

# Stabilization of Stochastic Iterative Methods for Singular and Nearly Singular Linear Systems

Mengdi Wang  
mdwang@MIT.EDU

Dimitri P. Bertsekas\*  
dimitrib@mit.edu

## Abstract

We consider linear systems of equations,  $Ax = b$ , with an emphasis on the case where  $A$  is singular. Under certain conditions, necessary as well as sufficient, linear deterministic iterative methods generate sequences  $\{x_k\}$  that converge to a solution, as long as there exists at least one solution. This convergence property can be impaired when these methods are implemented with stochastic simulation, as is often done in important classes of large-scale problems. We introduce additional conditions and novel algorithmic stabilization schemes under which  $\{x_k\}$  converges to a solution when  $A$  is singular, and may also be used with substantial benefit when  $A$  is nearly singular.

**Key words:** stochastic algorithm, singular system, stabilization, projected equation, simulation, regularization, approximate dynamic programming.

## 1 Introduction

We consider the solution of the linear system of equations

$$Ax = b,$$

where  $A$  is an  $n \times n$  real matrix and  $b$  is a vector in  $\mathbb{R}^n$ , by using approximations of  $A$  and  $b$ , generated by simulation or other stochastic process. We allow  $A$  to be nonsymmetric and singular, but *we assume throughout that the system is consistent*, i.e., it has at least one solution.

We are interested in methods where in place of  $A$  and  $b$ , we use simulation-generated approximations  $A_k$  and  $b_k$  with  $A_k \rightarrow A$  and  $b_k \rightarrow b$ . In the case where  $A$  is nonsingular, a possible approach is to approximate the solution  $A^{-1}b$  with  $A_k^{-1}b_k$ , since the inverse  $A_k^{-1}$  will exist for sufficiently large  $k$ . In the case where  $A$  is singular, it may seem possible to adopt a pseudoinverse approach, whereby we may attempt to approximate a solution of  $Ax = b$  with a pseudoinverse solution  $A_k^\dagger b_k$ . Unfortunately, this approach does not work (for example, consider the case where  $A = 0$  and  $b = 0$ , so that  $A_k$  and  $b_k$  are equal to simulation noises, in which case  $A_k^\dagger b_k$  is unlikely to converge as  $k \rightarrow \infty$ ). This motivates the use of an iterative method of the form

$$x_{k+1} = x_k - \gamma G_k (A_k x_k - b_k), \tag{1}$$

---

\*Mengdi Wang and Dimitri Bertsekas are with the Department of Electrical Engineering and Computer Science, and the Laboratory for Information and Decision Systems (LIDS), M.I.T. Work supported by the Air Force Grant FA9550-10-1-0412 and by the LANL Information of Science and Technology Institute.

where  $\gamma$  is a positive stepsize, and  $G_k$  is an  $n \times n$  scaling matrix. This method is motivated by the classical iteration

$$x_{k+1} = x_k - \gamma G(Ax_k - b), \quad (2)$$

which has been considered for singular  $A$  by several authors (see e.g., the survey by Dax [Dax90], and the references cited there, such as Keller [Kel65], Young [You72], Tanabe [Tan74]), who have given necessary and sufficient conditions for convergence of  $\{x_k\}$  to a solution of the system  $Ax = b$ , assuming at least one solution exists. Similar conditions for convergence arise in the context of stability analysis of discrete-time linear dynamic systems.

Monte Carlo simulation has been proposed for solution of linear systems long ago, starting with a suggestion by von Neumann and Ulam, as recounted by Forsythe and Leibler [FoL50], and Wasow [Was52] (see also Curtiss [Cur54], [Cur57], and the survey by Halton [Hal70]). More recently, work on simulation methods has focused on using low-order calculations for solving large least squares and other problems. In this connection we note the papers by Strohmer and Vershynin [StV09], Censor and Herman [CeH09], and Leventhal and Lewis [LeL10] on randomized versions of coordinate descent and iterated projection methods for overdetermined least squares problems, and the series of papers by Drineas et. al. who consider the use of simulation methods for linear least squares problems and low-rank matrix approximation problems; see [DKM06a], [DKM06b], [DMM06], [DMM08], and [DMMS11].

Our motivation in this paper can be traced to the methodology of approximate dynamic programming (ADP for short), also known as reinforcement learning. The aim there is to solve forms of Bellman's equation of very large dimension (billions and much larger) by using simulation (see the books by Bertsekas and Tsitsiklis [BeT96], and by Sutton and Barto [SuB98]). In this context,  $Ax = b$  is a low-dimensional system obtained from a high-dimensional system through a Galerkin approximation or aggregation process. The formation of the entries of  $A$  and  $b$  often requires high-dimensional linear algebra operations that are impossible to perform without the use of simulation. In another related context, the entries of  $A$  and  $b$  may not be known explicitly, but only through a simulator, in which case using approximations  $(A_k, b_k)$  instead of  $(A, b)$  is the most we can expect.

To our knowledge, all of the existing works on simulation-based methods assume that the system  $Ax = b$  has a unique solution (an exception is [Ber11], which considers, among others, simulation-based algorithms where  $A$  is singular and has a special form that arises in ADP). By contrast, in this paper we address the convergence issue in the singular case. Our line of development is to first analyze the convergence of the deterministic iteration (2). Under a certain assumption that is necessary and sufficient for convergence, we decompose the error  $x_k - x^*$ , where  $x^*$  is a solution of  $Ax = b$ , into the sum of two orthogonal components:

$Uy_k$ : a component in  $\mathbf{N}(A)$ , the nullspace of  $A$  [ $U$  is a basis matrix for  $\mathbf{N}(A)$ ],

$Vz_k$ : a component in  $\mathbf{N}(A)^\perp$  [ $V$  is a basis matrix for  $\mathbf{N}(A)^\perp$ ].

The iterate  $z_k$  is uncoupled from  $y_k$ , and converges to 0 at a geometric rate, as it evolves according to a contractive iteration. As a result, the cumulative effect of  $y_k$  is to add a constant vector  $Uy_0$  from  $\mathbf{N}(A)$  to the component  $Vz_k$ , so  $\{x_k\}$  converges to some vector in  $x^* + \mathbf{N}(A)$ , the solution set of  $Ax = b$ .

The deterministic convergence analysis is the starting point for the analysis of stochastic variants, where  $A$ ,  $b$ , and  $G$  are replaced by convergent estimates  $A_k$ ,  $b_k$ , and  $G_k$ , as in iteration (1). Then, when  $A$  is singular, it turns out that there is a major qualitative change in the nature of the convergence process. The reason is that the iterates  $y_k$  and  $z_k$  become coupled in complicated ways, and the effects of the coupling may also depend on the rate of the convergence of  $A_k$ ,  $b_k$ , and  $G_k$ . As a result,  $\{x_k\}$  and even the residual sequence  $\{Ax_k - b\}$  need not converge, under natural simulation-based implementations. To guarantee satisfactory behavior, we impose additional assumptions and introduce stabilization schemes in iteration (1), which are novel and broadly applicable. With these schemes, we prove that  $\{x_k\}$  converges to a single special solution, which is independent of the initial condition  $x_0$ . The limit depends only on the algorithm and the type of stabilization used, and in important special cases is related to the Drazin inverse solution or the Moore-Penrose pseudoinverse solution of  $GAx = Gb$ . This is in sharp contrast with the deterministic iteration (2), whose limit strongly depends on  $x_0$  when  $A$  is singular.

While our analysis focuses explicitly on the case where  $A$  is singular, much of our algorithmic methodology applies to the important case where  $A$  is nearly singular. Indeed, when simulation is used, there is little difference from a practical point of view, between the cases where  $A$  is singular and  $A$  is nonsingular but highly ill-conditioned. This is particularly so in the early part of the iterative computation, when the standard deviation of the simulation noise  $(A_k - A, G_k - G, b_k - b)$  is overwhelming relative to the smallest singular value of  $A$ .

The paper is organized as follows. In Section 2 we derive a necessary and sufficient condition for the convergence of the deterministic iterative method (2) for singular linear systems. While this condition is essentially a restatement of known results (see e.g., [Dax90] and the references cited there), our line of development is different from those found in the literature (it has a strong linear algebra flavor), and paves the way for analysis of related stochastic methods. In Section 3 we discuss a common simulation framework and illustrate a few practical cases of simulation-based solution of large linear systems, and we discuss the convergence issues of the stochastic iterative method (1). In Section 4, we introduce our new stabilization schemes, we analyze their convergence under various conditions, and we discuss various instances of alternative stabilization schemes that are well suited to the special structures of particular problems and algorithms. In Section 5 we suggest that our stabilization schemes can substantially improve the performance of stochastic iterative methods in the case where  $A$  is nonsingular but highly ill-conditioned. Finally, in Section 6 we present computational experiments that support our analysis.

We summarize our terminology, our notation, and some basic facts regarding positive semidefinite matrices. In our analysis all calculations are done with real finite-dimensional vectors and matrices. A vector  $x$  is viewed as a column vector, while  $x'$  denotes the corresponding row vector. For a matrix  $A$ , we use  $A'$  to denote its transpose. The standard Euclidean norm of a vector  $x$  is  $\|x\| = \sqrt{x'x}$ .

The nullspace and range of a matrix  $A$  are denoted by  $\mathbf{N}(A)$  and  $\mathbf{R}(A)$ , respectively. For two square matrices  $A$  and  $B$ , the notation  $A \sim B$  indicates that  $A$  and  $B$  are related by a similarity transformation and therefore have the same eigenvalues. When we wish to indicate the similarity transformation  $P$ , we write  $A \stackrel{P}{\sim} B$ , meaning that  $A = PBP^{-1}$ . The spectral radius of  $A$  is denoted by  $\rho(A)$ . We denote by  $\|A\|$  the Euclidean matrix norm of a matrix  $A$ , so that  $\|A\|$  is the square root of the largest eigenvalue of  $A'A$ . We have  $\rho(A) \leq \|A\|$ , and we will use the fact that if  $\rho(A) < 1$ , there exists a weighted Euclidean norm  $\|\cdot\|_P$ , defined for an invertible matrix  $P$  as  $\|x\|_P = \|P^{-1}x\|$  for all  $x \in \mathfrak{R}^n$ , such that the corresponding induced matrix norm  $\|A\|_P = \max_{\|x\|_P=1} \|Ax\|_P = \|P^{-1}AP\|$  satisfies  $\|A\|_P < 1$  (see Ortega and Rheinboldt [OrR70], Th. 2.2.8 and its proof, or Stewart [Ste73], Th. 3.8 and its proof).

If  $A$  and  $B$  are square matrices, we write  $A \preceq B$  or  $B \succeq A$  to denote that the matrix  $B - A$  is positive semidefinite, i.e.,  $x'(B - A)x \geq 0$  for all  $x$ . Similarly, we write  $A \prec B$  or  $B \succ A$  to denote that the matrix  $B - A$  is positive definite, i.e.,  $x'(B - A)x > 0$  for all  $x \neq 0$ . We have  $A \succ 0$  if and only if  $A \succ cI$  for some positive scalar  $c$  [take  $c$  in the interval  $(0, \min_{\|x\|=1} x'Ax)$ ].

If  $A \succ 0$ , the eigenvalues of  $A$  have positive real parts (see Theorem 3.3.9, and Note 3.13.6 of Cottle, Pang, and Stone [CPS92]). Similarly, if  $A \succeq 0$ , the eigenvalues of  $A$  have nonnegative real parts (since if  $A$  had an eigenvalue with negative real part, then for sufficiently small  $\delta > 0$ , the same would be true for the positive definite matrix  $A + \delta I$  - a contradiction). For a singular matrix  $A$ , the algebraic multiplicity of the 0 eigenvalue is the number of 0 eigenvalues of  $A$ . This number is greater or equal to the dimension of  $\mathbf{N}(A)$  (the geometric multiplicity of the 0 eigenvalue, i.e., the dimension of the eigenspace corresponding to 0). We will use the fact that in case of strict inequality, there exists a vector  $v$  such that  $Av \neq 0$  and  $A^2v = 0$ ; this is a generalized eigenvector of order 2 corresponding to eigenvalue 0 (see [LaT85], Section 6.3).

The abbreviations “ $\xrightarrow{a.s.}$ ” and “ $\xrightarrow{i.d.}$ ” mean “converges almost surely to,” and “converges in distribution to,” respectively, while the abbreviation “i.i.d.” means “independent identically distributed.” For two sequences  $\{x_k\}$  and  $\{y_k\}$ , we use the abbreviation  $x_k = O(y_k)$  to denote that, there exists  $c > 0$  such that  $\|x_k\| \leq c\|y_k\|$  for all  $k$ . Moreover, we use the abbreviation  $x_k = \Theta(y_k)$  to denote that, there exists  $c_1, c_2 > 0$  such that  $c_1\|y_k\| \leq \|x_k\| \leq c_2\|y_k\|$  for all  $k$ .

## 2 Deterministic Iterative Methods for Singular Linear Systems

In this section, we analyze the convergence of the deterministic iteration

$$x_{k+1} = x_k - \gamma G(Ax_k - b). \quad (3)$$

For a given triplet  $(A, b, G)$ , with  $b \in \mathbf{R}(A)$ , we say that this iteration is *convergent* if there exists  $\bar{\gamma} > 0$  such that for all  $\gamma \in (0, \bar{\gamma}]$  and all initial conditions  $x_0 \in \mathfrak{R}^n$ , the sequence  $\{x_k\}$  produced by the iteration converges to a solution of  $Ax = b$ . *Throughout our subsequent analysis we assume that  $A$  is singular.* This is done for convenience, since a major part of our analytical technique (e.g., the nullspace decomposition of the subsequent Prop. 2) makes no sense if  $A$  is nonsingular, and it would be awkward to modify so that it applies to both the singular and the nonsingular cases. However, our algorithms and results have evident (and simpler) counterparts for the case where  $A$  is nonsingular. Indeed, a major motivation for our analysis is the case where  $A$  is nonsingular but is instead nearly singular, so that methods for solving  $Ax = b$  are highly susceptible to simulation noise.

The following condition, a restatement of conditions given in various contexts in the literature (e.g., [Kel65], [You72], [Tan74], [Dax90]), turns out to be equivalent to the iteration being convergent.

### Assumption 1

- (a) Each eigenvalue of  $GA$  either has a positive real part or is 0.
- (b) The dimension of  $\mathbf{N}(GA)$  is equal to the algebraic multiplicity of the eigenvalue 0 of  $GA$ .
- (c)  $\mathbf{N}(A) = \mathbf{N}(GA)$ .

We first show that the conditions above are necessary for convergence, which is relatively easy, and then prove that they are also sufficient for convergence, using a special decomposition that will be useful in our subsequent analysis.

**Proposition 1** *If the iteration (3) is convergent, the conditions of Assumption 1 must hold.*

*Proof.* If part (a) of Assumption 1 does not hold, some eigenvalue of  $I - \gamma GA$  will be strictly outside the unit circle for all  $\gamma > 0$ , so the iteration (3) cannot be convergent. If part (b) does not hold, there exists a vector  $w$  such that  $GAw \neq 0$  but  $(GA)^2w = 0$ ; this is a generalized eigenvector of order 2. Assuming that  $\gamma > 0$ ,  $b = 0$  and  $x_0 = w$ , we then have

$$x_k = (I - \gamma GA)^k x_0 = (I - k\gamma GA)x_0, \quad GAx_k = GAx_0 \neq 0, \quad k = 1, 2, \dots,$$

so the iterate sequence  $\{x_k\}$  diverges for all  $\gamma > 0$ . Finally, if part (c) does not hold,  $\mathbf{N}(A)$  is strictly contained in  $\mathbf{N}(GA)$ , so for any solution  $x^*$  of  $Ax = b$ , the iteration (3) when started at any  $x_0 \in x^* + \mathbf{N}(GA)$  with  $x_0 \notin x^* + \mathbf{N}(A)$ , will stop at  $x_0$ , which is not a solution of  $Ax = b$  - a contradiction. ■

To show that the iteration (3) is convergent under Assumption 1, we first derive a decomposition of  $GA$ , which will also be useful in the analysis for the stochastic iterations of Section 4.

**Proposition 2 (Nullspace Decomposition)** *Let Assumption 1 hold. The matrix  $GA$  can be written as*

$$GA = [U \ V] \begin{bmatrix} 0 & N \\ 0 & H \end{bmatrix} [U \ V]', \quad (4)$$

where

$U$  is a matrix whose columns form an orthonormal basis of  $\mathbf{N}(A)$ .

$V$  is a matrix whose columns form an orthonormal basis of  $\mathbf{N}(A)^\perp$ .

$N = U'GAV$ .

$H$  is the square matrix given by

$$H = V'GAV, \quad (5)$$

and its eigenvalues are equal to the eigenvalues of  $GA$  that have positive real parts.

*Proof.* Let  $U$  be a matrix whose columns form an orthonormal basis of  $\mathbf{N}(GA)$ , and let  $V$  be a matrix whose columns form an orthonormal basis of  $\mathbf{N}(GA)^\perp$ . We have

$$[U \ V]'GA[U \ V] = \begin{bmatrix} U'GAU & U'GAV \\ V'GAU & V'GAV \end{bmatrix} = \begin{bmatrix} 0 & U'GAV \\ 0 & V'GAV \end{bmatrix} = \begin{bmatrix} 0 & N \\ 0 & H \end{bmatrix}, \quad (6)$$

where we used the fact  $GAU = 0$ , so that  $U'GAU = V'GAU = 0$ . Clearly  $[U \ V]$  is orthogonal, since  $[U \ V][U \ V]' = UU' + VV' = I$  and  $[U \ V]'[U \ V] = \begin{bmatrix} U'U & 0 \\ 0 & V'V \end{bmatrix} = I$ . Therefore Eq. (4) follows from Eq. (6).

From Eq. (6), the eigenvalues of  $GA$  are the eigenvalues of  $H$  plus 0 eigenvalues whose number is equal to the dimension of  $\mathbf{N}(GA)$ . Thus by Assumption 1(a), the eigenvalues of  $H$  are either 0 or have positive real parts. If  $H$  had 0 as an eigenvalue, the algebraic multiplicity of the 0 eigenvalue of  $GA$  would be strictly greater than the dimension of  $\mathbf{N}(GA)$ , a contradiction of Assumption 1(b). Hence, all eigenvalues of  $H$  have positive real parts.  $\blacksquare$

The significance of the decomposition (4) is that in a scaled coordinate system defined using  $y = U'(x - x^*)$  and  $z = V'(x - x^*)$ , where  $x^*$  is a solution of  $Ax = b$ , the iteration (3) decomposes into two component iterations, one for  $y$ , generating a sequence  $\{y_k\}$ , and one for  $z$ , generating a sequence  $\{z_k\}$ , which is independent of the sequence  $\{y_k\}$ . This is formalized in the following proposition.

**Proposition 3 (Iteration Decomposition)** *Let Assumption 1 hold. If  $x^*$  is a solution of the system  $Ax = b$ , the iteration (3) can be written as*

$$x_k = x^* + Uy_k + Vz_k, \quad (7)$$

where  $y_k$  and  $z_k$  are given by

$$y_k = U'(x_k - x^*), \quad z_k = V'(x_k - x^*), \quad (8)$$

and are generated by the iterations

$$y_{k+1} = y_k - \gamma Nz_k, \quad z_{k+1} = z_k - \gamma Hz_k, \quad (9)$$

and  $U$ ,  $V$ ,  $N$ , and  $H$  are the matrices of Prop. 2. Moreover the corresponding residuals are given by

$$r_k = Ax_k - b = AVz_k. \quad (10)$$

*Proof.* The iteration (3) is written as  $x_{k+1} - x^* = (I - \gamma GA)(x_k - x^*)$ , which in view of Prop. 2 has the form

$$x_{k+1} - x^* = [U \ V] \begin{bmatrix} I & -\gamma N \\ 0 & I - \gamma H \end{bmatrix} [U \ V]'(x_k - x^*),$$

or since  $[U \ V]$  is orthogonal,

$$[U \ V]'(x_{k+1} - x^*) = \begin{bmatrix} I & -\gamma N \\ 0 & I - \gamma H \end{bmatrix} [U \ V]'(x_k - x^*). \quad (11)$$

Defining  $y_k = U'(x_k - x^*)$ , and  $z_k = V'(x_k - x^*)$ , and using Eq. (11), we obtain Eq. (9). In addition, using the facts  $x_k - x^* = Uy_k + Vz_k$ ,  $Ax^* = b$  and  $AU = 0$ , we have  $r_k = A(x_k - x^*) = A(Uy_k + Vz_k) = AVz_k$ . ■

Based on the preceding proposition, the iteration for  $z_k$  is independent of  $y_k$ , and has the form  $z_{k+1} = z_k - \gamma H z_k$  [cf. Eq. (9)]. It then follows that  $\{z_k\}$  converges to 0 (at a geometric rate) if and only if the matrix  $I - \gamma H$  is contractive, or equivalently if  $H$  has eigenvalues with positive real parts and  $\gamma$  is sufficiently small. If this is so, it can be seen that  $\{y_k\}$  involves a series of geometric powers of  $I - \gamma H$  and hence also converges. This, together with Prop. 1, proves the equivalence of Assumption 1 with the iteration (3) being convergent, as indicated in the following proposition.

**Proposition 4 (Convergence of Deterministic Iterative Methods)** *If Assumption 1 holds, the iteration (3) converges to the following solution of  $Ax = b$ :*

$$\hat{x} = (UU' - UNH^{-1}V')x_0 + (I + UNH^{-1}V')x^*, \quad (12)$$

where  $x_0$  is the initial iterate and  $x^*$  is the solution of  $Ax = b$  that has minimum Euclidean norm.

*Proof.* From Eq. (9),  $z_k$  and  $y_k$  are equal to

$$z_k = (I - \gamma H)^k z_0, \quad y_k = y_0 - \gamma N \sum_{i=0}^{k-1} (I - \gamma H)^i z_0.$$

By Prop. 2(a),  $I - \gamma H$  has eigenvalues within the unit circle for all sufficiently small  $\gamma$ . Therefore  $z_k$  converges to 0 at a geometric rate, and  $y_k$  also converges and its limit point is given by:

$$\lim_{k \rightarrow \infty} y_k = y_0 - NH^{-1}z_0.$$

Hence, by Eq. (7),  $\{x_k\}$  converges to the vector  $x^* + U(y_0 - NH^{-1}z_0)$ . By using the expression  $y_0 = U'(x_0 - x^*)$  and  $z_0 = V'(x_0 - x^*)$  we further have

$$\begin{aligned} \lim_{k \rightarrow \infty} x_k &= x^* + Uy_0 - UNH^{-1}z_0 \\ &= x^* + UU'(x_0 - x^*) - UNH^{-1}V'(x_0 - x^*) \\ &= (UU' - UNH^{-1}V')x_0 + (I + UNH^{-1}V')x^*, \end{aligned}$$

where the last equality uses the fact  $UU'x^* = 0$  [since  $x^*$  is the minimum norm solution, it is orthogonal to  $\mathbf{N}(A)$ , so  $U'x^* = 0$ ]. ■

The limit  $\hat{x}$  of the iteration can be characterized in terms of the Drazin inverse of  $GA$ , which is denoted by  $(GA)^D$  (see the book [CaM91] for its definition and properties). According to a known result ([EMN88] Theorem 1), it takes the form

$$\hat{x} = (I - (GA)(GA)^D)x_0 + (GA)^D Gb. \quad (13)$$

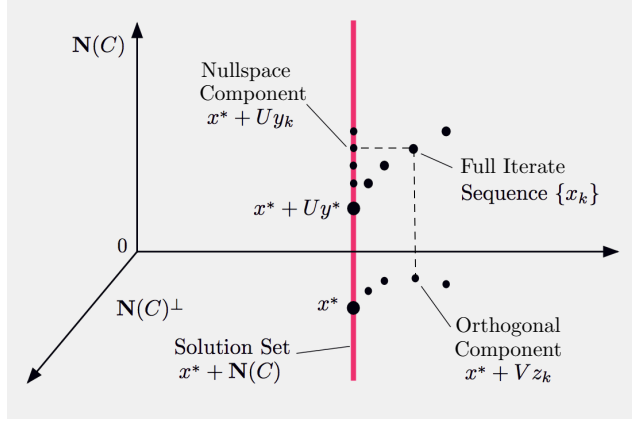


Figure 1: Illustration of the convergence process of the deterministic iteration (3). The iteration decomposes into two orthogonal components, on  $\mathbf{N}(A)$  and  $\mathbf{N}(A)^\perp$ , respectively, and we have  $x_k = x^* + Uy_k + Vz_k$ . In this figure,  $x^*$  is the solution of minimum norm, and  $\{x_k\}$  converges to  $x^* + Uy^*$ , where  $y^*$  is the limit of  $\{y_k\}$ .

Note that  $\hat{x}$  consists of two parts: a linear function of the initial iterate  $x_0$  and the Drazin inverse solution of the linear system  $GAx = Gb$ . Indeed we can verify that the expressions (12) and (13) for  $\hat{x}$  are equal.<sup>1</sup>

The preceding proof of deterministic convergence shows that under Assumption 1, the stepsizes  $\gamma$  that guarantee convergence are precisely the ones for which  $I - \gamma H$  has eigenvalues strictly within the unit circle. Figure 1 illustrates the convergence process.

<sup>1</sup>We use a formula for the Drazin inverse for  $2 \times 2$  block matrices (see [Cve08]), which yields

$$(GA)^D = [U \ V] \begin{bmatrix} 0 & N \\ 0 & H \end{bmatrix}^D [U \ V]' = [U \ V] \begin{bmatrix} 0 & NH^{-2} \\ 0 & H^{-1} \end{bmatrix} [U \ V]'$$

Then we have

$$\begin{aligned} UU' - UNH^{-1}V' &= [U \ V] \begin{bmatrix} I & -NH^{-1} \\ 0 & 0 \end{bmatrix} [U \ V]' \\ &= [U \ V] \left( I - \begin{bmatrix} 0 & N \\ 0 & H \end{bmatrix} \begin{bmatrix} 0 & NH^{-2} \\ 0 & H^{-1} \end{bmatrix} \right) [U \ V]' \\ &= I - (GA)(GA)^D. \end{aligned}$$

The minimum norm solution is given by

$$x^* = \left( [U \ V] \begin{bmatrix} 0 & 0 \\ 0 & H^{-1} \end{bmatrix} [U \ V]' \right) Gb,$$

so by using the equation  $UU' + VV' = I$ , we have

$$\begin{aligned} (I + UNH^{-1}V')x^* &= \left( [U \ V] \begin{bmatrix} I & NH^{-1} \\ 0 & I \end{bmatrix} [U \ V]' \right) \left( [U \ V] \begin{bmatrix} 0 & 0 \\ 0 & H^{-1} \end{bmatrix} [U \ V]' \right) Gb \\ &= [U \ V] \begin{bmatrix} 0 & NH^{-2} \\ 0 & H^{-1} \end{bmatrix} [U \ V]' Gb \\ &= (GA)^D Gb. \end{aligned}$$

By combining the preceding two equations with Eq. (12), we obtain

$$\hat{x} = (UU' - UNH^{-1}V')x_0 + (I + UNH^{-1}V')x^* = (I - (GA)(GA)^D)x_0 + (GA)^D Gb,$$

so the expressions (12) and (13) for  $\hat{x}$  coincide.

## 2.1 Classical Algorithms

We will now discuss some special classes of methods for which Assumption 1 is satisfied. Because this assumption is necessary and sufficient for the convergence of iteration (3) for some  $\gamma > 0$ , any set of conditions under which this convergence has been shown in the literature implies Assumption 1. In what follows we collect various conditions of this type, which correspond to known algorithms of the form (3) or generalizations thereof. These fall in three categories:

- (a) Projection algorithms, which are related to Richardson’s method.
- (b) Proximal algorithms, including quadratic regularization methods.
- (c) Splitting algorithms, including the Gauss-Seidel and related methods.

In their most common form, both projection and proximal methods for the system  $Ax = b$  require that  $A \succeq 0$ , and take the form of Eq. (3) for special choices of  $\gamma$  and  $G$ . Their convergence properties may be inferred from the analysis of their nonlinear versions (originally proposed by Sibony [Sib70], and Martinet [Mar70], respectively). Generally, these methods are used for finding a solution  $x^*$ , within a closed convex set  $X$ , of a variational inequality of the form

$$f(x^*)'(x - x^*) \geq 0, \quad \forall x \in X, \quad (14)$$

where  $f$  is a mapping that is monotone on  $X$ , in the sense that  $(f(y) - f(x))'(y - x) \geq 0$ , for all  $x, y \in X$ .

For the special case where  $f(x) = Ax - b$  and  $X = \mathfrak{R}^n$ , the projection method is obtained when  $G$  is positive definite symmetric, and is related to Richardson’s method (see e.g., [HaY81]). Then strong (or weak) monotonicity of  $f$  is equivalent to positive (or nonnegative, respectively) definiteness of  $A$ . The convergence analysis of the projection method for the variational inequality (14) generally requires strong monotonicity of  $f$  (see [Sib70]; also textbook discussions in Bertsekas and Tsitsiklis [BeT89], Section 3.5.3, or Facchinei and Pang [FaP03], Section 12.1). When translated to the special case where  $f(x) = Ax - b$  and  $X = \mathfrak{R}^n$ , the conditions for convergence are that  $A \succ 0$ ,  $G$  is positive definite symmetric, and the stepsize  $\gamma$  is small enough. A variant of the projection method for solving weakly monotone variational inequalities is the extragradient method of Korpelevich [Kor76]. A special case where  $f$  is weakly monotone [it has the form  $f(x) = \Phi' \bar{f}(\Phi x)$  for some strongly monotone mapping  $\bar{f}$ ] and the projection method is convergent was given by Bertsekas and Gafni [BeG82].

The proximal method, often referred to as the “proximal point algorithm,” uses  $\gamma \in (0, 1]$  and

$$G = (A + \beta I)^{-1},$$

where  $\beta$  is a positive regularization parameter. An interesting special case arises when the proximal method is applied to the system  $A'\Sigma^{-1}Ax = A'\Sigma^{-1}b$ , with  $\Sigma$ : positive semidefinite symmetric; this is the necessary and sufficient condition for minimization of  $(Ax - b)'\Sigma^{-1}(Ax - b)$ , so the system  $A'\Sigma^{-1}Ax = A'\Sigma^{-1}b$  is equivalent to  $Ax = b$ , for  $A$  not necessarily positive semidefinite, as long as  $Ax = b$  has a solution. Then we obtain the method  $x_{k+1} = x_k - \gamma G(Ax_k - b)$ , where  $\gamma \in (0, 1]$  and

$$G = (A'\Sigma^{-1}A + \beta I)^{-1}A'\Sigma^{-1}.$$

The proximal method was analyzed extensively by Rockafellar [Roc76] for the variational inequality (14) (and even more general problems), and subsequently by several other authors. It is well-known ([Mar70], [Roc76]) that when  $f$  is weakly monotone, the proximal method is convergent.

There are several splitting algorithms that under various assumptions can be shown to be convergent. For example, if  $A$  is positive semidefinite symmetric,  $(B, C)$  is a regular splitting of  $A$  (i.e.  $B + C = A$  and  $B - C \succ 0$ ), and  $G = B^{-1}$ , the algorithm

$$x_{k+1} = x_k - B^{-1}(Ax_k - b),$$



converges to a solution, as shown by Luo and Tseng [LuT89]. Convergent Jacobi and asynchronous or Gauss-Seidel iterations are also well known in dynamic programming, where they are referred to as *value iteration* methods (see e.g., [Ber07], [Put94]). In this context, the system to be solved has the form  $x = g + Px$ , with  $P$  being a substochastic matrix, and under various assumptions on  $P$ , the iteration

$$x_{k+1} = x_k - \gamma((I - P)x_k - g), \quad (15)$$

can be shown to converge asynchronously to a solution for some  $\gamma \in (0, 1]$ . Also asynchronous and Gauss-Seidel versions of iterations of the form (15) are known to be convergent, assuming that the matrix  $P$  is nonnegative (i.e., has nonnegative entries) and irreducible, with  $\rho(P) = 1$  (see [BeT89], p. 517). In the special case where  $P$  or  $P'$  is an irreducible transition probability matrix and  $g = 0$ , the corresponding system,  $x = Px$ , contains as special cases the problems of consensus (multi-agent agreement) and of finding the invariant distribution of an irreducible Markov chain (see [BeT89], Sections 7.3.1 and 7.3.2).

### 3 Stochastic Iterative Methods

We will now consider a stochastic version of the iterative method:

$$x_{k+1} = x_k - \gamma G_k(A_k x_k - b_k), \quad (16)$$

where  $A_k$ ,  $b_k$ , and  $G_k$ , are estimates of  $A$ ,  $b$ , and  $G$ , respectively. We will assume throughout the following condition.

**Assumption 2** *The sequence  $\{A_k, b_k, G_k\}$  is generated by a stochastic process such that*

$$A_k \xrightarrow{a.s.} A, \quad b_k \xrightarrow{a.s.} b, \quad G_k \xrightarrow{a.s.} G.$$

Beyond Assumption 2, we will also need in various parts of the analysis additional conditions, which we will introduce later. However, at this point it is worth pointing out another major class of stochastic algorithms, *stochastic approximation methods* of the Robbins-Monro type, which have an extensive theory (see e.g., Benveniste, Metivier, and Priouret [BMP90], Borkar [Bor08], Kushner and Yin [KuY03], and Meyn [Mey07]), and many applications, including some in ADP (see [BeT96] and the references quoted there). They have the form

$$x_{k+1} = x_k - \gamma_k G(Ax_k - b + w_k), \quad (17)$$

where  $w_k$  is additive zero-mean random noise, and  $\gamma_k > 0$  is a possibly time-varying stepsize. While this iteration also requires Assumption 1 for convergence, it differs in fundamental ways from iteration (16). For example, the method (17) cannot be convergent, even when  $A$  is invertible, unless  $\gamma_k \rightarrow 0$ , for otherwise the term  $\gamma_k G w_k$  will ordinarily have nonzero covariance. By contrast, iteration (16) involves a constant stepsize  $\gamma$  as well as multiplicative (rather than additive) noise. This both enhances its performance and complicates its analysis when  $A$  is singular, as it gives rise to large stochastic perturbations that must be effectively controlled in a stochastic setting. In what follows, we focus exclusively on iteration (16).

#### 3.1 Some Simulation Contexts

The use of simulation often aims to deal with large-scale linear algebra operations, which would be very time consuming or impossible if done exactly. A simulation framework, commonly used in many applications, is to generate an infinite sequence of random variables

$$\{(W_t, v_t) \mid t = 1, 2, \dots\},$$

where  $W_t$  is an  $n \times n$  matrix and  $v_t$  is a vector in  $\mathfrak{R}^n$ , and then estimate  $A$  and  $b$  with  $A_k$  and  $b_k$  given by

$$A_k = \frac{1}{k} \sum_{t=1}^k W_t, \quad b_k = \frac{1}{k} \sum_{t=1}^k v_t. \quad (18)$$

For simplicity, we have described the simulation contexts that use only one sample per iteration. In fact, using multiple samples per iteration is also allowed, as long as the estimates  $A_k$  and  $b_k$  possess appropriate asymptotic behaviors. While we make some probabilistic assumptions on  $A_k$ ,  $b_k$ , and  $G_k$ , the details of the simulation process are not material to our analysis.

We will illustrate some possibilities for obtaining  $A_k$ ,  $b_k$ , and  $G_k$  by simulation. In the first application we aim to solve approximately an overdetermined system by randomly selecting a subset of the constraints (see e.g., [DMM06], [DMMS11]).

**Example 1 (Overdetermined Least Squares Problem)** Consider the least squares problem

$$\min_{x \in \mathfrak{R}^n} \|Cx - d\|_{\xi}^2,$$

where  $C$  is an  $m \times n$  matrix with  $m$  very large and  $n$  is small, and  $\|\cdot\|_{\xi}$  is a weighted Euclidean norm with  $\xi$  being a vector with positive components, i.e.  $\|y\|_{\xi}^2 = \sum_{i=1}^m \xi_i y_i^2$ . Equivalently this is the  $n \times n$  system  $Ax = b$  where

$$A = C' \Xi C, \quad b = C' \Xi d,$$

and  $\Xi$  is the diagonal matrix with  $\xi$  along its diagonal. We generate a sequence of i.i.d. indices  $\{i_1, \dots, i_k\}$  according to a distribution  $\zeta$ , and estimate  $A$  and  $b$  using Eq. (18), where

$$W_t = \frac{\xi_{i_t}}{\zeta_{i_t}} c_{i_t} c_{i_t}', \quad v_t = \frac{\xi_{i_t}}{\zeta_{i_t}} c_{i_t} d_{i_t},$$

$\xi_i$  and  $\zeta_i$  denote the probabilities of the  $i$ th index according to distributions  $\xi$  and  $\zeta$  respectively, and  $c_i'$  is the  $i$ th row of  $C$ . It can be verified that  $A_k = \frac{1}{k} \sum_{t=1}^k W_t \xrightarrow{a.s.} A$  and  $b_k = \frac{1}{k} \sum_{t=1}^k v_t \xrightarrow{a.s.} b$  by the strong law of large numbers for i.i.d. random variables. The system  $Ax = b$  is likely to be singular or nearly singular when  $C$  is singular or nearly singular.

In the second application, we consider the projected equation approach for approximate solution of large linear systems. This approach is widely used in solving forms of Bellman's equation arising in ADP.

**Example 2 (Projected Equations with Galerkin Approximation)** Consider a projected version of the  $m \times m$  system  $y = Py + g$ :

$$\Phi x = \Pi_{\xi} (P\Phi x + g),$$

where  $\Phi$  is an  $m \times n$  matrix whose columns are viewed as features/basis vectors,  $\Pi_{\xi}$  denotes the orthogonal projection onto the subspace  $S = \{\Phi x \mid x \in \mathfrak{R}^n\}$  with respect to the weighted Euclidean norm  $\|\cdot\|_{\xi}$  as in Example 1. Equivalently this is the  $n \times n$  system  $Ax = b$  where

$$A = \Phi' \Xi (I - P) \Phi, \quad b = \Phi' \Xi g.$$

One approach is to generate a sequence of i.i.d. indices  $\{i_1, \dots, i_k\}$  according to a distribution  $\zeta$ , and generate a sequence of independent state transitions  $\{(i_1, j_1), \dots, (i_k, j_k)\}$  according to transition probabilities  $\theta_{ij}$ . We may then estimate  $A$  and  $b$  using Eq. (18), where

$$W_t = \frac{\xi_{i_t}}{\zeta_{i_t}} \phi_{i_t} \left( \phi_{i_t} - \frac{p_{i_t j_t}}{\theta_{i_t j_t}} \phi_{j_t} \right)', \quad v_t = \frac{\xi_{i_t}}{\zeta_{i_t}} \phi_{i_t} g_{i_t},$$

$\phi'_i$  denotes the  $i$ th row of  $\Phi$ , and  $p_{ij}$  denotes the  $(i, j)$ th component of the matrix  $P$ .

Projected equations are central in the theory of Galerkin approximation (see e.g., [Kra72]). They find applications in approximate dynamic programming, which are often solved by simulation-based methods (see e.g., [Ber07], [Put94]). In an important application to cost evaluation of average or discounted cost dynamic programming problems, the matrix  $P$  is the transition probability matrix of an irreducible Markov chain (multiplied with a discount factor in discounted problems). We use the Markov chain instead of i.i.d. sample indices for sampling. In particular, we take  $\xi$  to be the invariant distribution of the Markov chain. We then generate a sequence of indices  $\{i_1, \dots, i_k\}$  according to this Markov chain, and estimate  $A$  and  $b$  using Eq. (18), where

$$W_t = \phi_{i_t}(\phi_{i_t} - \phi_{i_{t+1}})', \quad v_t = \phi_{i_t} g_{i_t}.$$

It can be verified that  $A_k = \frac{1}{k} \sum_{t=1}^k W_t \xrightarrow{a.s.} A$  and  $b_k = \frac{1}{k} \sum_{t=1}^k v_t \xrightarrow{a.s.} b$  by the strong law of large numbers for irreducible Markov chains. This system  $Ax = b$  is typically near-singular for discounted cost problems when the discount factor is close to 1, and is singular or near-singular for average cost problems when the nullspace of  $I - P$  parallels or almost parallels the space spanned by  $\Phi$ .

Note that the simulation formulas used in Examples 1 and 2 satisfy Assumption 2. Moreover they involve low-dimensional linear algebra computations. In Example 1, this is a consequence of the low dimension  $n$  of the solution space of the overdetermined system. In Example 2, this is a consequence of the low dimension  $n$  of the approximation subspace defined by the basis matrix  $\Phi$ .

### 3.2 Convergence Issues in Stochastic Methods

We now provide an overview of the special convergence issues introduced by the stochastic errors, and set the stage for the subsequent analysis. Let us consider the simple special case where

$$A_k \equiv A, \quad G_k \equiv G, \tag{19}$$

so for any solution  $x^*$  of  $Ax = b$ , the iteration (16) is written as

$$x_{k+1} - x^* = (I - \gamma GA)(x_k - x^*) + \gamma G(b_k - b). \tag{20}$$

If we assume that  $G(b_k - b) \in \mathbf{N}(GA)$  for all  $k$ , it can be verified by simple induction that the algorithm evolves according to

$$x_{k+1} - x^* = (I - \gamma GA)^k(x_0 - x^*) + \gamma G \sum_{t=0}^k (b_t - b). \tag{21}$$

Since the last term on the right can accumulate uncontrollably, even under Assumption 2,  $\{x_k\}$  need not converge and may not even be bounded. What is happening here is that the iteration has no mechanism to damp the accumulation of stochastic noise components on  $\mathbf{N}(GA)$ .

Still, however, it is possible that the residual sequence  $\{r_k\}$ , where

$$r_k = Ax_k - b,$$

converges to 0, even though the iterate sequence  $\{x_k\}$  may diverge. To get a sense of this, note that in the deterministic case, by Props. 3 and 4,  $\{z_k\}$  converges to 0 geometrically, and since by Eq. (10) we have  $r_k = AVz_k$ , the same is true for  $\{r_k\}$ . In the special case of iteration (20), where  $A_k \equiv A$  and  $G_k \equiv G$  [cf. Eq. (19)] the residual sequence evolves according to

$$r_{k+1} = (I - \gamma AG)r_k + \gamma AG(b_k - b),$$

and since the iteration can be shown to converge geometrically to 0 when the noise term  $(b_k - b)$  is 0, it also converges to 0 when  $(b_k - b)$  converges to 0.

In the more general case where  $A_k \rightarrow A$  and  $G_k \rightarrow G$ , but  $A_k \neq A$  and  $G_k \neq G$ , the residual sequence evolves according to

$$r_{k+1} = (I - \gamma AG_k)r_k + \gamma AG_k((A - A_k)(x_k - x^*) + b_k - b),$$

a more sophisticated analysis is necessary, and residual convergence comes into doubt. If we can show, under suitable conditions, that the rightmost simulation noise term converges to 0, then the iteration converges to 0. For this it is necessary to show that  $(A_k - A)(x_k - x^*) \rightarrow 0$  so that the noise term converges to 0, i.e., that  $A_k - A$  converges to 0 at a faster rate than the “rate of divergence” of  $(x_k - x^*)$ . In particular, if  $A_k \equiv A$ , the residual sequence  $\{r_k\}$  converges to 0, even though the sequence  $\{x_k\}$  may diverge as indicated earlier for iteration (20).

For another view of the convergence issues, let us consider the decomposition

$$x_k = x^* + Uy_k + Vz_k,$$

where  $x^*$  is a solution of the system  $Ax = b$ , and  $U$  and  $V$  are the matrices of the decomposition of Prop. 2 [cf. Eq. (7)], and

$$y_k = U'(x_k - x^*), \quad z_k = V'(x_k - x^*). \quad (22)$$

In the presence of stochastic error,  $y_k$  and  $z_k$  are generated by an iteration of the form

$$y_{k+1} = y_k - \gamma N z_k + \zeta_k(y_k, z_k), \quad z_{k+1} = z_k - \gamma H z_k + \xi_k(y_k, z_k), \quad (23)$$

where  $\zeta_k(y_k, z_k)$  and  $\xi_k(y_k, z_k)$  are stochasticity-induced errors that are linear functions of  $y_k$  and  $z_k$  [cf. Eq. (9)]. Generally, these errors converge to 0 if  $y_k$  and  $z_k$  are bounded, in which case  $z_k$  converges to 0 (since  $I - \gamma H$  has eigenvalues with positive real parts by Prop. 2), and so does the corresponding residual  $r_k = AVz_k$  [cf. Eq. (10)]. However,  $y_k$  need not converge even if  $y_k$  and  $z_k$  are bounded. Moreover, because of the complexity of iteration (23), the boundedness of  $y_k$  is by no means certain and in fact  $y_k$  may easily become unbounded, as indicated by our earlier divergence example involving Eqs. (20) and (21).

In this paper we will analyze convergence in the general case where the coupling between  $y_k$  and  $z_k$  is strong, and the errors  $\zeta_k(y_k, z_k)$  and  $\xi_k(y_k, z_k)$  may cause divergence. For this case, we will introduce in the next section a modification of the iteration  $x_{k+1} = x_k - \gamma G_k(A_k x_k - b_k)$  in order to attenuate the effect of these errors. We will then show that  $\{x_k\}$  converges.

## 4 Stabilized Stochastic Iterative Methods

In the preceding section we saw that the stochastic iteration

$$x_{k+1} = x_k - \gamma G_k(A_k x_k - b_k), \quad (24)$$

need not be convergent under Assumptions 1 and 2, even though its deterministic counterpart (3) is convergent. Indeed, there are examples for which both iterates and residuals generated by iteration (24) are divergent with probability 1 (see [WaB11]). In the absence of special structure, divergence is common for iteration (24). To remedy this difficulty, we will consider in this section modified versions with satisfactory convergence properties.

### 4.1 A Simple Stabilization Scheme

We first consider a simple stabilization scheme given by

$$x_{k+1} = (1 - \delta_k)x_k - \gamma G_k(A_k x_k - b_k), \quad (25)$$

where  $\{\delta_k\}$  is a scalar sequence from the interval  $(0, 1)$ . For convergence, the sequence  $\{\delta_k\}$  will be required to converge to 0 at a rate that is sufficiently slow (see the following proposition).

The idea here is to stabilize the divergent iteration (24) by shifting the eigenvalues of the iteration matrix  $I - \gamma G_k A_k$  by  $-\delta_k$ , thereby moving them strictly inside the unit circle. For this it is also necessary that the simulation noise sequence  $\{G_k A_k - GA\}$  decreases to 0 faster than  $\{\delta_k\}$  does, so that the shifted eigenvalues remain strictly within the unit circle with sufficient frequency to induce convergence. This is the motivation for the following assumption on the simulation process and the subsequent assumptions on  $\{\delta_k\}$ . The stabilization scheme of Eq. (25) may also help to counteract the combined effect of simulation noise and eigenvalues of  $I - \gamma GA$  that are very close to the boundary of the unit circle, even if  $A$  is only nearly singular (rather than singular). We provide some related analytical and experimental evidence in Sections 5 and 6.

In order to formulate an appropriate assumption for the rate at which  $\delta_k$  converges to 0, we need the following assumption on the convergence rate of the simulation process.

**Assumption 3** *The simulation error sequence*

$$E_k = (A_k - A, G_k - G, b_k - b),$$

*viewed as a  $(2n^2 + n)$ -dimensional vector, satisfies*

$$\limsup_{k \rightarrow \infty} \sqrt{k^q} \mathbf{E} [\|E_k\|^q] < \infty,$$

*for some  $q > 2$ .*

Assumptions 2 and 3 are very general and apply to practical situations that involve a stochastic simulator or a Monte Carlo sampler. For instance, the sample sequence can be generated by independently sampling according to a certain distribution (e.g., [DMM06]); or it can be generated adaptively according to a sequence of importance sampling distributions. Also, the sample sequence can be generated through state transitions of an irreducible Markov chain, as for example in temporal difference methods for cost evaluation problems in ADP (see [BrB96], [Boy02], [NeB03], and [Ber10]), or for general projected equations ([BeY09], [Ber11]). Under natural conditions, all these simulation methods satisfy Assumption 2 through a strong law of large numbers, and Assumption 3 through forms of the central limit theorem (see the discussions in the preceding references). On the other hand, we note that our analysis can accommodate a rate of convergence of  $E_k$  different than  $1/\sqrt{k}$ , as long as the rate of convergence of  $\delta_k$  is appropriately adjusted.

The following proposition shows that if  $\delta_k$  decreases to 0 at a rate sufficiently slower than  $1/\sqrt{k}$ , the sequence of iterates  $\{x_k\}$  converges to a solution. Moreover, it turns out that in sharp contrast to the deterministic version of iteration (24), the stabilized version (25) may converge to only one possible solution of  $Ax = b$ , as the following proposition shows. This solution has the form

$$\hat{x} = (I + UNH^{-1}V')x^* = (GA)^D Gb, \quad (26)$$

where  $U$ ,  $V$ ,  $N$ , and  $H$  are as in the decomposition of Prop. 2, and  $x^*$  is the projection of the origin on the set of solutions (this is the unique solution of minimum Euclidean norm). Note that  $\hat{x}$  is the Drazin inverse solution of the system  $GAx = Gb$ , as noted following Prop. 4. A remarkable fact is that the limit of the iteration does not depend on the initial iterate  $x_0$  as is the case for the deterministic iteration where  $\delta_k \equiv 0$  [cf. Eq. (12) in Prop. 4]. Thus the parameter  $\delta_k$  provides a dual form of stabilization: it counteracts the effect of simulation noise and the effect of the choice of initial iterate  $x_0$ .

**Proposition 5** Let Assumptions 1, 2, and 3 hold, and let  $\{\delta_k\} \subset (0, 1)$  be a decreasing sequence of scalars satisfying

$$\lim_{k \rightarrow \infty} \delta_k = 0, \quad \lim_{k \rightarrow \infty} k^{(1/2 - (\varepsilon + 1)/q)} \delta_k = \infty,$$

where  $\varepsilon$  is some positive scalar and  $q$  is the scalar of Assumption 3. Then there exists  $\bar{\gamma} > 0$  such that for all  $\gamma \in (0, \bar{\gamma}]$  and all initial iterates  $x_0$ , the sequence  $\{x_k\}$  generated by iteration (25) converges with probability 1 to the solution  $\hat{x}$  of  $Ax = b$ , given by Eq. (26).

We will develop the proof of the proposition through a series of preliminary lemmas.

**Lemma 1** Let  $\{\delta_k\}$  satisfy the assumptions of Prop. 5. Then  $E_k/\delta_k \xrightarrow{a.s.} 0$ .

*Proof.* We first note that such  $\{\delta_k\}$  always exists. Since  $q > 2$ , there exists  $\varepsilon > 0$  such that  $1/2 - (1 + \varepsilon)/q > 0$  and  $k^{1/2 - (1 + \varepsilon)/q} \rightarrow \infty$ , implying the existence of  $\{\delta_k\}$  that satisfies the assumptions of Prop. 5 (e.g., we may take  $\delta_k = (k^{1/2 - (1 + \varepsilon)/q})^{-1/2}$ ).

Let  $\epsilon$  be an arbitrary positive scalar. We use the Markov inequality to obtain

$$\mathbf{P} \left( \frac{1}{\delta_k} \|E_k\| \geq \epsilon \right) = \mathbf{P} \left( \left( \frac{1}{\delta_k} \|E_k\| \right)^q \geq \epsilon^q \right) \leq \frac{1}{\epsilon^q} \mathbf{E} \left[ \left( \frac{1}{\delta_k} \|E_k\| \right)^q \right]. \quad (27)$$

By Assumption 3,  $\sqrt{k^q} \mathbf{E}[\|E_k\|^q]$  is a bounded sequence so there exists  $c > 0$  such that

$$\frac{1}{\epsilon^q} \mathbf{E} \left[ \left( \frac{1}{\delta_k} \|E_k\| \right)^q \right] \leq \frac{1}{\epsilon^q} \frac{c}{k^{q/2} \delta_k^q}. \quad (28)$$

From the assumption  $\lim_{k \rightarrow \infty} k^{(1/2 - (1 + \varepsilon)/q)} \delta_k = \infty$ , we have

$$\frac{1}{\epsilon^q} \leq (k^{(1/2 - (1 + \varepsilon)/q)} \delta_k)^q,$$

for sufficiently large  $k$ . Combining this relation with Eqs. (27) and (28), we obtain

$$\mathbf{P} \left( \frac{1}{\delta_k} \|E_k\| \geq \epsilon \right) \leq \frac{c k^{(1/2 - (1 + \varepsilon)/q)q}}{k^{q/2}} = \frac{c}{k^{1 + \varepsilon}},$$

and since  $\varepsilon > 0$  we obtain

$$\sum_{k=1}^{\infty} \mathbf{P} \left( \frac{1}{\delta_k} \|E_k\| \geq \epsilon \right) \leq c \sum_{k=1}^{\infty} \frac{1}{k^{1 + \varepsilon}} < \infty.$$

Using the Borel-Cantelli Lemma, it follows that the event  $\left\{ \frac{1}{\delta_k} \|E_k\| \geq \epsilon \right\}$  occurs only a finite number of times, so  $\frac{1}{\delta_k} \|E_k\| \leq \epsilon$  for  $k$  sufficiently large with probability 1. By taking  $\epsilon \downarrow 0$  we obtain  $E_k/\delta_k \xrightarrow{a.s.} 0$ . ■

**Lemma 2** Let  $\{\delta_k\}$  satisfy the assumptions of Prop. 5, and let  $f$  be a function that is Lipschitz continuous within a neighborhood of  $(A, G, b)$ . Then

$$\frac{1}{\delta_k} \|f(A_k, G_k, b_k) - f(A, G, b)\| \xrightarrow{a.s.} 0.$$

*Proof.* If  $L$  is the Lipschitz constant of  $f$  within a neighborhood of  $(A, G, b)$ , we have within this neighborhood

$$\frac{1}{\delta_k} \|f(A_k, G_k, b_k) - f(A, G, b)\| \leq \frac{L}{\delta_k} \|(A_k - A, G_k - G, b_k - b)\| = \frac{L}{\delta_k} \|E_k\|,$$

for all sufficiently large  $k$  with probability 1. Thus the result follows from Lemma 1.  $\blacksquare$

We will next focus on iteration (25), which is equivalent to

$$x_{k+1} = T_k x_k + g_k,$$

where

$$T_k = (1 - \delta_k)I - \gamma G_k A_k, \quad g_k = \gamma G_k b_k,$$

so that we have

$$T_k \xrightarrow{a.s.} T = I - \gamma GA, \quad g_k \xrightarrow{a.s.} g = \gamma Gb.$$

The key of our convergence analysis is to show that  $T_k$  is contractive with respect to some induced norm, and has modulus that is sufficiently smaller than 1 to attenuate the simulation noise. To be precise, we will find a matrix  $P$  such that

$$\|T_k\|_P = \|P^{-1}T_k P\| \leq 1 - c\delta_k, \quad (29)$$

for  $k$  sufficiently large with probability 1, where  $c$  is some positive scalar.

To this end, we construct a block diagonal decomposition of  $T$  and  $T_k$ . Let  $Q = [U \ V]$  be the orthogonal matrix defined in Prop. 2, and let  $R$  be the matrix  $R = \begin{bmatrix} I & NH^{-1} \\ 0 & I \end{bmatrix}$ . We have

$$GA \stackrel{Q}{\sim} \begin{bmatrix} 0 & N \\ 0 & H \end{bmatrix} \stackrel{R}{\sim} \begin{bmatrix} 0 & 0 \\ 0 & H \end{bmatrix}, \quad (30)$$

where the first similarity relation follows from the nullspace decomposition of Prop. 2, and the second follows by verifying directly that

$$\begin{bmatrix} 0 & N \\ 0 & H \end{bmatrix} R = \begin{bmatrix} 0 & N \\ 0 & H \end{bmatrix} \begin{bmatrix} I & NH^{-1} \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & NH^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & H \end{bmatrix} = R \begin{bmatrix} 0 & 0 \\ 0 & H \end{bmatrix}.$$

The matrix  $H$  has eigenvalues with positive real parts, so  $\rho(I - \gamma H) < 1$  for  $\gamma > 0$  sufficiently small. Thus there exists a matrix  $S$  such that

$$\|I - \gamma H\|_S = \|I - \gamma S^{-1}HS\| < 1.$$

Denoting  $\tilde{H} = S^{-1}HS$ , we obtain from the above relation

$$I \succ (I - \gamma \tilde{H})'(I - \gamma \tilde{H}) = I - \gamma (\tilde{H}' + \tilde{H}) + \gamma^2 \tilde{H}' \tilde{H},$$

implying that

$$\tilde{H} = S^{-1}HS \succ 0. \quad (31)$$

Defining

$$P = QR \begin{bmatrix} I & 0 \\ 0 & S \end{bmatrix} = [U \ V] \begin{bmatrix} I & NH^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & S \end{bmatrix}, \quad (32)$$

we can verify using Eq. (30) that

$$GA \stackrel{P}{\sim} \begin{bmatrix} I & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & H \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & S \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \tilde{H} \end{bmatrix}. \quad (33)$$

We are now ready to prove the contraction property of  $T_k$  [cf. Eq. (29)] by using the matrix  $P$  constructed above. Let us mention that although we have constructed  $P$  based on a single  $\gamma$ , any such  $P$  will work in the subsequent analysis.

**Lemma 3** *Let the assumptions of Prop. 5 hold, and let  $P$  be the matrix given by Eq. (32). There exist scalars  $c, \bar{\gamma} > 0$  such that for any  $\gamma \in (0, \bar{\gamma}]$  and all initial iterates  $x_0$ , we have*

$$\|(1 - \delta_k)I - \gamma G_k A_k\|_P \leq 1 - c\delta_k,$$

for  $k$  sufficiently large with probability 1.

*Proof.* By using Eq. (32) and Eq. (33), we obtain for all  $k$  that

$$(1 - \delta_k)I - \gamma GA \stackrel{P}{\approx} \begin{bmatrix} (1 - \delta_k)I & 0 \\ 0 & (1 - \delta_k)I - \gamma \tilde{H} \end{bmatrix}.$$

Since  $\tilde{H} \succ 0$  [cf. Eq. (31)], we have

$$((1 - \delta_k)I - \gamma \tilde{H})'((1 - \delta_k)I - \gamma \tilde{H}) = (1 - \delta_k)^2 I - \gamma(1 - \delta_k)(\tilde{H}' + \tilde{H}) + \gamma^2 \tilde{H}' \tilde{H} \prec (1 - \delta_k)^2 I,$$

for  $\gamma > 0$  sufficiently small. It follows that, there exists  $\bar{\gamma} > 0$  such that for all  $\gamma \in (0, \bar{\gamma}]$  and  $k$  sufficiently large

$$\|(1 - \delta_k)I - \gamma \tilde{H}\| < 1 - \delta_k.$$

Thus we have

$$\|(1 - \delta_k)I - \gamma GA\|_P = \left\| \begin{bmatrix} (1 - \delta_k)I & 0 \\ 0 & (1 - \delta_k)I - \gamma \tilde{H} \end{bmatrix} \right\| = 1 - \delta_k,$$

for all  $\gamma$  in some interval  $(0, \bar{\gamma}]$  and  $k$  sufficiently large. Finally, by using the triangle inequality and the fact  $O(E_k)/\delta_k \xrightarrow{a.s.} 0$  (cf. Lemma 2), we obtain

$$\|(1 - \delta_k)I - \gamma G_k A_k\|_P \leq \|(1 - \delta_k)I - \gamma GA\|_P + \gamma \|G_k A_k - GA\|_P = 1 - \delta_k + O(\|E_k\|) \leq 1 - c\delta_k,$$

for  $k$  sufficiently large with probability 1, where  $c$  is some positive scalar.  $\blacksquare$

**Lemma 4** *Under the assumptions of Prop. 5, the sequence  $\{x_k\}$  is bounded with probability 1.*

*Proof.* For any solution  $\hat{x}$  of  $Ax = b$ , using Eq. (25) we have

$$x_{k+1} - \hat{x} = ((1 - \delta_k)I - \gamma G_k A_k)(x_k - \hat{x}) - \delta_k \hat{x} + \gamma G_k (b_k - A_k \hat{x}),$$

from which we obtain

$$\|x_{k+1} - \hat{x}\|_P \leq \|(1 - \delta_k)I - \gamma G_k A_k\|_P \|x_k - \hat{x}\|_P + \delta_k \|\hat{x}\|_P + \gamma \|G_k (b_k - A_k \hat{x})\|_P,$$

where  $\|\cdot\|_P$  is the norm of Lemma 3. Since  $\|G_k (b_k - A_k \hat{x})\|_P = O(E_k)$  and  $O(E_k)/\delta_k \xrightarrow{a.s.} 0$ , there exists  $\bar{c} > 0$  such that

$$\delta_k \|\hat{x}\|_P + \gamma \|G_k (b_k - A_k \hat{x})\|_P \leq \bar{c}\delta_k,$$

for  $k$  sufficiently large with probability 1. Thus by using Lemma 3 in conjunction with the preceding two inequalities, we obtain

$$\|x_{k+1} - \hat{x}\|_P \leq (1 - c\delta_k) \|x_k - \hat{x}\|_P + \bar{c}\delta_k.$$



Hence if  $\|x_k - \hat{x}\|_P \geq \bar{c}/c$ , we have

$$\|x_{k+1} - \hat{x}\|_P \leq \|x_k - \hat{x}\|_P - c\delta_k\|x_k - \hat{x}\|_P + \bar{c}\delta_k \leq \|x_k - \hat{x}\|_P,$$

implying that  $\{x_k - \hat{x}\}$  is bounded with probability 1, hence  $\{x_k\}$  is also bounded with probability 1.  $\blacksquare$

Our proof idea of Prop. 5 is to start with  $x^*$ , the solution of  $Ax = b$  that has minimum norm, and to decompose the sequence  $\{x_k - x^*\}$  into the two sequences  $\{Uy_k\}$  and  $\{Vz_k\}$ , which lie in the orthogonal subspaces  $\mathbf{N}(A)$  and  $\mathbf{N}(A)^\perp$ , respectively (cf. the decomposition of Prop. 3). Thus we will view iteration (25) as two interconnected subiterations, one for  $y_k$  and the other for  $z_k$ .

**Proof of Proposition 5:** We rewrite iteration (25) as

$$x_{k+1} - x^* = ((1 - \delta_k)I - \gamma GA)(x_k - x^*) - \delta_k x^* + \gamma(GA - G_k A_k)(x_k - x^*) + \gamma G_k(b_k - A_k x^*). \quad (34)$$

Since  $\{x_k\}$  is bounded with probability 1 (cf. Lemma 4), we have

$$\gamma(GA - G_k A_k)(x_k - x^*) + \gamma G_k(b_k - A_k x^*) = O(E_k), \quad w.p.1.$$

Let  $Q = [U \ V]$  be the orthogonal matrix used to construct the decomposition of  $I - \gamma GA$  in Section 2. We multiply both sides of Eq. (34) with  $Q'$  on the left, apply the above relation, and obtain

$$Q'(x_{k+1} - x^*) = Q'((1 - \delta_k)I - \gamma GA)QQ'(x_k - x^*) - \delta_k Q'x^* + O(E_k). \quad (35)$$

Let us define [cf. Eq. (8)]

$$y_k = U'(x_k - x^*), \quad z_k = V'(x_k - x^*), \quad Q'(x_k - x^*) = \begin{bmatrix} y_k \\ z_k \end{bmatrix}.$$

Then Eq. (35) can be rewritten as

$$\begin{bmatrix} y_{k+1} \\ z_{k+1} \end{bmatrix} = \begin{bmatrix} (1 - \delta_k)I & -\gamma N \\ 0 & (1 - \delta_k)I - \gamma H \end{bmatrix} \begin{bmatrix} y_k \\ z_k \end{bmatrix} - \delta_k \begin{bmatrix} 0 \\ V'x^* \end{bmatrix} + O(E_k), \quad (36)$$

where we have used the fact  $U'x^* = 0$ , which follows from the orthogonality of  $x^*$  to  $\mathbf{N}(A)$ , the subspace that is parallel to the set of solutions of  $Ax = b$ . By letting  $\tilde{y}_k = y_k - NH^{-1}z_k - NH^{-1}V'x^*$ , we have from Eq. (36)

$$\tilde{y}_{k+1} = (1 - \delta_k)\tilde{y}_k + O(E_k). \quad (37)$$

We can now analyze the asymptotic behavior of  $\{z_k\}$ ,  $\{\tilde{y}_k\}$ , and  $\{y_k\}$  according to Eqs. (36)-(37).

- The  $z_k$ -portion is given by

$$z_{k+1} = ((1 - \delta_k)I - \gamma H)z_k - \delta_k V'x^* + O(E_k),$$

where  $\|(1 - \delta_k)I - \gamma H\|_S \xrightarrow{a.s.} \eta \in (0, 1)$  for some norm  $\|\cdot\|_S$  and  $\gamma$  within a sufficiently small interval  $(0, \bar{\gamma}]$ . Since  $\delta_k \downarrow 0$  and  $E_k \xrightarrow{a.s.} 0$ , we then obtain  $z_k \xrightarrow{a.s.} 0$ .

- The  $\tilde{y}_k$ -portion satisfies that

$$\|\tilde{y}_{k+1}\| \leq (1 - \delta_k)\|\tilde{y}_k\| + O(\|E_k\|).$$

Since  $\sum_{k=0}^{\infty} \delta_k = \infty$  and  $O(E_k)/\delta_k \xrightarrow{a.s.} 0$ , it follows from a well-known result that  $\lim_{k \rightarrow \infty} \|\tilde{y}_k\| = 0$  (cf. [Ber99], Lemma 1.5.1) with probability 1. Therefore  $\tilde{y}_k \xrightarrow{a.s.} 0$ .

- The  $y_k$ -portion is given by

$$y_k = \tilde{y}_k + NH^{-1}z_k + NH^{-1}V'x^*.$$

By using the earlier result  $z_k \xrightarrow{a.s.} 0$  and  $\tilde{y}_k \xrightarrow{a.s.} 0$ , we obtain  $y_k \xrightarrow{a.s.} NH^{-1}V'x^*$ .

To summarize, we have shown that  $z_k \xrightarrow{a.s.} 0$  and  $y_k \xrightarrow{a.s.} NH^{-1}V'x^*$ . Therefore  $x_k = x^* + Uy_k + Vz_k$  converges with probability 1 to the vector  $\hat{x} = (I + UNH^{-1}V')x^*$  given by Eq. (26).  $\blacksquare$

To understand the convergence mechanism of the algorithm, we may review the line of proof, for the simpler case where there is no simulation noise, i.e.,  $A_k \equiv A$ ,  $b_k \equiv b$ ,  $G_k \equiv G$ . Then the stabilized iteration (25) is equivalent to the decomposed version

$$y_{k+1} = (1 - \delta_k)y_k - \gamma Nz_k, \quad z_{k+1} = ((1 - \delta_k)I - \gamma H)z_k - \delta_k V'x^*$$

[cf. Eq. (36)]. The iteration for  $z_k$  is subject to the slowly varying driving term  $-\delta_k V'x^*$ , but has geometrically/fast converging dynamics. As a result the iteration “sees” the driving term as being essentially constant, and we have  $z_k \approx -\delta_k(\gamma H)^{-1}V'x^*$  for sufficiently large  $k$ . After substituting this expression in the preceding iteration for  $y_k$ , we obtain

$$y_{k+1} \approx (1 - \delta_k)y_k + \gamma\delta_k N(\gamma H)^{-1}V'x^*,$$

which yields  $y_k \rightarrow NH^{-1}V'x^*$  and

$$x_k = x^* + Uy_k + Vz_k \rightarrow (I + UNH^{-1}V')x^* = \hat{x}.$$

The preceding argument also provides insight into the rate of convergence of the algorithm. When there is no stochastic noise, the iterates  $y_k$  and  $z_k$  operate on two different time scales. The iteration of  $z_k$  is naturally contractive, and can be equivalently written as

$$\frac{z_{k+1}}{\delta_{k+1}} + (\gamma H)^{-1}V'x^* \approx \frac{\delta_k}{\delta_{k+1}} ((1 - \delta_k)I - \gamma H) \left( \frac{z_k}{\delta_k} + (\gamma H)^{-1}V'x^* \right),$$

where both sides of the above relation are approximately equal and differ only by a term decreasing to 0. This implies  $z_k/\delta_k \xrightarrow{a.s.} -(\gamma H)^{-1}V'x^*$ . Therefore  $z_k$  converges linearly to the slowly decreasing bias  $-\delta_k(\gamma H)^{-1}V'x^*$ . On the other hand, the iteration of  $y_k$  is convergent due to the stabilization with modulus  $(1 - \delta_k)$ . Thus  $y_k$  converges to its limit at a rate much slower than the geometric rate.

In the case where the stochastic noise satisfies Assumption 3, we have  $E_k/\delta_k \xrightarrow{a.s.} 0$ . Thus the effect of the noise eventually becomes negligible compared with the effect of stabilization. This suggests that the asymptotic behavior of the stochastic stabilized iteration is the same with that of the stabilized iteration in the absence of stochastic noise. We will address this issue in a future publication. Let us also mention that, there may exist stochastic noise whose asymptotic behavior does not conform to Assumption 3. In this case, as long as we choose a sequence of  $\{\delta_k\}$  such that  $E_k/\delta_k \xrightarrow{a.s.} 0$ , the convergence results of the stabilized iterations still follow.

## 4.2 A Class of Stabilization Schemes and a Unified Convergence Analysis

While the stabilization scheme of the preceding section is simple and general, there are variations of this scheme that may be better suited for specific iterative algorithms. In this section, we will introduce a broad class of stabilization schemes, within which we will embed the algorithm of the preceding section. We will provide a unified convergence analysis that can be applied to that algorithm and other alternative algorithms as well. We will then discuss several such algorithms in Section 4.3.

We first write the deterministic iteration (2) in the form

$$x_{k+1} = Tx_k + g,$$

where we define

$$T = I - \gamma GA, \quad g = \gamma Gb.$$

We consider the modified/stabilized stochastic iteration

$$x_{k+1} = T_k x_k + g_k, \tag{38}$$

where the  $n \times n$  matrix  $T_k$  and the  $n$ -dimensional vector  $g_k$  are approximations of  $T$  and  $g$  of the form

$$T_k = T + \delta_k D + O(\delta_k^2 + \|E_k\|), \quad g_k = g + \delta_k d + O(\delta_k^2 + \|E_k\|). \quad (39)$$

Here  $D$  is an  $n \times n$  matrix and  $d$  is a vector in  $\mathfrak{R}^n$ ,  $\delta_k \in (0, 1)$  is the stabilization parameter, and  $E_k$  represents the simulation error, as earlier. We may view Eq. (39) as a form of first order expansion of  $T_k$  and  $g_k$  with respect to  $\delta_k$ . The algorithm (25) of the preceding section is obtained for  $D = -I$  and  $d = 0$ .

The following convergence result shows that if  $\|T_k\|_P$  has a certain contraction property for a suitable matrix  $P$ , and if  $D$  and  $d$  satisfy a certain consistency condition, the stochastic iteration (38) converges to a solution of  $Ax = b$  that is uniquely determined by  $D$  and  $d$ .

**Proposition 6 (Convergence of General Stabilization Schemes)** *Let Assumptions 1, 2, and 3 hold, and let  $\gamma$  be a sufficiently small scalar such that the deterministic iteration (3) converges. Assume that  $D$  and  $d$  are such that there exists  $c > 0$  and an invertible matrix  $P$  such that*

$$\|T_k\|_P \leq 1 - c\delta_k, \quad \forall k \text{ sufficiently large}, \quad (40)$$

*and let  $\{\delta_k\}$  satisfy the same assumptions as in Prop. 5. Then there is a unique solution  $\hat{x}$  to the system of equations*

$$\hat{\Pi}(Dx + d) = 0, \quad Ax = b, \quad (41)$$

*where  $\hat{\Pi}$  denotes orthogonal projection onto  $\mathbf{N}(A)$  with respect to the norm  $\|\cdot\|_P$ . Furthermore, for all initial iterates  $x_0$ , the sequence  $\{x_k\}$  generated by iteration (38) converges to  $\hat{x}$  with probability 1.*

We develop the proof of the proposition through some preliminary lemmas. We first establish the existence and uniqueness of the solution  $\hat{x}$ . Note that in the algorithm of the preceding section,  $\hat{x}$  is the Drazin inverse solution of  $Ax = b$ , but in the more general case considered at present  $\hat{x}$  depends on  $D$  and  $d$ .

**Lemma 5** *Under the assumptions of Prop. 6, the system (41) has a unique solution.*

*Proof.* It can be seen that an equivalent form of system (41) is

$$(Dx + d)'(P^{-1})'P^{-1}y = 0, \quad \forall y \in \mathbf{N}(A), \quad Ax = b,$$

which is also equivalent to

$$(Dx + d)'(P^{-1})'P^{-1}(y - x) = 0, \quad \forall y \in x + \mathbf{N}(A), \quad Ax = b.$$

Therefore the system (41) can be written equivalently as the following projected equation

$$x = \tilde{\Pi}(x + \beta(Dx + d)),$$

where  $\tilde{\Pi}$  is the orthogonal projection matrix with respect to  $\|\cdot\|_P$  on the solution set

$$X^* = \{x \mid Ax = b\} = \{x \mid Tx + g = x\},$$

and  $\beta$  is any nonzero scalar. By using Eq. (40), there exists  $\beta > 0$  and  $\eta \in (0, 1)$  such that

$$\|T + \beta D\|_P \leq \eta < 1.$$

For any  $x, y \in X^*$ , we have

$$\begin{aligned} \|x + \beta(Dx + d) - y - \beta(Dy + d)\|_P &= \|Tx + g + \beta(Dx + d) - Ty - g - \beta(Dy + d)\|_P \\ &= \|(T + \beta D)(x - y)\|_P \\ &\leq \eta \|x - y\|_P, \end{aligned}$$

where the first equality holds because  $x, y \in X^*$ , so  $x = Tx + g$  and  $y = Ty + g$ . By applying the projection  $\tilde{\Pi}$ , the contractive property is preserved, i.e.,

$$\|\tilde{\Pi}(x + \beta(Dx + d)) - \tilde{\Pi}(y + \beta(Dy + d))\|_P \leq \eta \|x - y\|_P, \quad \forall x, y \in X^*.$$

This implies that the projected mapping  $x \mapsto \tilde{\Pi}(x + \beta(Dx + d))$  is also a contraction on  $X^*$ . It follows that it has a unique fixed point in  $X^*$ , which is the unique solution of the system (41).  $\blacksquare$

We next construct a nullspace decomposition similar to the one of Prop. 2. Let  $\bar{U}$  be an orthonormal basis for the subspace  $P^{-1}\mathbf{N}(A)$ , which is equal to  $P^{-1}\mathbf{N}(GA)$  by Assumption 1. Since  $P^{-1}\mathbf{N}(GA) = \mathbf{N}(GAP) = \mathbf{N}(P^{-1}GAP)$ , we see that  $\bar{U}$  is an orthonormal basis for  $\mathbf{N}(P^{-1}GAP)$ . Let also  $\bar{V}$  be an orthonormal basis for the complementary subspace  $\mathbf{N}(P^{-1}GAP)^\perp$ , and let  $Q$  be the orthogonal matrix

$$\bar{Q} = [\bar{U} \ \bar{V}].$$

We use  $\bar{U}$  and  $\bar{V}$  to construct a nullspace decomposition of  $P^{-1}GAP$  and  $P^{-1}TP$ , with the procedure that was used for nullspace decomposition of  $GA$  and  $T$  (cf. Prop. 2). This yields

$$P^{-1}(I - T)P = \gamma(P^{-1}GAP) = \bar{Q} \begin{bmatrix} 0 & \gamma\bar{N} \\ 0 & \gamma\bar{H} \end{bmatrix} \bar{Q}', \quad P^{-1}TP = \bar{Q} \begin{bmatrix} I & -\gamma\bar{N} \\ 0 & I - \gamma\bar{H} \end{bmatrix} \bar{Q}', \quad (42)$$

where  $\bar{N}$  and  $\bar{H}$  are matrices defined analogously to the matrices  $N$  and  $H$  of Prop. 2. As in the case of  $H$ , it follows from Assumption 1 that the eigenvalues of  $\bar{H}$  have positive real parts, since  $GA$  and  $P^{-1}GAP$  have the same eigenvalues. Hence for all  $\gamma$  within a sufficiently small interval  $(0, \bar{\gamma}]$  such that the deterministic iteration (3) converges, the eigenvalues of  $I - \gamma\bar{H}$  lie strictly within the unit circle. In what follows, we will always assume that  $\gamma$  has been chosen within such an interval.

The following lemma shows that this decomposition is block diagonal. The lemma relies only on the assumption (40) and not on the detailed nature of  $D$  and  $d$ . Thus the lemma highlights the role of  $P$  as a matrix that block-diagonalizes the iteration mapping along two orthogonal subspaces, similar to the corresponding matrix  $P$  of the preceding section [cf. Eq. (32)].

**Lemma 6** *Under the assumptions of Prop. 6, the decomposition (42) is block diagonal, i.e.,  $\bar{N} = 0$ .*

*Proof.* Since

$$\|P^{-1}TP\| = \|T\|_P = \lim_{k \rightarrow \infty} \|T_k\|_P \leq \lim_{k \rightarrow \infty} (1 - c\delta_k) = 1,$$

and also  $\|T\|_P \geq \rho(T) = 1$  [since 1 is an eigenvalue of  $T$ , cf. Eq. (42)], we conclude that  $\|P^{-1}TP\| = 1$ . Assume to arrive at a contradiction, that some component of  $-\gamma\bar{N}$  is nonzero, say the component  $m_{ij}$  of the  $i$ th row and  $j$ th column of the matrix

$$M = \begin{bmatrix} I & -\gamma\bar{N} \\ 0 & I - \gamma\bar{H} \end{bmatrix},$$

(note that  $i$  is an index of the first block, and  $j$  is an index of the second block.) Consider a vector  $y$  with components  $y_i \neq 0$ ,  $y_j \neq 0$ ,  $y_\ell = 0$  for all  $\ell \neq i, j$ , such that the  $i$ th component of  $My$  satisfies

$$(My)_i^2 = (y_i + m_{ij}y_j)^2 > y_i^2 + y_j^2.$$

Let also  $x$  be such that  $\bar{Q}'x = y$ . Using Eq. (42) and the fact  $\bar{Q}'\bar{Q} = I$ , we have

$$\|P^{-1}TPx\|^2 = \|M\bar{Q}'x\|^2 = \|My\|^2 \geq (My)_i^2 > y_i^2 + y_j^2 = \|y\|^2 = \|\bar{Q}'x\|^2 = \|x\|^2,$$

where the last equality holds since  $\bar{Q}'$  is orthogonal. Thus we have  $\|P^{-1}TPx\| > \|x\|$ , which contradicts the fact  $\|P^{-1}TP\| = 1$  shown earlier.  $\blacksquare$

We are now ready to prove convergence of  $\{x_k\}$  to the solution of the system (41). Note that this system can be written in a more explicit form by observing that the first equation  $\hat{\Pi}(Dx + d) = 0$  means that  $(Dx + d)$  is orthogonal to  $\mathbf{N}(A)$  in the scaled geometry of the norm  $\|\cdot\|_P$ , i.e.,

$$P^{-1}(Dx + d) \perp P^{-1}\mathbf{N}(A). \quad (43)$$

Then the equation  $\hat{\Pi}(Dx + d) = 0$  or its equivalent form (43) is written as  $\bar{U}'P^{-1}(Dx + d) = 0$  and the system (41) is written as

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \bar{Q}'P^{-1}(Dx + d) = 0, \quad Ax = b. \quad (44)$$

**Proof of Proposition 6:** We first prove that  $\{x_k\}$  is bounded, using the solution  $\hat{x}$  of the system (41). Subtracting  $\hat{x}$  from both sides of Eq. (38), and using the relation  $\hat{x} = T\hat{x} + g$ , we obtain

$$x_{k+1} - \hat{x} = T_k(x_k - \hat{x}) + T_k\hat{x} - \hat{x} + g_k = T_k(x_k - \hat{x}) + (T_k - T)\hat{x} + g_k - g,$$

and finally, using Eq. (39),

$$x_{k+1} - \hat{x} = T_k(x_k - \hat{x}) + \delta_k(D\hat{x} + d) + O(\delta_k^2 + \|E_k\|). \quad (45)$$

By using the assumption  $\|T_k\|_P \leq 1 - c\delta_k$ , we further obtain

$$\|x_{k+1} - \hat{x}\|_P \leq \|T_k\|_P \|x_k - \hat{x}\|_P + \|\delta_k(D\hat{x} + d) + O(\delta_k^2 + \|E_k\|)\|_P \leq (1 - c\delta_k)\|x_k - \hat{x}\|_P + O(\delta_k).$$

It follows similar to the proof of Lemma 4 that  $\{x_k\}$  is bounded with probability 1.

We will use the decomposition of iteration (38) to prove convergence. From Lemma 6, we have

$$P^{-1}TP = \bar{Q} \begin{bmatrix} I & 0 \\ 0 & I - \gamma\bar{H} \end{bmatrix} \bar{Q}'.$$

Combining this equation with Eq. (39), we obtain

$$P^{-1}T_kP = P^{-1}(T + \delta_k D)P + O(\delta_k^2 + \|E_k\|) = \bar{Q} \begin{bmatrix} I - \delta_k J & O(\delta_k) \\ O(\delta_k) & I - \gamma\bar{H} + O(\delta_k) \end{bmatrix} \bar{Q}' + O(\delta_k^2 + \|E_k\|), \quad (46)$$

where  $J$  is the upper diagonal block of  $-P^{-1}DP$ . From Eqs. (40) and (46), we have

$$\|I - \delta_k J\| \leq \|P^{-1}T_kP\| + O(\delta_k^2 + \|E_k\|) = \|T_k\|_P + O(\delta_k^2 + \|E_k\|) \leq 1 - \bar{c}\delta_k, \quad (47)$$

for some positive scalar  $\bar{c}$ , for  $k$  sufficiently large with probability 1.

We now introduce scaled iterates  $y_k, z_k$ , and a vector  $\bar{x}$  defined by

$$\begin{bmatrix} y_k \\ z_k \end{bmatrix} = \bar{Q}'P^{-1}(x_k - \hat{x}), \quad \begin{bmatrix} 0 \\ \bar{x} \end{bmatrix} = \bar{Q}'P^{-1}(D\hat{x} + d),$$

where the top component of the vector in the equation on the right is 0 in light of Eq. (44). We rewrite Eq. (46) in the equivalent form

$$\bar{Q}'P^{-1}T_k = \begin{bmatrix} I - \delta_k J & O(\delta_k) \\ O(\delta_k) & I - \gamma\bar{H} + O(\delta_k) \end{bmatrix} \bar{Q}'P^{-1} + O(\delta_k^2 + \|E_k\|),$$

and we use it after applying the transformation  $\bar{Q}'P^{-1}$  to iteration (45), to obtain the following scaled form of this iteration:

$$\begin{bmatrix} y_{k+1} \\ z_{k+1} \end{bmatrix} = \begin{bmatrix} I - \delta_k J & O(\delta_k) \\ O(\delta_k) & I - \gamma\bar{H} + O(\delta_k) \end{bmatrix} \begin{bmatrix} y_k \\ z_k \end{bmatrix} + \delta_k \begin{bmatrix} 0 \\ \bar{x} \end{bmatrix} + O(\delta_k^2 + \|E_k\|). \quad (48)$$

We now analyze the asymptotic behavior of the sequences of scaled iterates  $\{y_k\}$  and  $\{z_k\}$ .

- The  $z_k$ -portion of Eq. (48) is

$$z_{k+1} = (I - \gamma\bar{H} + O(\delta_k))z_k + O(\delta_k)y_k + \delta_k\bar{x} + O(\delta_k^2 + \|E_k\|).$$

By the boundedness of  $\{x_k\}$ , the sequence  $\{y_k\}$  is also bounded with probability 1, implying

$$O(\delta_k)y_k + \delta_k\bar{x} + O(\delta_k^2 + \|E_k\|) = O(\delta_k).$$

Hence the  $z_k$ -portion can be simplified to

$$z_{k+1} = (I - \gamma\bar{H} + O(\delta_k))z_k + O(\delta_k),$$

where  $\delta_k \rightarrow 0$ , so that  $(I - \gamma\bar{H} + O(\delta_k)) \xrightarrow{a.s.} I - \gamma\bar{H}$ . Since  $I - \gamma\bar{H}$  is a contraction for  $\gamma \in (0, \bar{\gamma}]$ , it follows that  $z_k \xrightarrow{a.s.} 0$ .

- The  $y_k$ -portion of Eq. (48) is

$$y_{k+1} = (I - \delta_k J)y_k + O(\delta_k)z_k + O(\delta_k^2 + \|E_k\|).$$

From this equation and Eq. (47), it follows that

$$\|y_{k+1}\| \leq \|I - \delta_k J\| \|y_k\| + O(\delta_k^2 + \|E_k\| + \delta_k z_k) \leq (1 - \bar{c}\delta_k) \|y_k\| + O(\delta_k^2 + \|E_k\| + \delta_k z_k).$$

Using the assumption  $\sum_{k=0}^{\infty} \delta_k = \infty$ , and the fact  $O(\delta_k^2 + \|E_k\| + \delta_k z_k)/\delta_k \rightarrow 0$ , we obtain  $y_k \xrightarrow{a.s.} 0$ .

In summary, we have  $z_k \xrightarrow{a.s.} 0$  and  $y_k \xrightarrow{a.s.} 0$ , so  $x_k - \hat{x} \xrightarrow{a.s.} 0$ . ■

The preceding proposition can be used to prove convergence of a variety of iterations of the form (38): we only need to verify that the condition (40) is satisfied for some matrix  $P$ . The solution  $\hat{x}$  depends on  $D$  and  $d$ . In particular, if we let  $D = -I$  and  $d = 0$ , we recover the result of Prop. 5, and  $\hat{x}$  is the Drazin inverse solution of  $GAx = Gb$ . As will be illustrated by the subsequent analysis, other possible limit points exist.

### 4.3 Some Instances of Alternative Stabilization Schemes

The idea of the simple stabilization scheme of Section 4.1 is to shift all eigenvalues of  $I - \gamma GA$  by  $-\delta_k$ , so that the modified iteration has a modulus sufficiently smaller than 1. We will now discuss some other schemes that work similarly, and can be shown to be convergent using the unified analysis of Section 4.2.

#### 4.3.1 A Stabilization Scheme for Fixed Point Iterations

One alternative is to multiply the entire iteration with  $(1 - \delta_k)$ :

$$x_{k+1} = (1 - \delta_k)(x_k - \gamma G_k(A_k x_k - b_k)). \quad (49)$$

When this iteration is applied to the fixed point problem  $x = Fx + b$  with  $A = I - F$ ,  $G_k = I$ , and  $\gamma = 1$ , it yields the iteration

$$x_{k+1} = (1 - \delta_k)(F_k x_k + b_k),$$

where  $F_k$  and  $b_k$  are simulation-based estimates of  $F$  and  $b$ .

We may write iteration (49) as

$$x_{k+1} = T_k x_k + g_k,$$

[cf. Eq. (38)] where

$$T_k = (1 - \delta_k)(I - \gamma G_k A_k), \quad g_k = \gamma(1 - \delta_k)G_k b_k.$$

The first order approximations of  $T_k$  and  $g_k$  are

$$T_k \approx T + \delta_k D, \quad g_k \approx g + \delta_k d$$

[cf. Eq. (39)], where we can verify that

$$D = -(I - \gamma GA) = -T, \quad d = -\gamma Gb = -g.$$

We have the following convergence result by applying Prop. 6.

**Proposition 7** *Let Assumptions 1, 2, and 3 hold, and let  $\{\delta_k\}$  satisfy the assumptions of Prop. 5. Then for any  $\gamma$  within a sufficiently small interval  $(0, \bar{\gamma}]$  and all initial iterates  $x_0$ , the sequence  $\{x_k\}$  generated by iteration (49) converges with probability 1 to the solution  $\hat{x}$  of  $Ax = b$  given by Eq. (26).*

*Proof.* Let  $P$  be the transformation matrix defined in the analysis preceding Lemma 3:

$$P = [U \ V] \begin{bmatrix} I & NH^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & S \end{bmatrix},$$

where  $U, V, N, H$  are the matrices used to construct the decomposition of  $I - \gamma GA$  in Prop. 2, and  $S$  is the matrix defined in the analysis preceding Lemma 3. Using the line of analysis of Lemma 3, we obtain

$$T + \delta_k D = (1 - \delta_k)(I - \gamma GA) \stackrel{P}{\sim} (1 - \delta_k) \begin{bmatrix} I & 0 \\ 0 & I - \gamma \tilde{H} \end{bmatrix},$$

where  $\tilde{H} = S^{-1}HS$  [cf. Eq. (31)], and we also have

$$\|T + \delta_k D\|_P = (1 - \delta_k)\|I - \gamma GA\|_P \leq 1 - \delta_k.$$

Then there exists  $c > 0$  such that

$$\|T_k\|_P \leq \|T + \delta_k D\|_P + O(\delta_k^2 + \|E_k\|_P) \leq 1 - c\delta_k,$$

for  $k$  sufficiently large with probability 1, so the assumptions of Prop. 6 are satisfied. It follows that  $x_k$  converges to the unique solution  $\hat{x}$  of Eq. (41) or Eq. (44) with probability 1.

Now we consider the limit point. By applying Eq. (32) and the equation

$$Dx + d = -(Tx + g) = -x$$

to Eq. (44) with  $\bar{Q} = I$  [since  $P^{-1}TP$  already takes the desired form of nullspace decomposition, we let  $\bar{Q} = I$  so that Eq. (42) holds], we obtain

$$0 = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \left( \begin{bmatrix} I & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & -NH^{-1} \\ 0 & I \end{bmatrix} [U \ V]' \right) (-x), \quad Ax = b,$$

or equivalently

$$U'x - NH^{-1}V'x = 0, \quad V'x = V'x^*.$$

Thus  $\hat{x} = (I + UNH^{-1}V')x^*$ , as given by Eq. (26), is the unique solution to the above system. ■

### 4.3.2 A Stabilization Scheme by Selective Eigenvalue Shifting

Another alternative stabilization scheme is to shift by  $-\delta_k$  only those eigenvalues of  $I - \gamma GA$  that are equal to 1, instead of shifting all eigenvalues. This avoids the perturbation on those eigenvalues that are strictly contained in the unit circle, and reduces the bias induced by  $-\delta_k$  on the iterate portion  $V'x_k$  that lies in  $\mathbf{N}(A)^\perp$ . Note that this approach requires knowledge of the eigenspace of  $I - \gamma GA$  corresponding to eigenvalue 1, i.e., the nullspace  $\mathbf{N}(A)$ . In some cases, we can estimate a projection matrix of  $\mathbf{N}(A)$  based on stochastic simulation (we refer to the paper [WaB11] for more details).

Suppose that we can form a sequence of estimates  $\{\Pi_k\}$  such that

$$\Pi_k \xrightarrow{a.s.} \Pi_{\mathbf{N}(A)},$$

where  $\Pi_{\mathbf{N}(A)}$  is the orthogonal projection matrix onto  $\mathbf{N}(A)$  with respect to the Euclidean norm. We consider the stabilized iteration

$$x_{k+1} = (I - \delta_k \Pi_k)x_k - \gamma G_k(A_k x_k - b_k), \quad (50)$$

which can be written as

$$x_{k+1} = T_k x_k + g_k,$$

[cf. Eq. (38)] where

$$T_k = I - \delta_k \Pi_k - \gamma G_k A_k, \quad g_k = \gamma G_k b_k.$$

The first order approximations of  $T_k$  and  $g_k$  are

$$T_k = T + \delta_k D + O(\delta_k(\Pi_k - \Pi_{\mathbf{N}(A)}) + \delta_k^2 + \|E_k\|), \quad g_k = g + \delta_k d + O(\|E_k\|),$$

[cf. Eq. (39)] where

$$D = -\Pi_{\mathbf{N}(A)}, \quad d = 0.$$

By applying Prop. 6, we have the following convergence result for iteration (50).

**Proposition 8** *Let Assumptions 1, 2, and 3 hold. Let  $\{\delta_k\}$  satisfy the assumptions of Prop. 5, and let  $\Pi_k$  converge to  $\Pi_{\mathbf{N}(A)}$  with probability 1. Then for any  $\gamma$  within a sufficiently small interval  $(0, \bar{\gamma}]$  and all initial iterates  $x_0$ , the sequence  $\{x_k\}$  generated by iteration (50) converges with probability 1 to the solution  $x^*$  of  $Ax = b$  that has minimal Euclidean norm.*

*Proof.* Let  $U, V, N, H$  be the matrices used to construct the decomposition of  $I - \gamma GA$  in Prop. 2, and let  $P$  and  $S$  be the transformation matrices used in the proof of Lemma 3 [cf. Eq. (32)]. Since  $U$  is an orthonormal basis of  $\mathbf{N}(A)$ , we have  $\Pi_{\mathbf{N}(A)} = UU'$  so that

$$P^{-1}\Pi_{\mathbf{N}(A)}P = \left( \begin{bmatrix} I & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & -NH^{-1} \\ 0 & I \end{bmatrix} [U \ V]' \right) UU' \left( [U \ V] \begin{bmatrix} I & NH^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & S \end{bmatrix} \right) = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.$$

By using the line of analysis of Lemma 3, we obtain

$$T + \delta_k D = I - \gamma GA - \delta_k \Pi_{\mathbf{N}(A)} \stackrel{P}{\sim} \begin{bmatrix} I & 0 \\ 0 & I - \gamma \tilde{H} \end{bmatrix} - \delta_k \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} (1 - \delta_k)I & 0 \\ 0 & I - \gamma \tilde{H} \end{bmatrix},$$

where  $\tilde{H} = S^{-1}HS \succ 0$  [cf. Eq. (31)]. Thus by using the matrix norm  $\|\cdot\|_P$  we have for all  $k$

$$\|T + \delta_k D\|_P = \|I - \gamma GA - \delta_k \Pi_{\mathbf{N}(A)}\|_P \leq 1 - \delta_k,$$

and for some  $c > 0$  and all  $k$  sufficiently large,

$$\|T_k\|_P \leq \|T + \delta_k D\|_P + O((\Pi_k - \Pi_{\mathbf{N}(A)})\delta_k + \delta_k^2 + \|E_k\|) \leq 1 - c\delta_k.$$



Therefore the assumptions of Prop. 6 are satisfied, implying that the sequence  $\{x_k\}$  converges to the unique solution of Eq. (41) or Eq. (44) with probability 1.

Finally, let us solve for the limit point of  $\{x_k\}$ . We apply the definition of  $P$  given by Eq. (32),  $D = -\Pi_{\mathbf{N}(A)} = -UU'$  and  $d = 0$  to Eq. (44) [note that  $\bar{Q} = I$  since  $P^{-1}TP$  is already nullspace-decomposed]. This yields

$$0 = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \left( \begin{bmatrix} I & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & -NH^{-1} \\ 0 & I \end{bmatrix} [U \ V]' \right) (-UU'x), \quad Ax = b,$$

or equivalently

$$U'x = 0, \quad V'x = V'x^*.$$

We see that the unique solution to the above system is the minimal Euclidean norm solution.  $\blacksquare$

### 4.3.3 A Stabilization Scheme for the Proximal Iteration

Assume that  $A \succeq 0$ . When simulation is used, neither the iterate sequence  $\{x_k\}$  nor the residual sequence  $\{Ax_k - b\}$  generated by the natural analog of the proximal iteration

$$x_{k+1} = x_k - (A_k + \beta I)^{-1} (A_k x_k - b_k), \quad (51)$$

necessarily converge (for a divergent example we refer to [WaB11]). Moreover the sequence  $\{x_k\}$  generated by the proximal iteration applied to the system  $A'\Sigma^{-1}Ax = A'\Sigma^{-1}b$ , i.e.

$$x_{k+1} = x_k - (A'_k \Sigma^{-1} A_k + \beta I)^{-1} A'_k \Sigma^{-1} (A_k x_k - b_k),$$

also need not converge (see Example 6 in Section 6; on the other hand it is shown in [WaB11] that the residual sequence generated by this iteration does converge to 0). This is remarkable since proximal iterations are used widely for regularization of singular systems.

We may stabilize the proximal iteration (51) by shifting the eigenvalues of  $I - (A_k + \beta I)^{-1} A_k$  by  $-\delta_k$ , as discussed earlier in Section 4.1. However, for the special case of the proximal iteration we may consider an alternative scheme, which shifts instead the eigenvalues of the positive semidefinite matrix  $A$  into the positive half-plane by  $\delta_k$ . It has the form

$$x_{k+1} = x_k - (A_k + \delta_k I + \beta I)^{-1} ((A_k + \delta_k I)x_k - b_k). \quad (52)$$

In this way, we still have  $A_k + \delta_k I \xrightarrow{a.s.} A$ , and assuming that  $\delta_k \downarrow 0$  at a rate sufficiently slow, we will show that the sequence of iterates  $\{x_k\}$  converges with probability 1. Iteration (52) can be written as

$$x_{k+1} = T_k x_k + g_k,$$

where by using the identity  $I - (A_k + \delta_k I + \beta I)^{-1} (A_k + \delta_k I) = \beta (A_k + \delta_k I + \beta I)^{-1}$ , we have

$$T_k = \beta (A_k + \delta_k I + \beta I)^{-1}, \quad g_k = (A_k + \delta_k I + \beta I)^{-1} b_k.$$

The first order approximations of  $T_k$  and  $g_k$  are  $T_k \approx T + \delta_k D$  and  $g_k \approx g + \delta_k d$  as given by Eq. (39), where we can verify that<sup>2</sup>

$$D = -\beta(A + \beta I)^{-2}, \quad d = -(A + \beta I)^{-2}b. \quad (53)$$

---

<sup>2</sup>To see this, note that

$$I - T_k = (A_k + \delta_k I)(A_k + \delta_k I + \beta I)^{-1} = \frac{1}{\beta}(A_k + \delta_k I)T_k = \frac{1}{\beta}A(T + \delta_k D) + \frac{1}{\beta}\delta_k(T + \delta_k D) + O(\delta_k^2 + \|E_k\|),$$

and also note that

$$I - T_k = I - T - \delta_k D + O(\delta_k^2 + \|E_k\|).$$

By combining these two relations, we obtain

$$\frac{\delta_k}{\beta}AD + \frac{\delta_k}{\beta}T = -\delta_k D.$$

Thus the expression for  $D$  is

$$D = -(\beta I + A)^{-1}T = -\beta(A + \beta I)^{-2},$$

where we used  $T = \beta(A + \beta I)^{-1}$ . The expression for  $d$  can be obtained similarly.

We will apply Prop. 6 and show that iteration (52) converges. Moreover, we will show that the limit point, or the unique solution of the system (41), is the solution  $x^*$  with minimum Euclidean norm.

**Proposition 9** *Let Assumptions 2 and 3 hold, and assume that  $A \succeq 0$ ,  $\beta > 0$ . Let  $\{\delta_k\}$  satisfy the assumptions of Prop. 5. Then for all initial iterates  $x_0$ , the sequence  $\{x_k\}$  generated by iteration (52) converges with probability 1 to the solution  $x^*$  of  $Ax = b$  with minimum Euclidean norm.*

*Proof.* Since  $A \succeq 0$ , the proximal iteration with  $G = (A + \beta I)^{-1}$  and  $\beta > 0$  is convergent, implying that Assumption 1 is satisfied (this is the well-known convergence results for weakly monotone variational inequalities; see [Mar70], [Roc76]). We denote

$$G_k = (A_k + \delta_k I + \beta I)^{-1},$$

[cf. Eq. (52)], and we will show that the condition (40) of Prop. 6 is satisfied with  $P = I$ , i.e., there exists  $c > 0$  such that

$$\|I - G_k(A_k + \delta_k I)\| \leq 1 - c\delta_k,$$

for  $k$  sufficiently large with probability 1. Then the convergence will follow from Prop. 6.

Indeed, we have

$$I - G_k(A_k + \delta_k I) = \beta(A_k + \delta_k I + \beta I)^{-1} = \beta(A + \delta_k I + \beta I)^{-1} + \tilde{E}_k, \quad (54)$$

where we define

$$\tilde{E}_k = \beta(A_k + \delta_k I + \beta I)^{-1} - \beta(A + \delta_k I + \beta I)^{-1}.$$

First we consider  $\beta(A + \delta_k I + \beta I)^{-1}$ . We have

$$\begin{aligned} \|\beta(A + \delta_k I + \beta I)^{-1}\| &= \beta \left\| ((A' + \delta_k I + \beta I)(A + \delta_k I + \beta I))^{-1} \right\|^{1/2} \\ &= \beta \left\| (A'A + (\delta_k + \beta)(A' + A) + (\delta_k + \beta)^2 I)^{-1} \right\|^{1/2} \\ &\leq \beta \left\| ((\delta_k + \beta)^2 I)^{-1} \right\|^{1/2} \\ &= \frac{\beta}{\delta_k + \beta}, \end{aligned}$$

where the inequality uses the symmetry and positive semidefiniteness of  $A'A + (\delta_k + \beta)(A' + A)$ , and the fact  $\|(M_1 + M_2)^{-1}\| \leq \|M_1^{-1}\|$  if both  $M_1$  and  $M_2$  are symmetric positive semidefinite matrices.<sup>3</sup> Letting  $c_1 \in \left(0, \frac{1}{\delta_0 + \beta}\right)$  we have for all  $k$  that

$$\|\beta(A + \delta_k I + \beta I)^{-1}\| \leq \frac{\beta}{\delta_k + \beta} < 1 - c_1 \delta_k. \quad (55)$$

Second we consider the matrix  $\tilde{E}_k$ . Let  $f : \mathfrak{R}^{n \times n} \mapsto \mathfrak{R}^{n \times n}$  be the function defined by  $f(\tilde{A}) = (\tilde{A} + \beta I)^{-1}$ , which is Lipschitz continuous within a neighborhood of  $\tilde{A} = A$  [since  $A + \beta I$  is invertible]. Note that  $A_k + \delta_k I \xrightarrow{a.s.} A$  and  $A + \delta_k I \xrightarrow{a.s.} A$ , so we have

$$\|\tilde{E}_k\| = \|f(A_k + \delta_k I) - f(A + \delta_k I)\| = O(\|A_k - A\|) = O(\|E_k\|), \quad w.p.1. \quad (56)$$

By combining Eqs. (54)-(56), we obtain

$$\|I - G_k(A_k + \delta_k I)\| \leq \|\beta(A + \delta_k I + \beta I)^{-1}\| + O(\|E_k\|) \leq 1 - c_1 \delta_k + O(\|E_k\|).$$

---

<sup>3</sup> To prove this, we note that  $x'(M_1 + M_2)x \geq x'M_1x$  for any none-zero  $x \in \mathfrak{R}^n$  and symmetric positive semidefinite matrices  $M_1$  and  $M_2$ , so that  $\|(M_1 + M_2)^{-1}\| \leq \|M_1^{-1}\|$ .

Since  $O(\|E_k\|)/\delta_k \xrightarrow{a.s.} 0$  by Lemma 2, there exists  $c > 0$  such that

$$\|I - G_k(A_k + \delta_k I)\| \leq 1 - c\delta_k,$$

for all  $k$  sufficiently large with probability 1. Therefore we may apply Prop. 6 with  $P = I$  and obtain that  $x_k$  converges with probability 1 to the unique solution of Eq. (41).

Finally, let us verify that the solution to Eq. (41) is the solution  $x^*$  with minimum Euclidean norm. By using Eq. (53), we write the equivalent condition (44) [note that  $P = I$ , and  $\bar{Q} = Q$  since  $Q$  transforms  $P^{-1}TP = T$  into its nullspace decomposition] as

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Q'(A + \beta I)^{-2}(\beta x + b) = 0, \quad Ax = b.$$

We will now verify that the above condition is satisfied with  $x^*$ . The equation  $Ax^* = b$  clearly holds, so we will focus on the first part of the condition. By using the fact  $b = Ax^*$ , we rewrite the matrix involved as

$$\begin{aligned} 0 &= \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Q'(A + \beta I)^{-2}(\beta x^* + b) \\ &= \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Q'(A + \beta I)^{-1}(\beta(A + \beta I)^{-1}x^* + (A + \beta I)^{-1}Ax^*) \\ &= \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Q'(A + \beta I)^{-1}x^* \\ &= \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Q'(A + \beta I)^{-1}QQ'x^*, \end{aligned} \tag{57}$$

where the third equality uses the fact  $\beta(A + \beta I)^{-1} + (A + \beta I)^{-1}A = I$  and the fourth equality uses the fact  $QQ' = I$ . From the nullspace decomposition of  $I - \gamma GA$  given in Prop. 2, we have

$$\beta Q'(A + \beta I)^{-1}Q = Q'(I - GA)Q = \begin{bmatrix} I & -N \\ 0 & I - H \end{bmatrix}, \quad Q'x^* = [U \ V]'x^* = \begin{bmatrix} 0 \\ V'x^* \end{bmatrix}.$$

By using the fact  $\|\beta(A + \beta I)^{-1}\| \leq 1$  and using an analysis similar to that of Lemma 6, we obtain that the above decomposition is block diagonal, i.e.,  $N = 0$ . By combining the above relations with Eq. (57), we obtain

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Q'(A + \beta I)^{-1}QQ'x^* = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I & -0 \\ 0 & I - H \end{bmatrix} \begin{bmatrix} 0 \cdot V'x^* \\ V'x^* \end{bmatrix} = 0.$$

Therefore the system (41) is satisfied with the vector  $x^*$ , so  $x^*$  is the limit point. ■

## 5 Stabilization Schemes for Nearly Singular Systems

In this section we discuss the application of stabilization schemes to problems that are nearly singular (i.e., nonsingular but ill-conditioned, having smallest singular value that is close to 0). It is well known that such problems are highly susceptible to computational error, so it is not surprising that their solution by simulation can be almost as difficult as for their singular counterparts. This motivates the use of stabilization to solve systems that are nonsingular, but nearly singular. For some insight into this idea, we consider the following one-dimensional example and use a constant scalar  $\delta > 0$  instead of the decreasing sequence  $\{\delta_k\}$ .

**Example 3 (A One-Dimensional Problem with Constant  $\delta$ )** Consider the system  $x = \alpha x + b$  where  $\alpha \in (0, 1)$  is a scalar and  $b$  is a positive scalar. Suppose that we estimate  $\alpha$  with  $\alpha_k$  based on  $k$  i.i.d. samples with mean  $\alpha$  and variance  $\sigma^2$ , and estimate  $b$  with  $b_k$  based on  $k$  i.i.d. samples with mean  $b$  and variance  $\bar{\sigma}^2$ . Consider the stabilized iteration

$$x_{k+1} = (1 - \delta)(\alpha_k x_k + b_k),$$

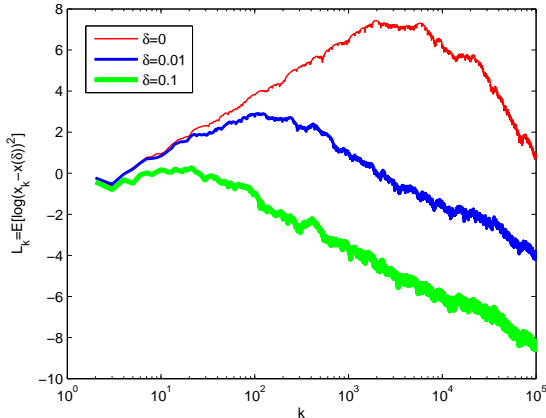


Figure 2: Comparison of stabilized iterations with  $\delta = 0$ ,  $\delta = 0.01$ , and  $\delta = 0.1$  in Example 3. The figure plots the error  $L_k = \mathbf{E} [\log |x_k - x(\delta)|^2]$ , which is relative to the biased limit.

where a constant positive scalar  $\delta$  is used instead of  $\delta_k$  [compare with Eq. (49) in Section 4.3.1]. The iteration converges to the biased limit

$$x(\delta) = \frac{(1 - \delta)b}{1 - (1 - \delta)\alpha}.$$

We let  $\alpha = 0.999$ ,  $b = 1$ ,  $\sigma^2 = 0.01$ ,  $\bar{\sigma}^2 = 1$ , and we consider three cases:  $\delta = 0$ ,  $\delta = 0.01$ , and  $\delta = 0.1$ . For each case, we start with  $x(0) = x(\delta) + 1$ , generate 100 independent trajectories  $\{\alpha_k, b_k, x_k\}$ , and plot the average logarithm of squared errors  $|x_k - x(\delta)|^2$  in Figure 2.

According to Figure 2, in all three cases the errors increase until  $k$  reaches a threshold, and then decrease to  $-\infty$  eventually. However, the errors of the iteration with  $\delta = 0$  increase by roughly four orders of magnitude until  $k$  reaches the order of  $10^4$ . By contrast, the iterations with  $\delta = 0.01$  and  $\delta = 0.1$  have much smaller peak errors, and their thresholds of  $k$  are of the order of  $10^2$  and  $10$  respectively.

For a simplified analysis that is qualitatively consistent with the results of Example 3, we rewrite the iteration as

$$x_{k+1} - x(\delta) = (1 - \delta)\alpha_k(x_k - x(\delta)) + (1 - \delta)((\alpha_k - \alpha)x(\delta) + b_k - b),$$

where  $(1 - \delta)\alpha_k \xrightarrow{a.s.} (1 - \delta)\alpha < 1$  and  $(1 - \delta)((\alpha_k - \alpha)x(\delta) + b_k - b) \xrightarrow{a.s.} 0$ . Intuitively, the multiplicative term  $(1 - \delta)\alpha_k$  has a more significant effect than the additive error does. To emphasize its role, we assume that  $(\alpha_k - \alpha)x(\delta) + b_k - b$  is negligibly small, which is true if the initial condition  $|x_0 - x(\delta)|$  is relatively large. Then the iteration becomes approximately

$$x_{k+1} - x(\delta) \approx (1 - \delta)\alpha_k(x_k - x(\delta)).$$

Defining

$$L_k = \mathbf{E} [\log (x_k - x(\delta))^2],$$

we obtain

$$L_{k+1} \approx L_k + 2 \log(1 - \delta) + \mathbf{E} [\log(\alpha_k^2)] \approx L_k - 2\delta + \mathbf{E} [\log(\alpha_k^2)]. \quad (58)$$

Note that by a first-order Taylor expansion we have  $\log(1 - \delta) \approx -\delta$  since  $\delta$  is close to 0.

We now analyze  $\mathbf{E} [\log(\alpha_k^2)]$ . Note that  $\alpha_k^2$  is a random variable with expectation satisfying

$$\mathbf{E} [\alpha_k^2] = \alpha^2 + \frac{\sigma^2}{k} > 1,$$

when  $k$  is small and  $\sigma^2$  is large. In many simulation contexts and under the same assumptions we often have

$$\mathbf{E} [\log(\alpha_k^2)] \approx \mathbf{E} [\alpha_k^2] - 1 > 0.$$

Indeed, let us suppose that  $\alpha_k$  is a Gaussian random variable with mean  $\alpha$  and variance  $\sigma^2/k$ . By using a result for the expected logarithm of chi-squares (Theorem 3 of [Mos07]), we have

$$\mathbf{E} [\log(\alpha_k^2)] = \log(\alpha^2) + \int_{\frac{k\alpha^2}{\sigma^2}}^{\infty} \frac{e^{-t}}{t} dt.$$

When  $\alpha \approx 1$  and  $\sigma^2/k$  is sufficiently large, we have the approximations

$$\log(\alpha^2) \approx \alpha^2 - 1, \quad \int_{\frac{k\alpha^2}{\sigma^2}}^{\infty} \frac{e^{-t}}{t} dt \approx \Theta\left(\frac{1}{k}\right),$$

so that

$$\mathbf{E} [\log(\alpha_k^2)] \approx \alpha^2 - 1 + \Theta\left(\frac{1}{k}\right) > 0.$$

In this case, Eq. (58) becomes

$$L_{k+1} \approx L_k + \Theta\left(\frac{1}{k}\right) - 2\delta + \alpha^2 - 1.$$

Assume that  $\sigma^2$  is sufficiently large. We analyze the behavior of  $L_k$  for two different values of  $k$ :

- (a) When  $k$  is small, the constant  $-2\delta + \alpha^2 - 1$  is negligible compared to  $\Theta\left(\frac{1}{k}\right)$ . Then we have

$$L_{k+1} \approx L_k + \Theta\left(\frac{1}{k}\right) \approx L_0 + \sum_{t=1}^{k-1} \Theta\left(\frac{1}{t}\right) \approx \Theta(\log k).$$

- (b) When  $k$  is large, the error of the iteration becomes decreasing when

$$k \geq \Theta\left(\frac{1}{1 + 2\delta - \alpha^2}\right). \quad (59)$$

As  $k \rightarrow \infty$ , we have  $L_k \rightarrow -\infty$  and  $x_k \xrightarrow{a.s.} x(\delta)$ .

Comparing the case where  $\delta = 0$  and the case where  $\delta \gg 1 - \alpha^2 \approx 2(1 - \alpha) > 0$ , we have

$$\frac{1}{1 + 2\delta - \alpha^2} \ll \frac{1}{1 - \alpha^2},$$

This suggests that a small  $\delta > 0$  greatly reduces the number of samples needed for the iteration error to decrease, and stabilizes the dramatic rise of the iteration error at the early stage. Our analysis is consistent with the performance of the stabilized iteration as illustrated in Figure 2, and the orders of estimated thresholds of  $k$  as given by Eq. (59) are consistent with those illustrated in Figure 2.

One point missing from the above analysis relates to the bias. In the preceding one-dimensional example, the bias  $|x^* - x(\delta)|$  can be very large. In particular, we have

$$x^* = 1000, \quad x(0.01) \approx 100, \quad x(0.1) \approx 10.$$

To eliminate the bias we must take  $\delta \downarrow 0$ , and the preceding analysis does not provide any conclusive information about this case. We note, however, that when  $\delta$  is kept constant, there may be substantial benefit in the use of stabilization in multi-dimensional problems, as we will demonstrate in the next example.

**Example 4 (A Multi-Dimensional Problem with Constant  $\delta$ )** Consider the stabilized iteration for an  $n \times n$  system:

$$\begin{bmatrix} x_{k+1}(1) \\ \bar{x}_{k+1} \end{bmatrix} = \begin{bmatrix} (1-\delta)\alpha_k & 0 \\ e_k & \bar{A} \end{bmatrix} \begin{bmatrix} x_k(1) \\ \bar{x}_k \end{bmatrix} + \begin{bmatrix} b_k \\ \bar{b} \end{bmatrix},$$

where  $\alpha_k$  is a scalar that converges to  $\alpha \in (0, 1)$  with  $\alpha \approx 1$ ,  $\bar{A}$  is an  $(n-1) \times (n-1)$  matrix with spectral radius strictly smaller than 1 (but not close to 1),  $x_k(1)$  is the first component of the iterate  $x_k \in \mathfrak{R}^n$ ,  $\bar{x}_k \in \mathfrak{R}^{n-1}$  is a subvector consisting of the remaining components of  $x_k$ ,  $b_k \in \mathfrak{R}$  is an estimate of  $b \in \mathfrak{R}$ ,  $\bar{b}$  is a vector in  $\mathfrak{R}^{n-1}$ , and  $e_k \in \mathfrak{R}^{n-1}$  is the simulation error.

To simplify the analysis, we have only applied stabilization to  $x_k(1)$  [this is essentially the selective eigenvalue shifting scheme, cf. Eq. (50)], and have assumed that the vector 0 and matrix  $\bar{A}$  in the iteration does not involve simulation error. It can be seen that the component  $x_k(1)$  evolves according to the same iteration as in the previous one-dimensional example. We will focus on the remaining components  $\bar{x}_k$ , which evolve according to

$$\bar{x}_{k+1} = \bar{A}\bar{x}_k + e_k x_k(1) + \bar{b}.$$

Let  $\bar{x}^* = (\bar{A})^{-1}\bar{b}$  be the corresponding subvector of the fixed point  $x^*$ . Then we have

$$\bar{x}_{k+1} - \bar{x}^* = \bar{A}(\bar{x}_k - \bar{x}^*) + e_k x_k(1).$$

Note that since  $\bar{A}$  has a spectral radius smaller than 1, the above iteration error “tracks” the slow driving term  $(I - \bar{A})^{-1} e_k x_k(1)$ . Thus  $x_k(1)$  may “pollute”  $\bar{x}_k$  with its substantial simulation error if no stabilization is used. As a result when  $k$  is small,  $\|\bar{x}_k - \bar{x}^*\|$  may increase very fast because  $|e_k x_k(1)|$  increases fast. However, if we choose  $\delta > 0$ , by stabilizing  $x_k(1)$ , we will also make  $e_k x_k(1)$  converge to 0 at a faster rate.

Let us test the above analysis on a  $2 \times 2$  example. We let  $\alpha = 0.999$ ,  $\bar{A} = 0.5$ ,  $b = \bar{b} = 1$ . We add noise to each entry of the matrix and vector involved, which is the empirical mean of i.i.d. Gaussian random variables with mean 0 and variance 0.01. We again consider three cases:  $\delta = 0$ ,  $\delta = 0.01$  and  $\delta = 0.1$ . In each case, we start with  $x_0 = x^* + [10 \ 10]'$  and generate 100 independent trajectories. In Figure 3, we plot the average logarithm of the iterate errors  $\|x_k - x^*\|$ , and compare it with the average logarithm of residual errors.

According to Figure 3, the residual is less susceptible to both the simulation error and the stabilization bias. Although the iterate  $x_k$  may have large bias (when  $\delta > 0$ ) or large noise (when  $\delta \approx 0$ ), the residual performs well with a reasonable  $\delta > 0$ : the residual has a small bias and converges fast even for small  $k$ . We have also tested the stabilization scheme that applies  $(1 - \delta)$  to the entire iteration instead of only to  $x_k(1)$ , and obtained results almost identical with the ones of Figure 3. This observation suggests that the  $(1 - \delta)$  stabilization scheme has minor effects on the “well-conditioned” components.

The preceding two examples have illustrated a key beneficial effect of stabilization: in a multi-dimensional ill-conditioned system, it reduces the effects of the noisy “poorly-conditioned” components on the relatively “well-conditioned” components, yielding a good performance in terms of the residual error. If we replace  $\delta > 0$  with a sequence  $\{\delta_k\}$  that decreases to 0 at an appropriate rate, we achieve both a stabilization effect and asymptotic unbiasedness. Additional computational examples will be presented in the next section to justify the use of stabilization in nearly singular problems.

## 6 Computational Illustrations

To illustrate the nature of our convergence results, we will use a  $2 \times 2$  problem for computational experimentation, where

$$A = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad X^* = \{x \mid Ax = b\} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} + \mathbf{R} \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\},$$

and we will artificially add stochastic error to the entries of  $A$  and  $b$ . We will test each of several algorithms with 100 randomly generated iteration trajectories and we will plot, as a function of  $k$ , the corresponding

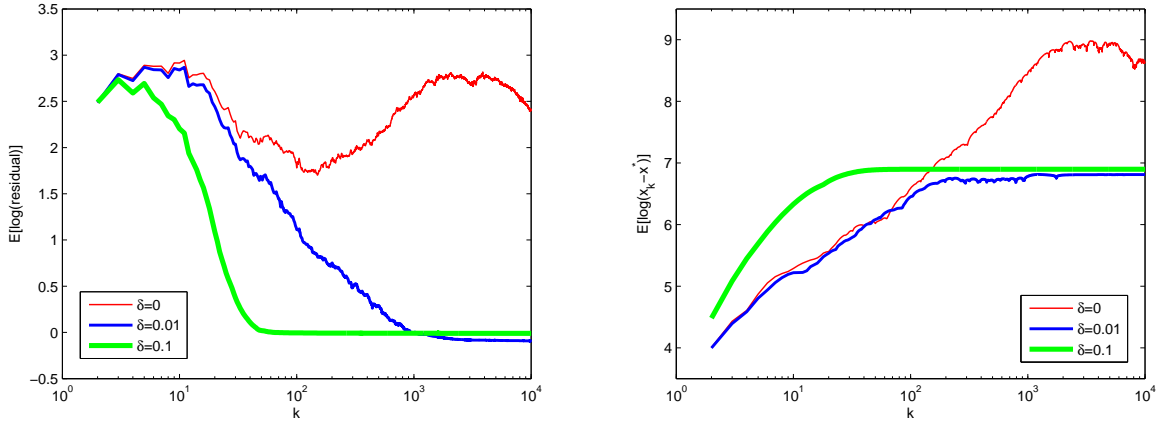


Figure 3: Comparison of stabilized iterations with  $\delta = 0$ ,  $\delta = 0.01$ , and  $\delta = 0.1$  in Example 4. The left figure plots the average logarithm of residual errors, and the right figure plots the average logarithm of iterate errors. Comparing the left and right, we observe that although the iterates may either be very noisy or be severely biased, the residual errors can be stabilized to converge fast with a small bias.

“95% confidence interval” of various quantities of interest, which is the range of the 95 values that are in the “middle” of all 100 values.

**Example 5 (Stabilized Iteration vs. Unstabilized Iteration for a Singular System)** Let  $A_k$  and  $b_k$  be given by

$$A_k = A + \frac{1}{k} \sum_{t=1}^k W_t, \quad b_k = b + \frac{1}{k} \sum_{t=1}^k w_t,$$

where the entries of  $W_t$  and  $w_t$  are i.i.d. Gaussian random variables with mean 0 and variance 0.1. We generate  $\{x_k\}$  with the iteration

$$x_{k+1} = (1 - \delta_k)x_k - \gamma G_k(A_k x_k - b_k),$$

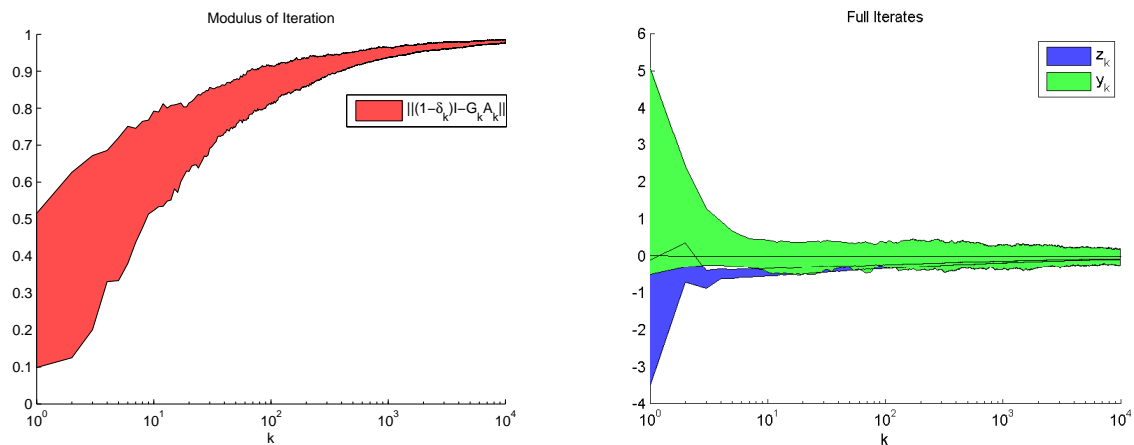
where  $\gamma = 1$ ,  $G_k = I$ , and we consider the two cases (i)  $\delta_k = k^{-1/3}$ , and (ii)  $\delta_k = 0$ . We start with  $x_0 = [10, 10]'$ , and we generate 100 independent trajectories of  $\{A_k, b_k, x_k\}$ .

In Figure 4 we plot the 95% confidence intervals of the spectral radius of the iteration matrix  $(1 - \delta_k)I - G_k A_k$ , and the sequence of components  $\{y_k\}$  and  $\{z_k\}$  respectively. The left side of Figure 4(i) shows that the spectral radius converges to 1, while staying below 1 for  $k$  sufficiently large. The right side of Figure 4(i) shows that both  $\{z_k\}$  and  $\{y_k\}$  are convergent, implying that  $\{x_k\}$  converges to a solution of  $Ax = b$ . For comparison, Figure 4(ii) shows that the unstabilized iteration has spectral radius that is frequently above 1, while  $\{z_k\}$  and  $\{y_k\}$  both diverge, implying that  $\{Ax_k - b\}$  and  $\{x_k\}$  both diverge.

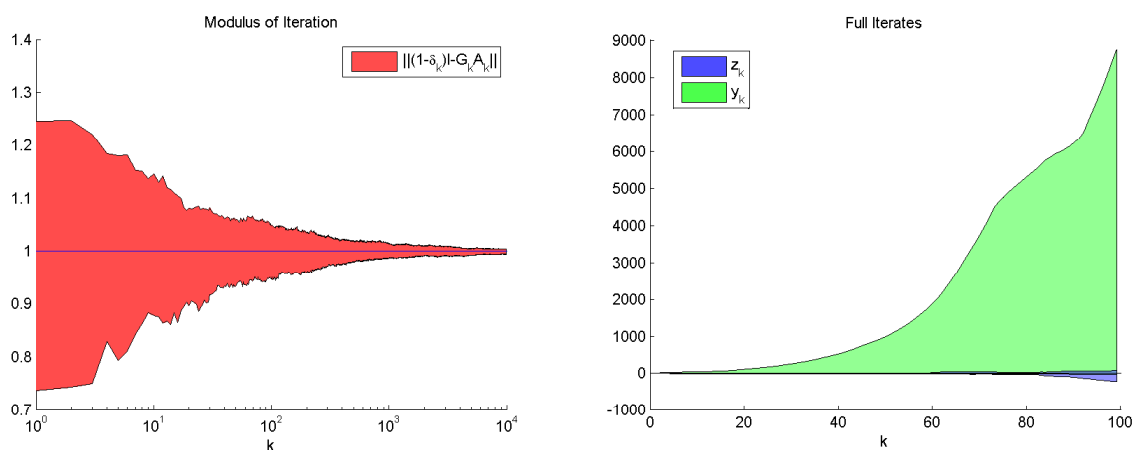
**Example 6 (Comparison of Stabilization Schemes for Proximal Iterations)** Let  $A_k$  and  $b_k$  be given as in Example 5, and consider the following cases of the proximal iteration:

- (i)  $x_{k+1} = (1 - \delta_k)x_k - (A_k + I)^{-1}(A_k x_k - b_k)$  with  $\delta_k = k^{-1/3}$ .
- (ii)  $x_{k+1} = x_k - (\tilde{A}_k + I)^{-1}(\tilde{A}_k x_k - b_k)$  with  $\tilde{A}_k = A_k + \delta_k I$  and  $\delta_k = k^{-1/3}$ .
- (iii)  $x_{k+1} = x_k - (A'_k A_k + I)^{-1}A'_k(A_k x_k - b_k)$ .<sup>4</sup>

<sup>4</sup> This is the proximal algorithm applied to the positive semidefinite symmetric system  $A'Ax = A'b$ .



(i)  $\delta_k = k^{-1/3}$



(ii)  $\delta_k = 0$

Figure 4: Convergence of stabilized iteration with  $\delta_k = k^{-1/3}$  in Example 5 [figure (i)], compared with the unstabilized iteration with  $\delta_k = 0$  [figure (ii)]. The spectral radius of the stabilized iteration converges to 1 from below, and the iterate  $x_k = x^* + Uy_k + Vz_k$  converges to a solution of  $Ax = b$ . For the unstabilized iteration, the spectral radius converges to 1 but crosses frequently above 1, and both  $\{y_k\}$  and  $\{z_k\}$  diverge, implying that the residual and the iterate sequences both diverge.



For each algorithm, we start with  $x_0 = [10, 10]'$ , and we generate 100 independent trajectories of  $\{A_k, b_k, x_k\}$ . The trajectories of “95% confidence intervals” of  $\{z_k\}$  and  $\{y_k\}$  are plotted in Figure 5.

According to Figure 5, the stabilized versions of proximal iteration [cases (i) and (ii)] are convergent and are subject to a decreasing bias. Comparing algorithms (i) and (ii), we notice that (ii) is subject to less bias in  $z_k$ , thus also in its residual. A possible reason is that adding  $\delta_k I$  to  $A$  has a reduced biasing effect on  $V'x_k$ , relative to adding  $-\delta_k I$  to the iteration matrix  $I - \gamma GA$ . This implies that the stabilization approach used in (ii) is preferable to the one used in (i) for this example.

In algorithm (iii), where quadratic regularization without stabilization is used, the component  $z_k$  and thus also the residual seem to be unbiased, but are subject to larger variance. Most importantly, the sequence  $\{y_k\}$  diverges. This is an example of a stochastic iterative method, which generates iterates that diverge and residuals that converge to 0.

In the next two examples, we test numerically the stabilization scheme in nearly singular  $2 \times 2$  systems. The first example involves a diagonal system with a diagonal component particularly small. The second example involves a rotation matrix with an added small multiple of the identity such that it “nearly violates” Assumption 1.

**Example 7 (Stabilized Iteration for a Nearly Singular System I)** Let  $A_k$  and  $b_k$  be given by

$$A_k = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.01 \end{bmatrix} + \frac{1}{k} \sum_{t=1}^k W_t, \quad b_k = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \frac{1}{k} \sum_{t=1}^k w_t,$$

where the entries of  $W_t$  and  $w_t$  are i.i.d. Gaussian random variables with mean 0 and variance 0.1. Let  $\{x_k\}$  be generated by

$$x_{k+1} = (1 - \delta_k)x_k - \gamma G_k(A_k x_k - b_k),$$

with  $\gamma = 1$ ,  $G_k = I$  and consider two cases: (i)  $\delta_k = k^{-1/3}$ ; (ii)  $\delta_k = 0$ . We start with  $x_0 = [10, 10]'$ , and we generate 100 independent trajectories of  $\{A_k, b_k, x_k\}$ . The trajectories of  $\{x_k(1)\}$  and  $\{x_k(2)\}$  are plotted in Figure 6.

According to Figure 6, both iterations (i) and (ii) converge eventually to the unique solution  $x^* = A^{-1}b$ . We notice a threshold effect in the trajectories of iteration (ii) with  $\delta_k = 0$ : the iterates are subject to substantial simulation error when  $k$  is small, and behave well when  $k$  is above a certain value. This phenomenon can be explained by using the analysis of Section 5. By contrast, iteration (i) with  $\delta_k = k^{-1/3}$  has moderate variance for all values of  $k$ .

**Example 8 (Stabilized Iteration for a Nearly Singular System II)** Let  $A_k$  and  $b_k$  be given by

$$A_k = \begin{bmatrix} 10^{-3} & 1 \\ -1 & 10^{-3} \end{bmatrix} + \frac{1}{k} \sum_{t=1}^k W_t, \quad b_k = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \frac{1}{k} \sum_{t=1}^k w_t,$$

where the entries of  $W_t$  and  $w_t$  are i.i.d. Gaussian random variables with mean 0 and variance 0.1. Let  $\{x_k\}$  be generated by

$$x_{k+1} = (1 - \delta_k)x_k - \gamma G_k(A_k x_k - b_k),$$

with  $\gamma = 1$ , and consider three cases: (i)  $\delta_k = k^{-1/3}$  and  $G_k = (A_k + I)^{-1}$ ; (ii)  $\delta_k = 0$  and  $G_k = (A_k + I)^{-1}$ ; (iii)  $\delta_k = 0$  and  $G_k = (A_k' A_k + I)^{-1} A_k'$  [this is the quadratic regularization algorithm as in Example 6(iii)]. We start with  $x_0 = [10, 10]'$ , and we generate 100 independent trajectories of  $\{A_k, b_k, x_k\}$ . The iterate errors are plotted in Figure 7. According to Figure 7, all three iterations are convergent, since the problem is nonsingular, but the stabilized iteration clearly has the smallest variance and bias.

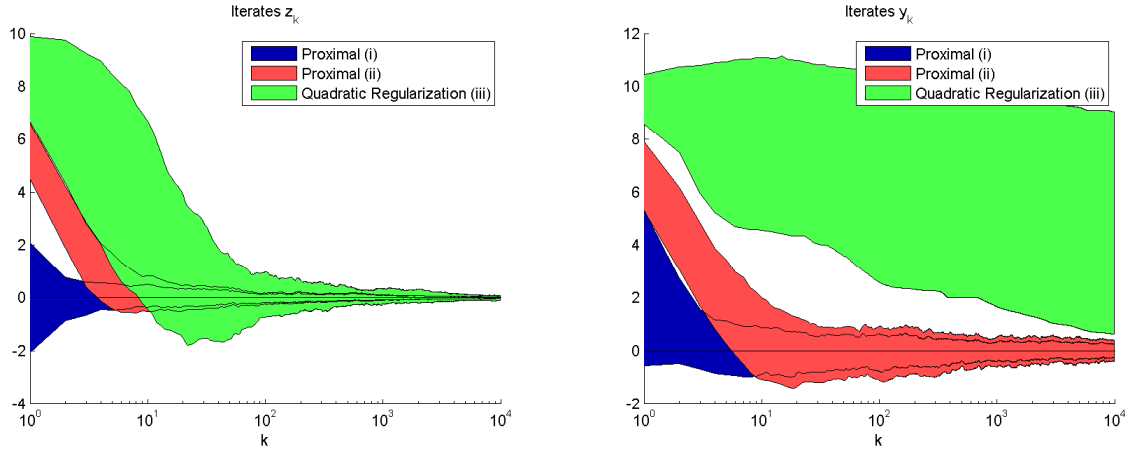


Figure 5: Comparison of the stabilized proximal iterations in Example 6. The left figure plots  $\{z_k\}$ , indicating that all three iterations generate residuals converging to 0. The right figure plots  $\{y_k\}$ , indicating that both versions of stabilized proximal iteration have convergent iterates, while the quadratic regularized iteration has divergent iterates.

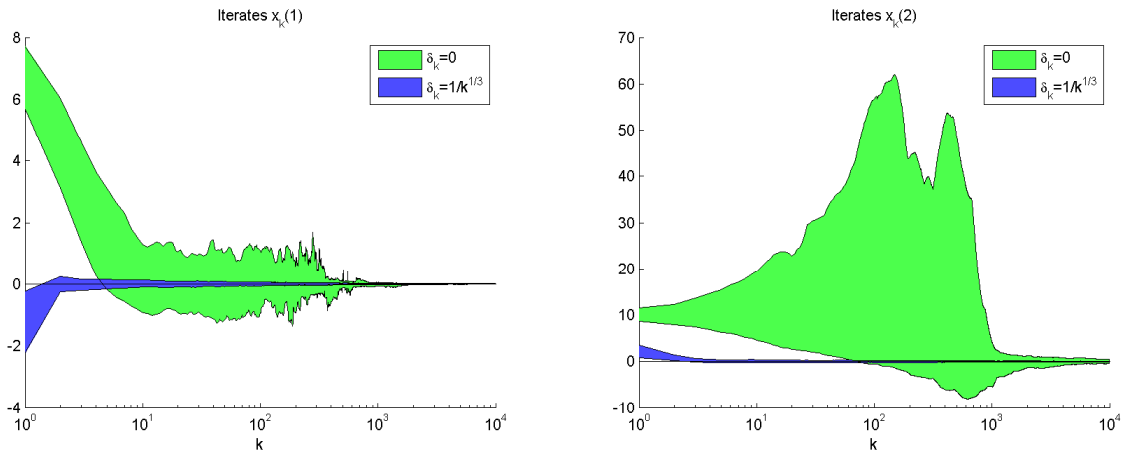


Figure 6: Stabilized iteration for the nearly singular system of Example 7, compared with the unstabilized iteration. The left and right figures plot trajectories of  $x_k(1)$  and  $x_k(2)$  respectively. We notice that  $x_k(2)$  is very noisy without stabilization, and cannot be estimated accurately when  $k$  is small due to the small second diagonal component of the matrix. However, the iterates  $x_k(1)$  obtained using stabilization is less affected by the simulation noise, as opposed to the case without using stabilization.

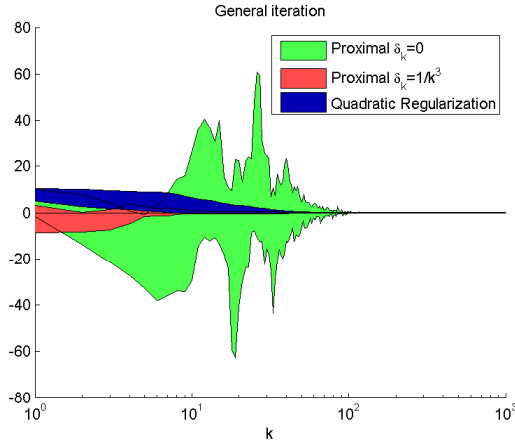


Figure 7: Stabilized iteration for the nearly singular system of Example 8, compared with the unstabilized iteration and quadratic regularization.

## 7 Concluding Remarks

In this paper we have considered the convergence of iterative methods for solving singular linear systems  $Ax = b$ . We have constructed a framework for analysis of deterministic iterative methods, based on an iteration decomposition along the nullspace of  $A$  and its orthogonal subspace, and provided necessary and sufficient conditions for convergence. For the natural stochastic extensions, the residual sequence or the iterate sequence, or both, may not converge due to the fact that stochastic error can accumulate in the nullspace of  $A$ . We have developed new algorithmic modifications that stabilize the methods in the presence of stochastic noise, which ensure that the stochastic iteration converges to a specific solution of  $Ax = b$ . Besides the case where  $A$  is singular, our algorithms seem promising for problems where  $A$  is nonsingular but highly ill-conditioned. This seems to be an interesting area for further research.

### Acknowledgements

We would like to thank the associate editor and the two anonymous referees for their detailed and valuable comments.

## References

- [BeG82] Bertsekas, D. P., and Gafni, E. M., 1982. “Projection Methods for Variational Inequalities with Application to the Traffic Assignment Problem,” *Mathematical Programming Study*, No. 17, pp. 139-159.
- [Ber07] Bertsekas, D. P., 2007. *Dynamic Programming and Optimal Control*, Vol. I and II, Athena Scientific, Belmont, M. A.
- [Ber10] Bertsekas, D. P., 2010. “Approximate Policy Iteration: A Survey and Some New Methods,” LIDS-P-2820, MIT, Cambridge, MA; *J. of Control Theory and Applications*, Vol. 9, pp. 310-335.
- [Ber11] Bertsekas, D. P., 2011. “Temporal Difference Methods for General Projected Equations,” *IEEE Trans. on Automatic Control*, Vol. 56, pp. 2128-2139.
- [Ber99] Bertsekas, D. P., 1999. *Nonlinear Programming*, 2nd edition, Athena Scientific, Belmont, M. A.

- [BeT89] Bertsekas, D. P., and Tsitsiklis J. N., 1989. *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall; republished by Athena Scientific, 1997, Belmont, M. A.
- [BeT96] Bertsekas, D. P., and Tsitsiklis J. N., 1996. *Neuro-Dynamic Programming*, Athena Scientific, Belmont, M. A.
- [BeY09] Bertsekas, D. P., and Yu, H., 2009. “Projected Equation Methods for Approximate Solution of Large Linear Systems,” *J. of Computational and Applied Mathematics*, Vol. 227, pp. 27-50.
- [BMP90] Benveniste, A., Metivier, M., and Priouret, P., 1990. *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, N. Y.
- [Bor08] Borkar, V. S., 2008. *Stochastic Approximation: A Dynamical Systems Viewpoint*, Cambridge University Press, M. A.
- [Boy02] Boyan, J. A., 2002. “Technical Update: Least-Squares Temporal Difference Learning,” *Machine Learning*, Vol. 49, pp. 233-246.
- [BrB96] Bradtke, S. J., and Barto, A. G., 1996. “Linear Least-Squares Algorithms for Temporal Difference Learning,” *Machine Learning*, Vol. 22, pp. 33-57.
- [CaM91] Campbell, S. L., and Meyer, C. D., 1991. *Generalized Inverses of Linear Transformations*, Dover, N. Y.
- [CeH09] Censor, J., Herman, G. T., and Jiang, M., 2009. “A Note on the Behavior of the Randomized Kaczmarz Algorithm of Strohmer and Vershynin,” *J. Fourier Analysis and Applications*, Vol. 15, pp. 431-436.
- [CPS92] Cottle, R. W., Pang, J. S., and Stone, R. E., 1992. *The Linear Complementarity Problem*, Academic Press, Boston, M. A.
- [Cur54] Curtiss, J. H., 1954. “A Theoretical Comparison of the Efficiencies of Two Classical Methods and A Monte Carlo Method for Computing One Component of the Solution of A Set of Linear Algebraic Equations,” *Proc. Symposium on Monte Carlo Methods*, pp. 191-233.
- [Cur57] Curtiss, J. H., 1957. “A Monte Carlo Methods for the Iteration of Linear Operators,” *Uspekhi Mat. Nauk*, Vol. 12, pp. 149-174.
- [Cve08] Cvetkovic-llic, D. S., 2008. “A Note on the Representation for the Drazin Inverse of  $2 \times 2$  Block Matrices,” *Linear Algebra and its Applications*, Vol. 429, Issue 1, pp. 242-248.
- [Dax90] Dax, A., 1990. “The Convergence of Linear Stationary Iterative Processes for Solving Singular Unstructured Systems of Linear Equations,” *SIAM Review*, Vol. 32, pp. 611-635.
- [DKM06a] Drineas, P., Kannan, R., and Mahoney, M. W., 2006. “Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication,” *SIAM J. Computing*, Vol. 35, pp. 132-157.
- [DKM06b] Drineas, P., Kannan, R., and Mahoney, M. W., 2006. “Fast Monte Calo Algorithms for Matrices II: Computing a Low-Rank Approximation to a Matrix,” *SIAM J. Computing*, Vol. 36, pp. 158-183.
- [DMM06] Drineas, P., Mahoney, M. W., and Muthukrishnan, S., 2006. “Sampling Algorithms for L2 Regression and Applications,” *Proc. 17th Annual SODA*, pp. 1127-1136.
- [DMM08] Drineas, P., Mahoney, M. W., and Muthukrishnan, S., 2008. “Relative-Error CVR Matrix Decompositions,” *SIMAX*, pp. 844-881.
- [DMMS11] Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlos, T., 2011. “Faster Least Squares Approximation,” *Numerische Mathematik*, Vol. 117, pp. 219-249.

- [EMN88] Eiermann, M., Marek, I., and Niethammer, W., 1988. "On the Solution of Singular Linear Systems of Algebraic Equations by Semiiterative Methods," *Numerische Mathematik*, Vol. 53, pp. 265-283.
- [FaP03] Facchinei, F., and Pang, J. S., 2003. *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer, N. Y.
- [FoL50] Forsythe, G. E., and Leibler, R. A., 1950. "Matrix Inversion By A Monte Carlo Method," *Mathematical Tables and Other Aids to Computation*, Vol. 4, pp. 127-129.
- [Hal70] Halton, J. H., 1970. "A Retrospective and Prospective Survey of the Monte Carlo Method," *SIAM Review*, Vol. 12, pp. 1-63.
- [HaY81] Hageman, L. A., and Young, D. M., 1981. *Applied Iterative Methods*, Academic Press, N. Y.
- [Kel65] Keller, H. B., 1965. "On the Solution of Singular and Semidefinite Linear Systems by Iteration," *J. SIAM: Series B, Numerical Analysis*, Vol. 2, pp. 281-290.
- [Kor76] Korpelevich, G. M., 1976. "An Extragradient Method for Finding Saddle Points and For Other Problems," *Matecon*, Vol. 12, no. 4, pp. 747-756.
- [Kra72] Krasnoselskii, M. A., 1972. *Approximate Solution of Operator Equations*, translated by Louvish, D. Wolters-Noordhoff Pub., Groningen.
- [KuY03] Kushner, H. J., and Yin, G., 2003. *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, N. Y.
- [LaT85] Lancaster, P., and Tismenetsky, M., 1985. *The Theory of Matrices*, Academic Press, N. Y.
- [LeL10] Leventhal, D., and Lewis, A. S., 2010. "Randomized Methods for Linear Constraints: Convergence Rates and Conditioning," *Mathematics of Operations Research*, Vol. 35, pp. 641-654.
- [LuT89] Luo, Z. Q., and Tseng P., 1989. "On the Convergence of a Matrix Splitting Algorithm for the Symmetric Monotone Linear Complementarity Problem," *SIAM J. Control Optim.*, Vol. 29, pp. 1037-1060.
- [Mar70] Martinet, B., 1970. "Regularisation D'Inequation Variationelles par Approximations Successives," *Rev. Francaise Inf. Rech. Oper.*, pp. 154-159.
- [Mey07] Meyn, S. P., 2007. *Control Techniques for Complex Systems*, Cambridge University Press, M. A.
- [Mos07] Moser, S. M., 2007. "Some Expectations of a Non-central Chi-square Distribution with an Even Number of Degrees of Freedom," *TENCON, IEEE Region 10 Conference*.
- [NeB03] Nedic, A., and Bertsekas, D. P., 2003. "Least-Squares Policy Evaluation Algorithms with Linear Function Approximation," *J. of Discrete Event Systems*, Vol. 13, pp. 79-110.
- [OrR70] Ortega, J. M., and Rheinboldt, W. C., 1970. *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, N. Y.
- [Put94] Puterman, M. L., 1994. *Markovian Decision Problems*, J. Wiley, N. Y.
- [Roc76] Rockafellar, R. T., 1976. "Monotone Operators and the Proximal Point Algorithm," *SIAM J. on Control and Optimization*, Vol. 14, pp. 877-898.
- [Sib70] Sibony M., 1970. "Methodes Iteratives Pour les Equations et Inequations aux Derivees Partielles Non Lineaires de Type Monotone," *Calcolo*, Vol. 7, pp. 65-183.
- [Ste73] Stewart, G. W., 1973. *Introduction to Matrix Computations*, Academic Press, N. Y.

- [StV09] Strohmer, T., and Vershynin, R., 2009. “A Randomized Kaczmarz Algorithm with Exponential Convergence”, *J. of Fourier Analysis and Applications*, Vol. 15, pp. 262-178.
- [SuB98] Sutton, R. S., and Barto, A. G., 1998. *Reinforcement Learning*, MIT Press, Cambridge, M. A.
- [Tan74] Tanabe, K., 1974. “Characterization of Linear Stationary Iterative Processes for Solving a Singular System of Linear Equations,” *Numerische Mathematik*, Vol. 22, pp. 349-359.
- [WaB11] Wang, M., and Bertsekas, D. P., 2011. “On the Convergence of Simulation-Based Iterative Methods for Singular Linear Systems,” *Lab. for Information and Decision Systems Report LIDS-P-2879*, MIT, Cambridge, M. A.
- [Was52] Wasow, W. R., 1952. “A Note on Inversion of Matrices by Random Walks,” *Mathematical Tables and Other Aids to Computation*, Vol. 6, pp. 78-81.
- [You72] Young, D. M., 1972. “On the Consistency of Linear Stationary Iterative Methods,” *J. SIAM on Numerical Analysis*, Vol. 9, pp. 89-96.