}_m$ converge to a solution u of (1.3) at a similar speed as the standard iteration if the convergence $g_m \to Nf$ is fast enough.

The proposed modification (1.5) of the fixed-point iteration is an interesting and easily realizable variation of the standard iteration. While several questions could be considered, we focus on its impact on the rank accumulation in certain model cases for M and Nf to show that it can be beneficial compared to the standard iteration. Specifically, to limit the scope we investigate only the cases that the approximation $g_m \to Nf$ converges exponentially fast, and that the rank amplification by the repeated application of the iteration matrix is either exponential or linear in the number of steps. The latter scenario is motivated by the examples mentioned above.

Besides its implications on the computational cost in iterative methods, our rank estimates also serve a theoretic purpose of characterizing low-rank approximability of structured linear equations since they yield upper bounds on the corresponding approximation numbers

$$\tau_r(u) = \inf_{\operatorname{rank}(v) \leq r} \frac{\|u - v\|}{\|u\|} \tag{1.6}$$

for the solution u in terms of the rank parameter r. However, the asymptotic rates obtained in this way by using linear fixed fixed-point iterations (or modifications of them) are not necessarily optimal and can be slow, and hence mainly relevant if $\mathbf{V}$ is of very large or infinite dimension. For the two mentioned examples of sparse approximation of systems with banded matrices, and matrix or tensor equations with Sylvester-type operators, better rates than those implied by our analysis (covered by Examples 3.3 and 3.7) are available for Hilbert spaces based on different and more problem related approximation schemes; see, e.g., [12, 4] and [22, 14, 20, 15, 10, 26], respectively.

Furthermore, by taking the fixed-point formulation (1.3) as the point of departure we avoid any discussion under which conditions it is actually possible to design for a given linear equation (1.1) an iterative solver (1.2) for which M is highly contractive and mildly rank increasing at the same time (which could be conflicting targets), while Nf admits fast converging and available low-rank approximations. The existence of such a linear iteration would directly imply that the solution can be well approximated in the dictionary. Clearly, if it is not available, it can be more efficient to use few steps steps of a method with a general spectrally equivalent preconditioner N and then apply truncation. This difficult question is at the very core of understanding low-rank approximability and preconditioning of linear systems for a given rank function, and should not be treated in a general setup like in this paper. There is, however, an interesting converse logic to this. If the solution u of a given linear equation (1.1) does not satisfy the approximability estimates that would be implied by certain assumptions on M and N, then it means that there cannot exist a linear iterative solver with the desired properties. While this may sound trivial, it can actually be seen as a remarkable non-existence result, say for matrix decompositions $A = N^{-1}(I - M)$.

Our results extend the studies in [21], where mainly the Richardson iteration and the steepest descent method have been considered, to the broad class of linear iterations (1.2). Of course, a great amount of other iterative methods for low-rank solutions of linear systems has been developed in the literature, among them those based on variational formulations, nonlinear optimization, or greedy methods. One common feature of such methods, that should also be applied in a practical implementation of the modified iterations considered in this paper, is the rank truncation of intermediate iterates. Truncation usually occurs either as an adaptive projection (hard thresholding) or as a prox-operator (soft thresholding), and its combination with fixed-point iterations can be studied in a quite general context; see, e.g. [11, 8, 7, 3] for sparse vectors and [18, 6, 1, 23, 2, 24] for low-rank tensors. In particular, if a certain low-rank approximability of the solution is already known or assumed, then suitable adaptive truncation schemes can lead to refined and near-optimal error estimates. This is, however, outside the scope of the present paper.

The paper is outlined as follows. In section 2 the convergence rate of the modified iteration (1.5) is estimated for the model case that the g_m approximate Nf exponentially fast. In sections 3.1–3.4 rank estimates for approximate solutions obtained from the standard and modified iteration are derived from assumptions on the rank increasing properties of the iteration matrix M. Section 3.5 presents numerical comparisons of the obtained bounds.

2 Convergence of the modified iteration

Let u be a solution to the linear fixed-point equation (1.3). Given $0 < \varepsilon \leq 1$ we seek an approximation u_ε to u of relative accuracy ε, that is,

$$\frac{\|u - u_\varepsilon\|}{\|u\|} \leq \varepsilon. \tag{2.1}$$

Such a u_ε will be called an ε-solution of the fixed-point equation (1.3). Here the choice of the norm (and hence of the Banach space $\mathbf{V}$) is usually problem dependent and can already account, e.g., for a large condition number or unboundedness of the operator A in the initial linear equation (1.1). In particular, we assume that M satisfies

$$\|M\| \leq \zeta < 1, \tag{2.2}$$

in the corresponding operator norm. It guarantees that u is the unique solution of (1.3) and that the standard iteration (1.2) converges to u for every starting point u_0 at a linear rate:

$$\|u_{m+1} - u\| \leq \zeta \|u_m - u\|.$$

Note that ζ is a property of the chosen iteration (1.2) *and* of the chosen norm.[1]

It will be convenient to use the starting value

$$u_0 = 0$$

throughout the paper. It leads to the relative error

$$\frac{\|u - u_m\|}{\|u\|} \leq \zeta^m \tag{2.3}$$

after m steps of the iteration, and hence the number of steps needed for an ε-solution is upper bounded by

$$m_{(1.2)}(\varepsilon) = \left\lceil \frac{\ln \varepsilon}{\ln \zeta} \right\rceil. \tag{2.4}$$

Recall that it is always assumed that $0 < \varepsilon \leq 1$.

We now consider the modified iteration (1.5) in which in the $(m+1)$-th step Nf is replaced by some (simpler) approximation g_{m+1}. A first general statement on this approach is the following.

Proposition 2.1. *Let u be a solution to* (1.3) *and assume* (2.2). *If $g_m \to Nf$, then for the iterates* (1.5) *$\hat{u}_m \to u$.*

[1] For a bounded operator M the iteration converges for every starting point u_0 if and only if the spectral radius $\rho(M)$ is strictly less than one. Then for any fixed $\zeta \in (\rho(M), 1)$ there exists an equivalent norm on $\mathbf{V}$ satisfying (2.2); see, e.g., [27, Appx. (58)]. While $\rho(M) < 1$ describes the asymptotic behaviour (R-linear rate) of the error in any equivalent norm, the condition (2.2) allows for error estimates in the given norm.

Proof. Let $\epsilon > 0$. For some m_ϵ we have $\|Nf - g_{m+1}\| \leq \epsilon$ for all $m \geq m_\epsilon$. Taking the difference of (1.3) with (1.5) gives

$$u - \hat{u}_{m+1} = M\left(u - \hat{u}_m\right) + Nf - g_{m+1}. \tag{2.5}$$

Hence, the error $\hat{\eta}_m := \|u - \hat{u}_m\|$ satisfies

$$\hat{\eta}_{m+1} \leq \zeta \hat{\eta}_m + \epsilon \qquad \text{for all } m \geq m_\epsilon. \tag{2.6}$$

Define the recursive sequence $\hat{\eta}'_{m+1} = \zeta \hat{\eta}'_m + \epsilon$ by fixing $\hat{\eta}'_{m_\epsilon} = \hat{\eta}_{m_\epsilon}$. Then $\hat{\eta}'_m \to \epsilon/(1-\zeta)$ and hence

$$\limsup_{m\to\infty} \hat{\eta}_m \leq \limsup_{m\to\infty} \hat{\eta}'_m = \frac{\epsilon}{1-\zeta}$$

by (2.6). Since ϵ was arbitrary, this proves the assertion. □

In the following, we assume that the g_m approximate Nf exponentially fast and satisfy

$$\|Nf - g_m\| \leq C\,\|Nf\|\,\xi^m, \qquad \text{where } \xi \leq \zeta \tag{2.7}$$

and $C > 0$ is a constant (that may depend on Nf). Then the error after m steps of the modified iteration (1.5) can be estimated. As for the standard iteration it will be convenient to consider here and in the following only the starting point

$$\hat{u}_0 = 0$$

for our estimates.

Proposition 2.2. *Let u be a solution to* (1.3)*. Assume* (2.2) *and* (2.7)*. Then it holds for the modified iteration* (1.5) *with starting point $\hat{u}_0 = 0$ that*

$$\|u - \hat{u}_m\| \leq \zeta^m \|u\| + C\,\|Nf\| \sum_{\ell=0}^{m-1} \zeta^\ell \xi^{m-\ell} = \zeta^m \|u\| + \zeta^m C\,\|Nf\| \begin{cases} \frac{1-(\xi/\zeta)^m}{(\zeta/\xi)-1} & \text{if } \xi < \zeta, \\ m & \text{if } \xi = \zeta. \end{cases}$$

The proof is an immediate induction from (2.5) and (2.7). Now note that due to $u - Mu = Nf$ it holds that

$$(1-\zeta)\,\|u\| \leq \|Nf\| \leq (1+\zeta)\,\|u\|\,.$$

From Proposition 2.2 we thus obtain the following estimate on the relative error for the modified iteration with starting point $\hat{u}_0 = 0$:

$$\frac{\|u - \hat{u}_m\|}{\|u\|} \leq \zeta^m + \zeta^m C(1+\zeta) \begin{cases} \frac{1-(\xi/\zeta)^m}{(\zeta/\xi)-1} & \text{if } \xi < \zeta, \\ m & \text{if } \xi = \zeta. \end{cases} \tag{2.8}$$

This should be compared with (2.3).

Example 2.3. *If we choose the g_{m+1} such that $C = 1$ and $\xi = \zeta/2$ in* (2.7)*, then* (2.8) *implies*

$$\frac{\|\hat{u}_m - u\|}{\|u\|} \leq \zeta^m (2+\zeta).$$

To obtain a relative error ε, the modified iteration (1.5) *hence requires at most*

$$m_{(1.5)}(\varepsilon) = \left\lceil \frac{\ln \varepsilon - \ln(2+\zeta)}{\ln \zeta} \right\rceil \tag{2.9}$$

steps. The standard iteration (1.2) *needs only* $m_{(1.2)}(\varepsilon) = \lceil \frac{\ln \varepsilon}{\ln \zeta} \rceil$ *steps. Note, however, that* $\ln(2+\zeta) < \ln(3) < 1.1$*. Therefore in the case of a fast standard iteration, when* $\ln \zeta$ *is not close to zero, the number of additional steps in the modified iteration is very small. For instance with* $\zeta = \frac{1}{2}$ *the additional term* $\frac{\ln(2+\zeta)}{|\ln \zeta|}$ *equals* 1.322*, so the modified iteration needs at most two steps more than the standard iteration.*

Generalizing this example we can derive a bound for the required number of steps for reaching a certain accuracy by estimating the inverse function of the right-hand side in (2.8). In the worst case $\xi = \zeta$ this requires some effort. We have the following result.

Proposition 2.4. *Let* u *be a solution to* (1.3)*. Assume* (2.2) *and* (2.7)*. For* $0 < \varepsilon \le 1$*, let* $m_{(1.5)}(\varepsilon)$ *be the number of steps needed for the modified iteration* (1.5) *to reach an* ε*-solution satisfying* (2.1)*. In case* $\xi < \zeta$ *it holds that*

$$m_{(1.5)}(\varepsilon) \le \left\lceil \frac{\ln \varepsilon}{\ln \zeta} + K_1(\zeta, \xi, C) \right\rceil, \qquad K_1(\zeta, \xi, C) = \frac{\ln \left(1 + \frac{C(1+\zeta)}{(\zeta/\xi)-1}\right)}{|\ln \zeta|}, \tag{2.10}$$

whereas in case $\xi = \zeta$ *it holds that*

$$m_{(1.5)}(\varepsilon) \le \left\lceil \frac{\ln \varepsilon}{\ln \zeta} + K_2(\zeta, C) + \sqrt{\frac{2}{\ln \zeta^{-1}}} \sqrt{\frac{\ln \varepsilon}{\ln \zeta} + K_2(\zeta, C) + \frac{1}{\ln \zeta} + \frac{1}{C(1+\zeta)}} \right\rceil, \tag{2.11}$$

with

$$K_2(\zeta, C) = \frac{\ln \ln(\zeta^{-1/(C(1+\zeta))})}{\ln \zeta} = \frac{\left|\ln \ln \zeta^{-1}\right| + \ln(C(1+\zeta))}{|\ln \zeta|}.$$

The proof for $\xi < \zeta$ simply follows from (2.8) by omitting the term $(\xi/\zeta)^m$ and rearranging for m. The case $\xi = \zeta$ is treated as Lemma A.1 in the appendix, where also the accuracy of the bound (2.11) is illustrated in Figure A.1. It shows that the estimate is reasonably good, but too pessimistic for ζ close to one. Both constants $K_1(\zeta, \xi, C)$ and $K_2(\zeta, C)$ are unbounded for $\zeta \to 1$.

Recall that $m_{(1.2)}(\varepsilon) = \lceil \frac{\ln \varepsilon}{\ln \zeta} \rceil$ is the iteration bound for an ε-solution with the standard iteration. If ζ/ξ is sufficiently large, then (2.10) shows that the number of additional steps required by the modified iteration for reaching the same accuracy is effectively constant, and indeed small if ζ itself is very small, see Example 2.3. In the case $\xi = \zeta$ we can roughly state that

$$m_{(1.5)}(\varepsilon) \le m_{(1.2)}(\varepsilon) + K_2(\zeta, C) + O\left(\sqrt{m_{(1.2)}(\varepsilon) + K_2(\zeta, C)}\right),$$

but with a constant that behaves like $1/\ln \zeta^{-1}$ when $\zeta \to 1$. In practice, for a fixed ζ, say up to $\zeta \le 0.9$, and reasonable ε, the actual number of additional steps asserted by this bound is still effectively constant, as can be seen in Figure A.1 in the appendix. For example, for a fast iteration with $\zeta = \frac{1}{2}$ and $C = 1$, (2.11) provides the bound

$$m_{(1.5)}(\varepsilon) \le \left\lceil \frac{|\ln \varepsilon| + 0.772 + \sqrt{2\,|\ln \varepsilon| + 0.469}}{\ln 2} \right\rceil$$

for the case $\xi = \zeta$. For small ε this is considerably worse than (2.9), where $\xi = \zeta/2$, but in turn this bound is actually valid for all possible $\xi \le \zeta = \frac{1}{2}$.

We conclude this section by mentioning a further possible modification of the standard linear iteration, in which instead of a fixed iteration matrix M a sequence $M_m \to M$ is used. This leads to iterations of the form

$$\bar{u}_{m+1} = M_{m+1}\bar{u}_m + g_{m+1}.$$

The matrix M_m could be implicitly given by a fixed linear iterative solver applied to a family of approximations $A_m \to A$ of the linear system itself, or by a sequence $N_m \to N$ of preconditioners. Assuming (2.2), it is not difficult to prove that if $g_m \to Nf$ and $M_m \to M$, then $\bar{u}_m \to u$, the fixed point of (1.3), the argument is similar to the proof of Proposition 2.1. Based on suitable assumptions on the convergence speed $M_m \to M$ one can then study error estimates. In this work, however, we restrict our attention to the simpler variation (1.5) with M being fixed.

3 Rank growth in the standard and modified iteration

The modified iteration (1.5) will usually need some more steps than the standard iteration (1.2) to reach a target accuracy ε for the relative error, which is indicated by the error estimates stated in the previous section. In turn the rank of the iterates may grow a little bit less per step, since we are adding g_{m+1} instead of Nf. In this section we compare the achievable accuracy with the accumulated representation ranks of the approximate solutions generated by the standard iteration and its modification in simplified model cases. While the results are perhaps too generic (and thereby too pessimistic) to use when studying a particular linear equation, our aim is to show that the modified iteration can provide some improvement. We mention again that rank truncation during the iteration is not considered in our analysis, but recommended in computations. The required ranks for a certain accuracy in practice hence can be much smaller than the rank bounds obtained below.

3.1 Rank growth in the standard iteration

Due to the representation

$$u_m = \left(\sum_{\ell=0}^{m-1} M^\ell \right) Nf \tag{3.1}$$

the ranks for the iterates of the standard iteration (with $u_0 = 0$) can be estimated in terms of the following, in general unknown, constants:

$$\nu_m = \nu_m(M) = \sup_{v \neq 0} \frac{\operatorname{rank}\left(\sum_{\ell=0}^{m-1} M^\ell v \right)}{\operatorname{rank}(v)}.$$

Using these constants we have the following obvious estimate from (3.1).

Proposition 3.1. *Consider the standard linear fixed-point iteration* (1.2) *with starting point* $u_0 = 0$. *Then*

$$\operatorname{rank}(u_m) \leq \nu_m \operatorname{rank}(Nf). \tag{3.2}$$

In general all the ν_m could be infinite or equal to the dimension of $\mathbf{V}$. A basic assumption in our paper is that at least for small m the ν_m are small compared to the dimension of $\mathbf{V}$ and do not grow too fast. But even then, since in the definition of ν_m we have taken a supremum, the estimate (3.2) is quite generic and our results will not account for any additional structure that could be exploited when applying powers of M to Nf in a particular instance. Another issue is that the estimate (3.2) is only reasonable if $\operatorname{rank}(Nf)$ is finite. Let us assume this, then together with the convergence speed (2.4) one obtains rank bounds for an ε-solution of the fixed-point equation (1.3), depending on the behaviour of the constants ν_m.

When the constants ν_m are not known, it is possible to estimate them by the constants

$$\mu_\ell = \mu_\ell(M) = \sup_{v \neq 0} \frac{\operatorname{rank}(M^\ell v)}{\operatorname{rank}(v)},$$

which can be easier to determine. In most cases we may rightfully assume $\mu_1 > 1$. Then clearly,

$$\mu_\ell \leq \mu_1^\ell,$$

and therefore

$$\nu_m \leq \sum_{\ell=0}^{m-1} \mu_\ell \leq \sum_{\ell=0}^{m-1} \mu_1^\ell. \tag{3.3}$$

We call the upper bound in this estimate the worst-case behaviour, since in the typical case $\mu_1 > 1$ it indicates exponential rank growth in the standard iteration. It leads to rather pessimistic results.

Example 3.2. *Consider the worst-case behaviour* (3.3) *with* $\mu_1 > 1$. *Then* (3.2) *yields*

$$\operatorname{rank}(u_m) \leq \left(\frac{\mu_1^m - 1}{\mu_1 - 1}\right) \operatorname{rank}(Nf).$$

With (2.4) *it implies that for* $\varepsilon > 0$, *there exists an* ε*-solution* u_ε *for the linear equation* (1.1) *satisfying* (2.1) *and with a rank bounded by*

$$\operatorname{rank}(u_\varepsilon) \leq \left(\frac{\mu_1^{\lceil \frac{\ln \varepsilon}{\ln \zeta} \rceil} - 1}{\mu_1 - 1}\right) \operatorname{rank}(Nf) = O\left(\varepsilon^{\frac{\ln \mu_1}{\ln \zeta}}\right) \operatorname{rank}(Nf)$$

for $\varepsilon \to 0$. *If* Nf *has finite rank we can deduce an algebraic decay rate for the best low-rank approximation error of* u *with respect to the rank, namely*

$$\tau_r(u) = O\left(r^{\frac{\ln \zeta}{\ln \mu_1}}\right)$$

for $r \to \infty$, *where* τ_r *are the approximation numbers defined in* (1.6).

There exist interesting examples for which the ν_m do not increase exponentially. As mentioned in the introduction, for sparse approximation in $\mathbb{R}^n$ (rank being number of nonzero elements) a banded matrix M with bandwidth $1 + b$ will increase the number of nonzero elements of a vector by at most a factor $\mu_1 \leq 1 + b$. However, it holds $\mu_\ell \leq 1 + \ell b$, since M^ℓ has bandwidth $1 + \ell b$. Indeed, the band support for different powers of M is nested so that

$$\nu_m \leq 1 + (m - 1)b$$

As another example, assume M is of the form

$$M = M_1 + M_2$$

where both M_1 and M_2 do not increase the rank when applied to any u, that is, $\mu_1(M_1) \leq 1$ and $\mu_1(M_2) \leq 1$. Assume furthermore that M_1 and M_2 commute. Then

$$M^\ell = \sum_{k=0}^{\ell} \binom{\ell}{k} M_1^k M_2^{\ell-k}$$

shows that in such a case we have

$$\mu_\ell \leq \ell + 1.$$

This implies

$$\nu_m \leq \frac{m(m+1)}{2}.$$

However, in special cases one can go further. Assume additionally that $p(M_2)$ is rank-preserving for any polynomial p. Then

$$\sum_{\ell=0}^{m-1} M^\ell = \sum_{\ell=0}^{m-1}\sum_{k=0}^{\ell}\binom{\ell}{k} M_1^k M_2^{\ell-k} = \sum_{k=0}^{m-1} M_1^k\left(\sum_{\ell=k}^{m-1}\binom{\ell}{k} M_2^{\ell-k}\right)$$

implies

$$\nu_m \leq m. \tag{3.4}$$

For matrix equations an operator of the form,

$$M = \tilde{M}_1 \otimes \tilde{M}_2 + I \otimes \tilde{M}_3,$$

has the considered properties, provided $\tilde{M}_2$ and $\tilde{M}_3$ commute. This includes the Kronecker product operators ($\tilde{M}_3 = 0$) and Sylvester-type operators ($\tilde{M}_2 = I$). Both examples can be generalized to operators of such form on tensor spaces, but in the case of the Sylvester-like structure, ν_m becomes a polynomial of higher order.

Example 3.3. *If the linear rank growth* (3.4) *is assumed, then* (3.2) *becomes*

$$\operatorname{rank}(u_m) \leq m \operatorname{rank}(Nf).$$

Hence, if $\operatorname{rank}(Nf)$ *is finite, an ε-solution exists satisfying*

$$\operatorname{rank}(u_\varepsilon) \leq \left\lceil \frac{\ln\varepsilon}{\ln\zeta} \right\rceil \operatorname{rank}(Nf)$$

It implies a super-algebraic decay rate

$$\tau_r(u) = O\left(\zeta^r\right)$$

for the best rank-r approximation error of the solution u to a fixed-point equation (1.3).

More generally, for a polynomial growth $\nu_m \leq p(m)$, where p is a polynomial of degree q, one obtains $\operatorname{rank}(u_\varepsilon) = O((\ln\varepsilon/\ln\zeta)^q)$ *and* $\tau_r(u) = O(\zeta^{(r^{1/q})})$.

3.2 Standard iteration with fixed approximation of Nf

Now we discuss the case that Nf has very large or infinite rank. In practice, when an approximation

$$\frac{\|Nf - N\tilde{f}\|}{\|Nf\|} \leq \delta \tag{3.5}$$

is available, where $N\tilde{f}$ has finite rank, we can simply use $N\tilde{f}$ in the standard iteration. This is equivalent to solving a perturbed fixed-point equation

$$v = Mv + N\tilde{f}, \tag{3.6}$$

which in case that N is invertible corresponds to a linear equation

$$Av = \tilde{f}$$

instead of (1.1). The corresponding standard iteration reads

$$v_{m+1} = Mv_m + N\tilde{f}, \qquad v_0 = 0, \tag{3.7}$$

and converges to $v = (I - M)^{-1}N\tilde{f}$. The relative error to the original fixed point $u = (I - M)^{-1}Nf$ can be estimated as follows:

$$\frac{\|v_m - u\|}{\|u\|} \leq \frac{\|u - v\|}{\|u\|} + \frac{\|v - v_m\|}{\|\tilde{u}\|} \cdot \frac{\|v\|}{\|u\|} \leq \left(\frac{1+\zeta}{1-\zeta}\right)\delta + \frac{\|v - v_m\|}{\|v\|}\left[1 + \left(\frac{1+\zeta}{1-\zeta}\right)\delta\right]. \tag{3.8}$$

For simplicity, let us choose target accuracies of the form

$$\delta \leq \left(\frac{1-\zeta}{1+\zeta}\right)\frac{\varepsilon}{2+\varepsilon}, \qquad \frac{\|v - v_m\|}{\|\tilde{u}\|} \leq \frac{\varepsilon}{2} \tag{3.9}$$

for (3.5) and (3.7). Then (3.8) becomes

$$\frac{\|v_m - u\|}{\|u\|} \leq \varepsilon, \tag{3.10}$$

where v_m satisfies

$$\operatorname{rank}(v_m) \leq \nu_m \operatorname{rank}(N\tilde{f}) \tag{3.11}$$

by Proposition 3.1. The second inequality in (3.9) is satisfied after at most

$$m = \left\lceil \frac{\ln\varepsilon - \ln 2}{\ln\zeta} \right\rceil$$

iterations. One can now proceed as above by assuming different cases for ν_m and $\operatorname{rank}(N\tilde{f})$.

Example 3.4. *For later comparison with the modified iteration, we assume* (2.7) *holds with* $C = 1$ *and* $\operatorname{rank}(g_m) \leq m_0 \cdot m$. *Then we choose* $N\tilde{f} = g_{\tilde{m}}$ *such that* (3.5) *is satisfied for* $\delta = \left(\frac{1-\zeta}{1+\zeta}\right)\frac{\varepsilon}{2+\varepsilon}$ *as required in* (3.9). *For this, we need to truncate* (2.7) *after* $\tilde{m} = \lceil \frac{\ln\delta}{\ln\xi} \rceil$ *terms so that*

$$\operatorname{rank}(N\tilde{f}) \leq m_0 \cdot \left\lceil \frac{\ln\delta}{\ln\xi} \right\rceil = m_0 \cdot \left\lceil \frac{\ln\varepsilon + \kappa(\zeta,\varepsilon)}{\ln\xi} \right\rceil, \qquad \kappa(\zeta,\varepsilon) := \ln\left(\frac{1-\zeta}{(2+\varepsilon)(1+\zeta)}\right).$$

Assuming the worst case (3.3) *of exponential rank growth, we conclude from* (3.10) *and* (3.11) *that there exists an* ε*-solution* u_ε *for the initial fixed-point equation* (1.3) *that satisfies*

$$\operatorname{rank}(u_\varepsilon) \leq \left(\frac{\mu_1^{\left\lceil \frac{\ln\varepsilon - \ln 2}{\ln\zeta} \right\rceil} - 1}{\mu_1 - 1}\right)\operatorname{rank}(N\tilde{f}) = O\left(\mu_1^{\frac{\ln 2}{\ln\zeta^{-1}}}\,\varepsilon^{\frac{\ln\mu_1}{\ln\zeta}}\left(\frac{\ln\varepsilon + \kappa(\zeta,\varepsilon)}{\ln\xi}\right)\right), \tag{3.12}$$

where the constant only depends on μ_1.

Correspondingly, in case of linear rank growth $\mu_m \leq m$, *we obtain*

$$\operatorname{rank}(u_\varepsilon) \leq \left\lceil \frac{\ln\varepsilon - \ln 2}{\ln\zeta} \right\rceil \operatorname{rank}(N\tilde{f}) = O\left(\left(\frac{\ln\varepsilon}{\ln\zeta}\right)\left(\frac{\ln\varepsilon + \kappa(\zeta,\varepsilon)}{\ln\xi}\right)\right). \tag{3.13}$$

The bounds (3.12) *and* (3.13) *are depicted in Figures 3.1 and 3.2 further below for some values of* ζ *and* ξ.

3.3 Rank growth in the modified iteration

In the modified iteration (1.5) we can deal with the case that Nf has large rank by replacing it with a sequence g_m with growing ranks. In non-recursive form, the modified iteration with $\hat{u}_0 = 0$ reads

$$\hat{u}_m = M^{m-1} g_1 + M^{m-2} g_2 + \cdots + M^0 g_m.$$

This can also be written as

$$\hat{u}_m = \left(\sum_{\ell=0}^{m-1} M^\ell \right) g_1 + \left(\sum_{\ell=0}^{m-2} M^\ell \right) (g_2 - g_1) + \cdots + M^0 (g_m - g_{m-1}).$$

Instead of Proposition 3.1 we hence have the following rank estimates.

Proposition 3.5. *Consider the modified iteration* (1.5) *with starting point* $\hat{u}_0 = 0$. *Then*

$$\operatorname{rank}(\hat{u}_m) \leq \sum_{\ell=0}^{m-1} \mu_\ell \operatorname{rank}(g_{m-\ell}) \tag{3.14}$$

and

$$\operatorname{rank}(\hat{u}_m) \leq \sum_{\ell=1}^{m} \nu_\ell \operatorname{rank}(g_{m-\ell+1} - g_{m-\ell}) \tag{3.15}$$

where $g_0 = 0$.

For the standard iteration, knowing the behaviour of the constants ν_m or μ_ℓ is sufficient for deriving approximation results in terms of ζ. In the case of the modified iteration we also need to know how fast the ranks of g_m grow in relation to how fast the error $\|Nf - g_m\|$ tends to zero. We keep the assumption (2.7), but restrict to $C = 1$, that is,

$$\|Nf - g_m\| \leq \|Nf\| \xi^m \tag{3.16}$$

for some $\xi \leq \zeta$, and consider the simplest case that the rank of the g_m grow linearly, that is,

$$\operatorname{rank}(g_m) \leq m_0 \cdot m \tag{3.17}$$

where m_0 is a fixed constant. In combination with (2.7) this assumption is equivalent to Nf belonging to a certain approximation class defined by

$$\tau_r(Nf) \lesssim \xi^r,$$

where τ_r again are the approximation numbers (1.6). Note however that in a practical method the g_m must be available. When the rank function is defined by a dictionary $\mathcal{D}$, a most reasonable model for (3.17) is that Nf admits an initial expansion

$$Nf = \sum_{i=1}^{R} h_i, \qquad h_i \in \mathcal{D}, \tag{3.18}$$

and then approximating it by batches of m_0 terms taking

$$g_m = \sum_{j=1}^{m} (h_{(j-1)m_0+1} + \cdots + h_{jm_0}). \tag{3.19}$$

In this case

$$\operatorname{rank}(g_m - g_{m-1}) \le m_0 \tag{3.20}$$

for all m. A related approach that arises in practice is that a dictionary expansion of $f = \sum_i f_i$ is given. Then assuming that the operator N does not increase rank by more than a factor μ_N, one could take $g_m = N(f_1 + \cdots + f_m)$ so that $\operatorname{rank}(g_m - g_{m-1}) \le \mu_N$ for all m.

Clearly in (3.19) we can trade a larger batch size m_0 for a faster approximation rate. In general, if (3.16) and (3.17) hold for some sequence g_m one can define for an integer $t > 1$ the sequence

$$g'_m = g_{t\cdot m}, \tag{3.21}$$

which will satisfy (3.16) and (3.17), too, but with different constants $\xi' = \xi^t$ and $m'_0 = tm_0$. In particular, in the case that $\xi = \zeta$ we can always pass to $\xi' = \zeta^2 < \zeta$ which enables the more accurate estimate (2.10) for the required number of steps in Proposition 2.4. The difference will be illustrated in some numerical comparisons further below.

With (3.17) the rank estimate (3.14) simplifies to

$$\operatorname{rank}(\hat{u}_m) \le m_0 \sum_{\ell=0}^{m-1} \mu_\ell (m - \ell). \tag{3.22}$$

If also (3.20) holds, (3.15) simplifies to

$$\operatorname{rank}(\hat{u}_m) \le m_0 \sum_{\ell=1}^{m} \nu_\ell. \tag{3.23}$$

We next consider the same two examples for the behaviour of ν_m as in section 3.1.

Example 3.6. *As in Example 3.2 assume the worst-case scenario* (3.3) *of exponential growth. In this case both simplified bounds* (3.22) *and* (3.23) *yield*

$$\operatorname{rank}(\hat{u}_m) \le m_0 \left(\frac{\mu_1}{\mu_1 - 1}\right) \left(\frac{\mu_1^m - 1 - m + \frac{m}{\mu_1}}{\mu_1 - 1}\right). \tag{3.24}$$

For rigorous bounds on the rank of an ε-solution we can insert the estimates for $m_{(1.5)}(\varepsilon)$ provided in (2.10) *and* (2.11) *in the right-hand side of* (3.24)*. We omit the resulting formulas. The asymptotic behaviour is* $\operatorname{rank}(u_\varepsilon) \lesssim \mu_1^{\frac{\ln \varepsilon}{\ln \zeta}} \sim \varepsilon^{\frac{\ln \mu_1}{\ln \zeta}}$ *when $\xi < \zeta$, but with a constant deteriorating for $\xi \to \zeta$ and $\zeta \to 1$. It implies again $\tau_r(u) = O\big(r^{\frac{\ln \zeta}{\ln \mu_1}}\big)$ with corresponding constants. If $\xi = \zeta$ the estimate* (2.11) *technically yields* $\operatorname{rank}(u_\varepsilon) \lesssim \mu_1^{\frac{\ln \varepsilon}{\ln \zeta} + \sqrt{\frac{\ln \varepsilon}{\ln \zeta}}} \sim \varepsilon^{\frac{\ln \mu_1}{\ln \zeta}\left(1 + \sqrt{\frac{\ln \zeta}{\ln \varepsilon}}\right)}$ *(with a constant deteriorating for $\zeta \to 1$), but as explained above, in the considered model* (3.16) *and* (3.17) *we can always assume $\xi = \zeta^2 < \zeta$. Figure 3.1 further below contains the precise bounds for some combinations of ζ and ξ.*

Example 3.7. *If we proceed with the bound* (3.23) *and assume a linear rank growth, $\nu_\ell \le \ell$, as in Example 3.3, then we get the rank estimate*

$$\operatorname{rank}(\hat{u}_m) \le m_0 \left(\frac{m(m+1)}{2}\right). \tag{3.25}$$

From (2.10) *and* (2.11) *we then obtain the asymptotic bound*

$$\operatorname{rank}(u_\varepsilon) \lesssim \left(\frac{\ln \varepsilon}{\ln \zeta}\right)^2$$

for an ε-solution of (1.3)*, with different constants for the cases $\xi < \zeta$ and $\xi = \zeta$. The constants deteriorate with $\xi \to \zeta$ and $\zeta \to 1$, respectively. Some concrete values are plotted in Figure 3.2 below. We omit more detailed formulas and just note the implied approximation rate $\tau_r(u) = O(\zeta^{\sqrt{r}})$ for $r \to \infty$, with constants depending on ζ, ξ and m_0.*

3.4 Modified iteration with target accuracy for Nf

Since one seeks only for an ε-solution to the fixed-point equation (1.3) it may not be necessary to approximate Nf with higher and higher rank. Similar to the discussion for the standard iteration in section 3.2, one can terminate at some $g_{\tilde{m}} = N\tilde{f}$ satisfying $\|Nf - g_{\tilde{m}}\| \le \delta \|Nf\|$ as in (3.5) and then proceed with $g_m = g_{\tilde{m}}$ for all subsequent iterations.

We can analyse such an approach as a modified iteration

$$\hat{v}_{m+1} = M\hat{v}_m + \tilde{g}_{m+1} \tag{3.26}$$

for the perturbed fixed-point equation $v = Mv + N\tilde{f}$ as in (3.6), where we use $\tilde{g}_{m+1} = g_{m+1}$ for $m = 1, \dots, \tilde{m} - 1$ as approximations of $N\tilde{f}$, and $\tilde{g}_{m+1} = g_{\tilde{m}} = N\tilde{f}$ for $m \ge \tilde{m}$. Hence the first $\tilde{m}$ iterates are identical to the modified iteration with g_m. The competitor for this strategy is the standard iteration (3.7) that uses $N\tilde{f}$ from the start. Since the error analysis (3.8) remains valid (with $\hat{v}_m$ instead of v_m), we can aim at the same target accuracies (3.9) (with $\hat{v}_m$ instead of v_m) as the standard iteration for guaranteeing an ε-solution for the initial fixed-point equation. If (3.16) holds, this means we have to take $\tilde{m} = \lceil \frac{\ln \delta}{\ln \xi} \rceil$ as in Example 3.4. We can expect a similar estimate as (3.16) for $N\tilde{f} = g_{\tilde{m}}$, that is,

$$\|N\tilde{f} - \tilde{g}_m\| \le \tilde{C} \|N\tilde{f}\| \xi^m, \tag{3.27}$$

where $\tilde{C}$ is some not too large constant. For example, we may assume the h_i in a dictionary expansion (3.18) of Nf to be pairwise orthogonal, as would be the case in sparse approximation in $\mathbb{R}^n$ or in a singular value decomposition of a matrix. Then we can take $\tilde{C} = (1 - \xi^{2\tilde{m}})^{-1/2}$, since $\|N\tilde{f} - g_m\| \le \|Nf - g_m\| \le \|Nf\| \xi^m$ and

$$\frac{\|\tilde{N}f\|^2}{\|Nf\|^2} \ge 1 - \frac{\|Nf - g_{\tilde{m}}\|^2}{\|Nf\|^2} \ge 1 - \xi^{2\tilde{m}},$$

which yields (3.27) for $m \le \tilde{m}$ (for larger m the left side of (3.27) is zero anyway).

Let $\hat{m}$ be the number of steps required for (3.26) to reach an $(\varepsilon/2)$-solution for v, which by (3.8) will be an ε-solution for the original u. Assuming $\hat{m} = \tilde{m} + k \ge \tilde{m}$,[2] then according to (3.15) the final rank estimate will be

$$\operatorname{rank}(\hat{v}_{\hat{m}}) \le \sum_{\ell=k+1}^{\hat{m}} \nu_\ell \operatorname{rank}(g_{\hat{m}-\ell+1} - g_{\hat{m}-\ell}).$$

Example 3.8. *For exponential rank growth* (3.3)*, and assuming the model* (3.20)*, this means*

$$\operatorname{rank}(\hat{v}_{\hat{m}}) = m_0 \left(\frac{\mu_1}{\mu_1 - 1} \right) \left(\frac{\mu_1^{\tilde{m}} - \mu_1^k - \tilde{m} + \frac{\tilde{m}}{\mu_1}}{\mu_1 - 1} \right). \tag{3.28}$$

In the case of linear rank growth $\nu_m \le m$ one gets

$$\operatorname{rank}(\hat{v}_{\hat{m}}) \le m_0 \left(\frac{\hat{m}(\hat{m}+1) - k(k+1)}{2} \right). \tag{3.29}$$

[2] In principle $\hat{m}$ could be less than $\tilde{m}$, then the original rank estimates apply.

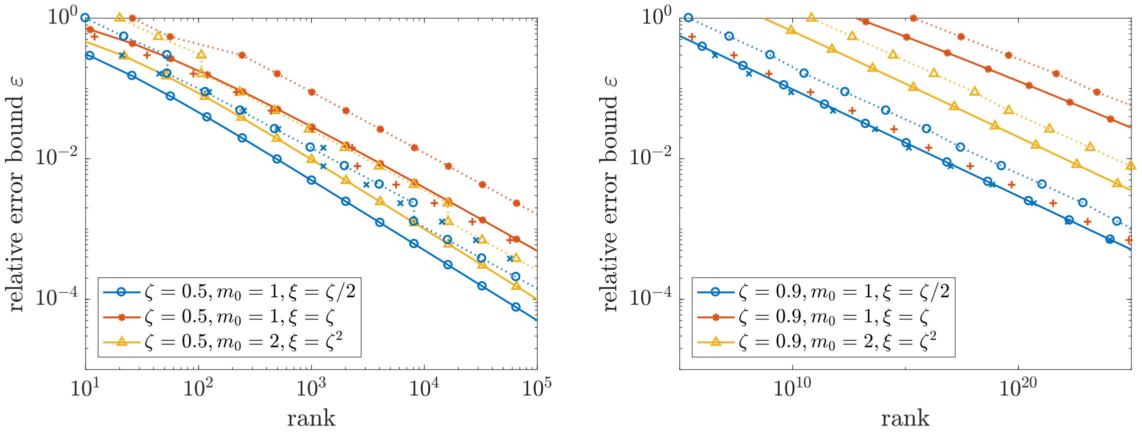

Figure 3.1: Rank bounds for ε-solutions with the modified iteration (1.5) and exponential growth $\mu_\ell \leq 2^\ell$, assuming (3.16) and (3.17). Left: $\zeta = 0.5$, right: $\zeta = 0.9$. Solid lines: rank bound (3.24) using the minimal $m_{(1.5)}(\varepsilon)$ (solid lines in Fig. A.1). Dotted lines: rank bound for a modified iteration (3.26) with target accuracy $N\tilde{f} = g_{\tilde{m}}$ according to (3.28). Cross and plus markers: standard iteration using the same $N\tilde{f}$ according to (3.12) for $\xi = \zeta/2$ (cross) and $\xi = \zeta$ (plus).

We omit the formulas for rank estimates in terms of target accuracy ε. Numerical values are provided in Figures 3.1 and 3.2. They indicate that using the modified iteration until some $g_{\tilde{m}} = N\tilde{f}$ as derived above can outperform the standard iteration with fixed $N\tilde{f}$.

3.5 Numerical illustration of error bounds

In the numerical illustrations in Figures 3.1 and 3.2 we compare the derived rank estimates for achieving an ε-solution with the modified iteration in the two scenarios of an exponential growth $\nu_m \leq 2^m - 1$, i.e. $\mu_1 = 2$ in (3.3), and for a linear rank growth $\nu_m \leq m$.

Both scenarios are evaluated for the values $\zeta = 0.5$ and $\zeta = 0.9$ (spectral norm of M). We consider the approximation rate (2.7) for Nf with $C = 1$ in the two cases $\xi = \zeta/2$ and $\xi = \zeta$ in (3.16), where we assume $m_0 = 1$ in (3.20), that is, g_m is obtained from g_{m-1} by a rank-one update. To check the potential merit of a larger batch size in the case $\xi = \zeta$, we also consider $m_0 = 2$ (rank-two updates) with squared rate $\xi = \zeta^2$ (i.e. $t = 2$ in (3.21)). According to the plots, the larger batch size can be slightly beneficial for the case of exponential rank growth, but does not help in the case of linear rank growth. The following functions are shown in Figures 3.1 and 3.2:

– The rank bounds (3.24) (in Fig. 3.1) and (3.25) (in Fig. 3.2) for the modified iteration as solid lines, when using as m the minimal number of steps $m_{(1.5)}(\varepsilon)$ such that the right-hand side of (2.8) is less than ε. The values for $m_{(1.5)}(\varepsilon)$ are determined numerically and are depicted in Fig. A.1 in the appendix (as solid lines). One could use instead the derived upper bounds (2.10)–(2.11) for $m_{(1.5)}(\varepsilon)$ (these can be seen in in Fig. A.1 as dotted lines). One then obtains slightly worse rank bounds, especially in the case $\xi = \zeta$.

– The rank bounds for a modified iteration (3.26) as dotted lines, using some truncation $N\tilde{f} = g_{\tilde{m}}$ as a final approximation, according to (3.28) (Fig. 3.1) and (3.29) (Fig. 3.2). We used $\tilde{C} = (1 - \xi^{2\tilde{m}})^{-1/2}$ in (3.27), as motivated above.

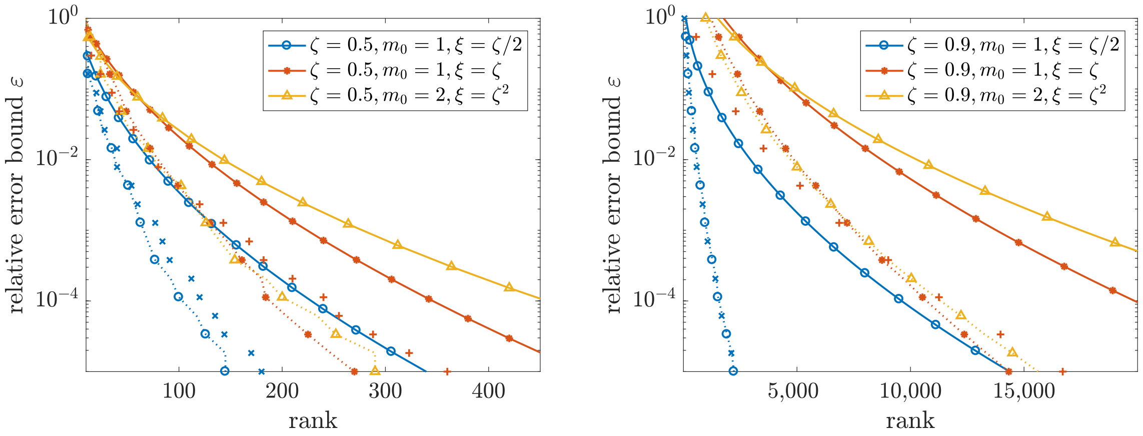

Figure 3.2: Rank bounds for ε-solutions with the modified iteration and linear growth $\mu_\ell \leq \ell + 1$. Left: $\zeta = 0.5$, right: $\zeta = 0.9$. Solid lines: rank bound (3.25) using the minimal $m_{(1.5)}(\varepsilon)$ (solid lines in Fig. A.1). Dotted lines: rank bound for a modified iteration (3.26) with target accuracy $N\tilde{f} = g_{\tilde{m}}$ according to (3.29). Cross and plus markers: standard iteration using the same $N\tilde{f}$ according to (3.13) for $\xi = \zeta/2$ (cross) and $\xi = \zeta$ (plus).

- The rank bound for the standard iteration when using the same truncation $N\tilde{f} = g_{\tilde{m}}$, according to (3.12) (Fig. 3.1) and (3.13) (Fig. 3.2), respectively. These are only given for batch size $m_0 = 1$ and are depicted as cross markers for $\xi = \zeta/2$ and plus markers for $\xi = \zeta$.

As can be seen from both figures, modified iterations can perform equally well or better than the standard iteration with truncated right-hand side. Especially for the case of linear rank growth, the modified iterations with a target accuracy for Nf (dotted lines) seem to provide a reasonable improvement for $\xi = \zeta$, in particular keeping in mind that they are more data-sparse. It appears that with exponential rank growth the modified iteration should not be terminated at a fixed $g_{\tilde{m}} = N\tilde{f}$, and that it helps to take a larger batch size to ensure fewer steps. However, the rank bounds, especially for $\zeta = 0.9$, (Figure 3.1, right) are ridiculously large and only of theoretical interest. This illustrates that if for a given linear equation (1.1) there does not exist an iteration that is either fast or not exponentially rank increasing, its solution might not admit a good low-rank approximation.

4 Numerical experiment

Finally, we include a small numerical experiment to compare the actual convergence of the methods for a particular problem. We generate a 1000×1000 tridiagonal matrix $A = L + D + R$, with diagonal entries in D uniformly distributed in the interval $[2, 3]$, whereas the lower and upper off diagonal entries in L and R are uniformly distributed in $[-2, -1]$ and $[-1, 0]$, respectively. The goal is to solve the linear equation $Au = f$, where f has exponentially decaying entries $f_i = (4/5)^i$. Since the exact solution can be well approximated by sparse vectors, we aim at iterations that build approximations with possibly few nonzero entries. Hence the rank function here is the number of nonzero entries in a vector.

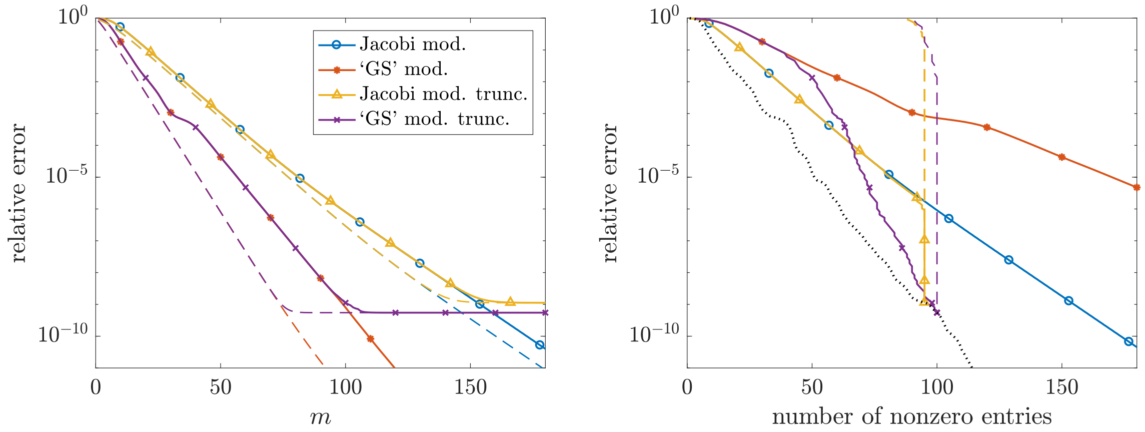

Figure 4.3: Numerical results for the solution of a tridiagonal linear system by the Jacobi method or an approximate Gauss-Seidel method ('GS'). Dashed lines: standard iterations (with and without truncation). Solid lines: modified iterations (with and without truncation). The dotted black line on the right shows the best possible relative error for a given number of nonzero entries.

We employ two iterative solvers (1.2), the Jacobi method, where $N = D^{-1}$, and an approximate Gauss-Seidel method, with

$$N = D^{-1}(I - LD^{-1} + (LD^{-1})^2),$$

which is an approximation of $(L + D)^{-1} = D^{-1}(I + LD^{-1})^{-1}$. Correspondingly,

$$M = I - NA$$

is a tridiagonal banded matrix for the Jacobi method (with zero diagonal), and a five-banded matrix (with only one upper diagonal) for the approximate Gauss-Seidel method. We use the standard iterations (1.2) and modified iterations (1.5). As approximations $g_m \to Nf$ we take $g_m = Nf_m$, where f_m contains the largest m entries of f in modulus. Compared to taking the m largest entries of Nf this has the advantage that the g_m can be recursively computed from the sparse columns of N without forming Nf. (Almost no difference was observed for the two approaches.) All four resulting methods are also tested in a truncated version, where after every step entries smaller than a fixed threshold are deleted from the iterate.

The results for various instances of A varied slightly but were overall quite similar. In Figure 4.3 we show one of the better outcomes. The left plot shows the decrease of the relative error (in Euclidean norm) to the exact solution u, with the dashed lines corresponding to the standard iteration and solid lines to the modified iteration. The numerically computed spectral radii and spectral norms in this instance were $\rho(M) \approx 0.934$ and $\zeta = \|M\| \approx 1.145$ for the Jacobi method, and $\rho(M) \approx 0.885$ and $\zeta = \|M\| \approx 0.981$ for the approximate Gauss-Seidel method. The threshold in the truncated versions was 10^{-9} to reach a relative accuracy below 10^{-8}.

In the right panel of Figure 4.3 we investigate the convergence speed with respect to the number of used nonzero entries. For the standard iterations only the truncated versions are shown (the vertical dashed lines), since without truncation the iterates immediately fill up. While all truncated methods eventually need about the same number of nonzero entries for a relative error in the magnitude of the threshold, the modified iterations need less nonzeros during the overall process. One should also recall that the standard methods operate with a full vector Nf throughout. Note that the Jacobi method without truncation (blue line), while being slower, is capable of constructing

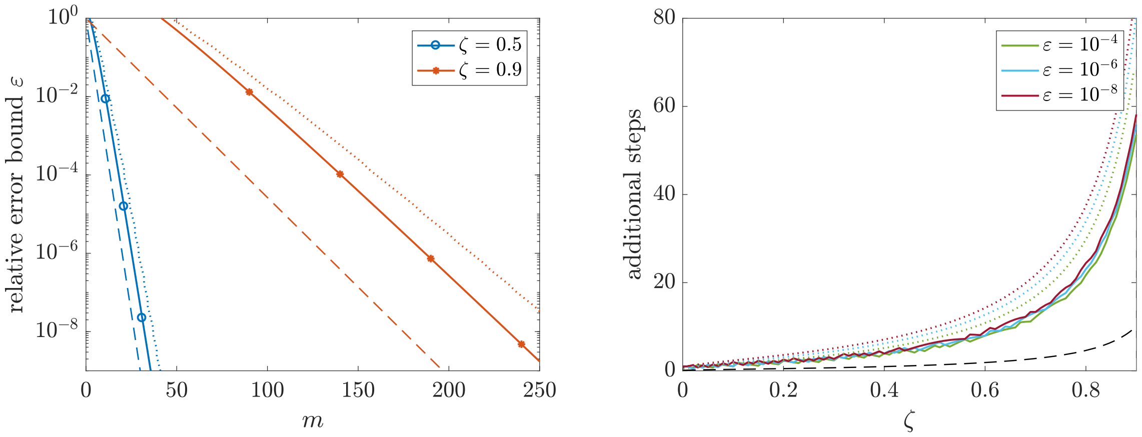

Figure A.1: Illustration of Lemma A.1. Left: The function $m \mapsto \zeta^m(1 + mC(1 + \zeta))$ for $C = 1$ and two different values of ζ as solid lines. Dotted lines: upper bounds (A.1) as a function of ε. Dashed lines: standard rates $\frac{\ln\varepsilon}{\ln\zeta}$. Right: Estimated number of additional steps in the modified iteration for $\xi = \zeta$ ($C = 1$) and different choices of ε compared to the standard iteration, i.e. $m(\varepsilon) - \ln\varepsilon/\ln\zeta$. Solid lines: using the smallest $m(\varepsilon)$ such that $\zeta^m(1 + mC(1 + \zeta)) \le \varepsilon$, dotted lines: using the proven bound (A.1) for $m(\varepsilon)$. The black dashed line shows the number of additional steps to reach $\varepsilon = 10^{-8}$ when $\xi = \zeta/2$ (according to the bound (2.10)). (Ceil operations omitted.)

relatively sparse solutions, whereas the approximate Gauß-Seidel method (red line) clearly requires the truncation. For comparison, the right panel also displays the best possible (relative) sparse approximation errors (i.e. the decay of $\tau_r(u)$) as a black dotted line, which are obtained from using the largest entries (in modulus) of the true solution u. The truncated approximate Gauss-Seidel method gets closest to this minimal error before reaching the number of non-zeros required for the accuracy specified by the threshold. Note that for both the truncated Jacobi and approximate Gauss-Seidel method the final error is essentially optimal with respect to the sparsity.

A Appendix

We provide a proof for (2.11). By (2.8), the statement we need to show is the following.

Lemma A.1. *For $0 < \varepsilon < 1$ the minimal integer $m(\varepsilon) \ge 1$ that satisfies*

$$\zeta^m(1 + mC(1 + \zeta)) \le \varepsilon$$

can be bounded by

$$m(\varepsilon) \le \left\lceil \frac{\ln\varepsilon}{\ln\zeta} + K_2(\zeta, C) + \sqrt{\frac{2}{\ln\zeta^{-1}}}\sqrt{\frac{\ln\varepsilon}{\ln\zeta} + K_2(\zeta, C) + \frac{1}{\ln\zeta} + \frac{1}{C(1+\zeta)}} \right\rceil, \qquad \text{(A.1)}$$

where

$$K_2(\zeta, C) = \frac{\ln\ln(\zeta^{-1/(C(1+\zeta))})}{\ln\zeta} = \frac{\left|\ln\ln\zeta^{-1}\right| + \ln(C(1+\zeta))}{|\ln\zeta|}.$$

In Figure A.1 the bound (A.1) is compared numerically to the true value of $m(\varepsilon)$.

Proof of Lemma A.1. Instead of the integer $m(\varepsilon)$ we consider the minimal real number $x = x(\varepsilon) \geq 1$ that satisfies $\zeta^x + x\zeta^x C(1+\zeta) \leq \varepsilon$. Denoting $a = \zeta^{1/C(1+\zeta)}$ this is equivalent with

$$a^{1+xC(1+\zeta)}(1 + xC(1+\zeta)) \leq \varepsilon a,$$

or, with $x' = 1 + xC(1+\zeta)$,

$$a^{x'} x' \leq \varepsilon a.$$

We rewrite this as

$$x' \ln a \cdot e^{x' \ln a} \geq \varepsilon a \ln a. \tag{A.2}$$

Note for the right-hand side that $0 > \varepsilon a \ln a \geq -\varepsilon e^{-1}$. We denote the inverse relation of $ze^z = y$ for $-e^{-1} \leq y \leq 0$ and $z \leq -1$ by $z = W_{-1}(y)$. It is called the W_{-1} branch of the *Lambert W function*. Since W_{-1} is monotonically decreasing, (A.2) will be satisfied if

$$x' \ln a \leq W_{-1}\left(\varepsilon a \ln a\right) \quad \Leftrightarrow \quad x' \geq \frac{1}{\ln a} W_{-1}\left(\varepsilon a \ln a\right). \tag{A.3}$$

Writing

$$\varepsilon a \ln a = -e^{-1-b} \quad \Leftrightarrow \quad b = -\ln\left(e\varepsilon a \ln a^{-1}\right) = -\ln \varepsilon - \ln a - \ln\ln a^{-1} - 1$$

the following bound is known [9]:

$$W_{-1}(-e^{-b-1}) \geq -1 - \sqrt{2b} - b$$

(with strict inequality when $b > 0$). The condition

$$x' \geq \frac{1}{\ln a^{-1}}\left(b + \sqrt{2b} + 1\right)$$

is therefore stronger than (A.3) and hence also sufficient for (A.2). Using the definition of x' and a we rewrite this as

$$x \geq \frac{1}{C(1+\zeta)\ln a^{-1}}\left(b + \sqrt{2b} + 1\right) - \frac{1}{C(1+\zeta)} = \frac{b + \sqrt{2b}}{\ln \zeta^{-1}} - \frac{1}{\ln \zeta} - \frac{1}{C(1+\zeta)}.$$

Now noting that

$$\frac{b}{\ln \zeta^{-1}} = \frac{\ln \varepsilon}{\ln \zeta} + \frac{1}{C(1+\zeta)} + \frac{\ln\ln a^{-1}}{\ln \zeta} + \frac{1}{\ln \zeta}$$

and setting $K_2(\zeta, C) = \frac{\ln\ln a^{-1}}{\ln \zeta}$ proves the assertion. □

Acknowledgement

We thank Daniel Kressner for helpful discussions on parts of this work and for providing additional references.

References

[1] M. Bachmayr and W. Dahmen. Adaptive near-optimal rank tensor approximation for high-dimensional operator equations. *Found. Comput. Math.*, 15(4):839–898, 2015.

[2] M. Bachmayr and R. Schneider. Iterative methods based on soft thresholding of hierarchical tensors. *Found. Comput. Math.*, 17(4):1037–1083, 2017.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[4] M. Benzi and G. H. Golub. Bounds for the entries of matrix functions with applications to preconditioning. *BIT*, 39(3):417–438, 1999.

[5] G. Beylkin and M. J. Mohlenkamp. Algorithms for numerical analysis in high dimensions. *SIAM J. Sci. Comput.*, 26(6):2133–2159, 2005.

[6] M. Billaud-Friess, A. Nouy, and O. Zahm. A tensor approximation method based on ideal minimal residual formulations for the solution of high-dimensional problems. *ESAIM Math. Model. Numer. Anal.*, 48(6):1777–1806, 2014.

[7] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, 27(3):265–274, 2009.

[8] K. Bredies and D. A. Lorenz. Linear convergence of iterative soft-thresholding. *J. Fourier Anal. Appl.*, 14(5-6):813–837, 2008.

[9] I. Chatzigeorgiou. Bounds on the Lambert function and their application to the outage analysis of user cooperation. *IEEE Commun. Lett.*, 17(8):1505–1508, 2013.

[10] W. Dahmen, R. DeVore, L. Grasedyck, and E. Süli. Tensor-sparsity of solutions to high-dimensional elliptic partial differential equations. *Found. Comput. Math.*, 16(4):813–874, 2016.

[11] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[12] S. Demko, W. F. Moss, and P. W. Smith. Decay rates for inverses of band matrices. *Math. Comp.*, 43(168):491–499, 1984.

[13] M. Espig, W. Hackbusch, T. Rohwedder, and R. Schneider. Variational calculus with sums of elementary tensors of fixed rank. *Numer. Math.*, 122(3):469–488, 2012.

[14] L. Grasedyck. Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure. *Computing*, 72(3-4):247–265, 2004.

[15] L. Grubišić and D. Kressner. On the eigenvalue decay of solutions to operator Lyapunov equations. *Systems Control Lett.*, 73:42–47, 2014.

[16] W. Hackbusch. Solution of linear systems in high spatial dimensions. *Comput. Vis. Sci.*, 17(3):111–118, 2015.

[17] W. Hackbusch. *Tensor spaces and numerical tensor calculus.* Springer, Cham, second edition, 2019.

[18] W. Hackbusch, B. N. Khoromskij, and E. E. Tyrtyshnikov. Approximate iterations for structured matrices. *Numer. Math.*, 109(3):365–383, 2008.

[19] B. N. Khoromskij. *Tensor numerical methods in scientific computing.* De Gruyter, Berlin, 2018.

[20] D. Kressner and C. Tobler. Krylov subspace methods for linear systems with tensor product structure. *SIAM J. Matrix Anal. Appl.*, 31(4):1688–1714, 2009/10.

[21] D. Kressner and A. Uschmajew. On low-rank approximability of solutions to high-dimensional operator equations and eigenvalue problems. *Linear Algebra Appl.*, 493:556–572, 2016.

[22] T. Penzl. Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case. *Systems Control Lett.*, 40(2):139–144, 2000.

[23] H. Rauhut, R. Schneider, and Ž. Stojanac. Tensor completion in hierarchical tensor representations. In *Compressed sensing and its applications*, pages 419–450. Birkhäuser/Springer, Cham, 2015.

[24] H. Rauhut, R. Schneider, and Ž. Stojanac. Low rank tensor recovery via iterative hard thresholding. *Linear Algebra Appl.*, 523:220–262, 2017.

[25] R. A. Smith. Matrix equation $XA + BX = C$. *SIAM J. Appl. Math.*, 16:198–201, 1968.

[26] A. Townsend and H. Wilber. On the singular values of matrices with high displacement rank. *Linear Algebra Appl.*, 548:19–41, 2018.

[27] E. Zeidler. *Nonlinear functional analysis and its applications. I.* Springer-Verlag, New York, 1986.