On Iterative Solution of the Extended Normal Equations

Henri Calandra, Serge Gratton, Elisa Riccietti, Xavier Vasseur

ON ITERATIVE SOLUTION OF THE EXTENDED NORMAL EQUATIONS*

HENRI CALANDRA†, SERGE GRATTON‡, ELISA RICCIETTI‡, AND XAVIER VASSEUR§

Abstract. Given a full-rank matrix $A \in \mathbb{R}^{m\times n}$ ($m \geq n$), we consider a special class of linear systems $A^TAx = A^Tb + c$ with $x, c \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, which we refer to as the extended normal equations. The occurrence of c gives rise to a problem with a different conditioning from the standard normal equations and prevents direct application of standard methods for least squares. Hence, we seek more insights on theoretical and practical aspects of the solution of such problems. We propose an explicit formula for the structured condition number, which allows us to compute a more accurate estimate of the forward error than the standard one used for generic linear systems, which does not take into account the structure of the perturbations. The relevance of our estimate is shown on a set of synthetic test problems. Then, we propose a new iterative solution method that, as in the case of normal equations, takes advantage of the structure of the system to avoid unstable computations such as forming A^TA explicitly. Numerical experiments highlight the increased robustness and accuracy of the proposed method compared to standard iterative methods. It is also found that the new method can compare to standard direct methods in terms of solution accuracy.

Key words. linear systems, conjugate gradient method, forward error, least squares problems

AMS subject classifications. 15A06, 65F10, 65F35, 65G50

DOI. 10.1137/19M1288644

1. Introduction. Given $A \in \mathbb{R}^{m\times n}$, $m \geq n$, with $\operatorname{rank}(A) = n$, $b \in \mathbb{R}^m$, and $x, c \in \mathbb{R}^n$, we consider the *extended least squares problem*

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 - c^Tx, \tag{ELS}$$

whose solution satisfies the *extended normal equations*

$$A^TAx = A^Tb + c \tag{ENE}$$

or, equivalently, what is often known as the augmented system

$$\begin{bmatrix} \xi I_m & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} y \\ x \end{bmatrix} = \begin{bmatrix} b \\ -c/\xi \end{bmatrix}, \quad r = \xi y = b - Ax, \tag{1.1}$$

where ξ is a scaling parameter that can be chosen to minimize the condition number of the augmented matrix. The optimal value of the scaling parameter is $\xi^* = \frac{1}{\sqrt{2}}\sigma_{\min}(A)$ [3]. The vectors x and r are uniquely determined and independent of ξ. Equation

*Received by the editors September 20, 2019; accepted for publication (in revised form) by M. Tůma July 16, 2020; published electronically October 8, 2020.

https://doi.org/10.1137/19M1288644

Funding: The work of the authors was partially supported by the 3IA Artificial and Natural Intelligence Toulouse Institute, French "Investing for the Future - PIA3" program under grant ANR-19-PI3A-0004 and by the TOTAL R&D under grant DS3416.

†TOTAL, Centre Scientifique et Technique Jean Féger, F-64000 Pau, France (Henri.Calandra@total.com).

‡INPT-IRIT, University of Toulouse and ENSEEIHT, F-31071 Toulouse Cedex 7, France (serge.gratton@enseeiht.fr, elisa.riccietti@enseeiht.fr).

§ISAE-SUPAERO, University of Toulouse, F-31055 Toulouse Cedex 4, France (xavier.vasseur@isae-supaero.fr).

(1.1) also gives the first-order optimality conditions for the problems

$$\text{(ELS-primal)} \qquad \min_{x,r} \frac{1}{2}\|r\|^2 - c^T x \quad \text{subject to} \quad Ax + r = b$$

and

$$\text{(ELS-dual)} \qquad \min_{r} \frac{1}{2}\|r\|^2 - b^T r \quad \text{subject to} \quad A^T r = -c.$$

Using the QR factorization $[A\ b] = Q\left[\begin{smallmatrix} R & d_1 \\ 0 & d_2 \end{smallmatrix}\right]$ with $R \in \mathbb{R}^{n\times n}$, we could obtain the solution from

$$x = A^\dagger b + A^\dagger (A^\dagger)^T c = R^{-1} d_1 + (R^T R)^{-1} c.$$

At first sight (ENE) evokes a least squares problem in the normal equations form

$$A^T A x = A^T b,$$

but the vector c makes the situation fundamentally different. Unlike pure least squares problems, a very well-studied topic [4], [19, Chapter 5], [25, Chapter 10], problem (ENE) has not been the object of much study in the literature. We mention Björck [3], who studies the numerical solution of the augmented system by Gaussian elimination.

The solution of problem (ENE) is required in various applications in optimization, such as the multilevel Levenberg–Marquardt methods [8], or certain formulations based on penalty function approaches [13], [14, section 7.2], [15], which we describe in section 2. Motivated by these important applications, we seek more insight on theoretical and practical aspects of the numerical solution of (ENE). Having in mind the possibility of seeking an approximate solution, we especially focus on iterative methods. We expect to encounter issues similar to those reported in the literature on normal equations.

First, it is well known that the product $A^T A$ should not be explicitly formed, because the accuracy attainable by methods for the solution of the normal equations may be much lower than for a backward stable method for least squares. If the matrix is formed, the best forward error bound for the normal equations can be obtained by the classical sensitivity analysis of linear systems. It is of order $\kappa^2(A)\ \epsilon$ with ϵ the machine precision and $\kappa(A) = \|A\|\|A^\dagger\|$ the condition number of A in the Euclidean norm [23, 28]. This is an underwhelming result, as from Wedin's theorem [23, Theorem 20.1] the sensitivity of a least squares problem is measured by $\kappa^2(A)$ only when the residual is large, and by $\kappa(A)$ otherwise.

However, practical solution methods do not form the product $A^T A$, and the following key observation is rather exploited. The special structure of the normal equations allows us to write

$$A^T A x - A^T b = A^T (Ax - b),$$

which makes it possible to either employ a factorization of A rather than of $A^T A$ (in the case of direct methods) or to perform matrix-vector multiplications of the form Ax and $A^T y$ rather than $A^T A x$ (in the case of iterative methods). The standard analysis of linear systems is unsuitable for predicting the error for such methods, as it is based on the assumption that the linear system is subject to normwise perturbations on $A^T A$. If the product is not formed explicitly, such global perturbations are not generated in finite precision and a condition number useful for predicting the error should rather take into account structured perturbations, such as perturbations in

the matrix A only. A structured analysis is then more relevant and indeed leads to the same conclusions as for least squares problems [20].

It is therefore possible to devise stable implementations of methods for normal equations. In [22, section 10], Hestenes and Stiefel propose a specialized implementation of the conjugate gradient (CG) method for the normal equations, now known as CGLS. This has been deeply investigated in the literature [5, 19, 22, 27] and CGLS was shown to be more stable than CG applied directly to the standard normal equation with $c = 0$. However, not all the considerations made for normal equations apply to (ENE). The presence of c has important theoretical and practical consequences.

On one hand, c results in a different mapping for the condition number and a different set of admissible perturbations for the backward error. Consequently, the existing perturbation theory for least squares problems [1, 20, 29] does not apply. A proper analysis of the structured condition number of the problem should be developed.

From a practical perspective, even though the system matrix is the same, the presence of c in the right-hand side prevents direct application of standard methods for the normal equations. Successful algorithmic procedures used for normal equations can, however, be tailored to obtain stable solution methods.

Contributions. We propose CGLSc, a modification of standard CGLS that results in a stable method for the solution of (ENE). We provide an expression of the structured condition number for (ENE), which allows us to compute first-order estimates of the forward error in the computed solution. We report on the numerical performance of the method on a relevant set of test problems. The experimentation confirms improved stability of the proposed method compared to standard iterative methods such as CG and MINRES [26], and it is shown to provide solutions almost as accurate as those obtained by stable direct methods. The estimate of the forward error is also validated numerically and shown to be a sharper upper bound than the standard bound from the theory of linear systems.

Structure. In section 2 we present two applications arising in optimization where the solution of (ENE) is required. In section 3, we report on the conditioning of the problem and backward error analysis. These results are employed to propose a first-order estimate of the forward error in the solution computed by a method that does not form matrix A^TA. In section 4 we introduce a new stable iterative method for the solution of (ENE). Extensive numerical experiments are described in section 5. Conclusions are drawn in section 6.

Notation. Given a matrix $A \in \mathbb{R}^{m\times n}$, we denote by $A^\dagger$ its pseudoinverse, by $\kappa(A) = \|A\|\|A^\dagger\|$ its condition number ($\|\cdot\|$ being the Euclidean norm), and by $\sigma_{\min}(A), \sigma_{\max}(A)$ its smallest and largest singular values. The Frobenius norm is defined as

$$\|A\|_F := \left(\sum_{i,j} |A_{i,j}|^2\right)^{1/2} = tr(A^TA)^{1/2}.$$

Given $b \in \mathbb{R}^m, c \in \mathbb{R}^n$, and $\alpha, \beta, \gamma > 0$, we define the following parameterized Frobenius norm

$$\|(A, b, c)\|_{F(\alpha,\beta,\gamma)} := \left\| \begin{bmatrix} \alpha A & \beta b \\ \gamma c^T & 0 \end{bmatrix} \right\|_F = \sqrt{\alpha^2\|A\|_F^2 + \beta^2\|b\|^2 + \gamma^2\|c\|^2}.$$

We denote by $I_n \in \mathbb{R}^{n\times n}$ the identity matrix of order n, by $\otimes$ the Kronecker product of two matrices, and by vec the operator that stacks the columns of a matrix into a vector of appropriate dimension [20].

2. Two motivating applications. We describe two different applications in which problems of the form (ELS) arise, motivating our interest in their solution.

2.1. Fletcher's exact penalty function approach. The first applicative context arises in equality constrained minimization. Consider a problem of the form

$$\min_x f(x) \quad \text{s.t.} \quad g(x) = 0$$

for twice differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$. The solution of systems of the form (ENE) is needed to evaluate the following penalty function and its gradient [15], [14, section 7.2]:

$$\Phi_\lambda(x) = f(x) - g(x)^T y_\lambda(x),$$

where $y_\lambda(x) \in \mathbb{R}^m$ is defined as the solution of the minimization problem

$$\min_y \|A(x)^T y - \nabla f(x)\|^2 + \lambda g(x)^T y$$

with $A(x)$ the Jacobian matrix of $g(x)$ at x, and $\lambda > 0$ a given real-valued penalty parameter.

2.2. Multilevel Levenberg–Marquardt method. The solution of (ENE) is required in multilevel Levenberg–Marquardt methods, which are specific members of the family of multilevel optimization methods recently introduced in Calandra et al. [8] and further analysed in Calandra et al. [7]. The multilevel Levenberg–Marquardt method is intended to solve nonlinear least squares problems of the form

$$\min_x f(x) = \frac{1}{2}\|F(x)\|^2$$

with $F : \mathbb{R}^n \to \mathbb{R}^m$ a twice continuously differentiable function. In the two-level setting, the multilevel method allows two different models to compute the step at each iteration: the classical Taylor model or a cheaper model m_k^H, built from a given approximation $f^H(x^H) = \frac{1}{2}\|F^H(x^H)\|^2$ to the objective function

$$m_k^H(x_k^H, s^H) = \frac{1}{2}\|J^H(x_k^H)s^H + F^H(x_k^H)\|^2 + \frac{\lambda_k}{2}\|s^H\|^2 + (R\nabla f(x_k) - \nabla f^H(x_0^H))^T s^H$$

with $J^H(x_k^H)$ the Jacobian matrix of F^H at x_k^H, $\lambda_k > 0$ a real-valued regularization parameter, R a full-rank linear restriction operator, and $x_0^H = Rx_k$ with x_k denoting the current iterate at a fine level.

While minimizing the Taylor model amounts to solving a formulation based on the normal equations, minimizing m_k^H requires the solution of a problem of the form (ELS) because of the correction term $(R\nabla f(x_k) - \nabla f^H(x_0^H))^T s^H$ needed to ensure coherence between levels. Approximate minimization of the model is sufficient to guarantee convergence of the method. Thus if the coarse problem is not small, an iterative method is well suited for minimizing the model.

3. Conditioning of the problem and backward error analysis. We first propose an explicit formula for the structured condition number of problem (ENE). This is useful to compute a first-order estimate of the forward error for methods that do not form matrix A^TA explicitly. Contrary to the classical sensitivity analysis of linear systems, which is based on the assumption that the linear system is subject to normwise perturbations on the matrix A^TA, our result indeed considers perturbations on matrix A only. We also propose theoretical results related to the backward error.

3.1. Conditioning of the problem. The conditioning of problem (ENE) is the sensitivity of the solution x to perturbations in the data A, b, c. We give an explicit formula for the structured condition number for perturbations on all of A, b, and c. In the following, we define the *condition number* of a function; see [28].

DEFINITION 3.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed vector spaces. If F is a continuously differentiable function*

$$F:\ \mathcal{X} \to \mathcal{Y}, \quad x \longmapsto F(x),$$

the absolute condition number of F at x is the scalar $\|F'(x)\| := \sup_{\|v\|_{\mathcal{X}}=1} \|F'(x)v\|_{\mathcal{Y}}$, where $F'(x)$ is the Fréchet derivative of F at x. The relative condition number of F at x is

$$\frac{\|F'(x)\|\,\|x\|_{\mathcal{X}}}{\|F(x)\|_{\mathcal{Y}}}.$$

We consider F as the function that maps A, b, c to the solution x of (ENE),

$$F : \mathbb{R}^{m\times n} \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n,$$
$$(A, b, c) \longmapsto F(A, b, c) = A^\dagger b + A^\dagger (A^\dagger)^T c.$$

The Fréchet derivative in finite-dimensional spaces is the usual derivative. In particular, it is represented in coordinates by the Jacobian matrix. If F is Fréchet differentiable at a point (A, b, c), then its derivative is

$$F'(A, b, c) : \mathbb{R}^{m\times n} \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n,$$
$$F'(A, b, c)(E, f, g) = J_F(A, b, c)(E, f, g),$$

where $J_F(A, b, c)(E, f, g)$ denotes the Jacobian matrix of F at (A, b, c) applied to (E, f, g). As in [20], we choose the Euclidean norm for the solution and the parameterized Frobenius norm for the data (as introduced in section 1). According to Definition 3.1 (cf. also [18]), the absolute condition number of F at the point (A, b, c) is given by

$$\|F'(A, b, c)\| = \sup_{\|(E,f,g)\|_{F(\alpha,\beta,\gamma)}=1} \|F'(A, b, c)(E, f, g)\|, \quad E \in \mathbb{R}^{m\times n}, f \in \mathbb{R}^m, g \in \mathbb{R}^n.$$

The parameterized Frobenius norm has been chosen for its flexibility. For instance, taking large values of γ allows us to perturb A and b only, and to include the case $c = 0$. This is because the condition $\gamma \to \infty$ implies $g \to 0$ from the constraint $\alpha^2\|E\|_F^2 + \beta^2\|f\|^2 + \gamma^2\|g\|^2 = 1$ in the definition of the condition number.

Let A be perturbed to $\tilde{A} = A + E$, the vector b to $\tilde{b} = b + f$, and vector c to $\tilde{c} = c + g$. A^TA is then perturbed to

$$\tilde{A}^T\tilde{A} = (A + E)^T(A + E) = A^TA + A^TE + E^TA, \tag{3.1}$$

neglecting the second-order terms. The solution $x = (A^TA)^{-1}(A^Tb + c)$ is then perturbed to $\tilde{x} = x + \delta x = (\tilde{A}^T\tilde{A})^{-1}(\tilde{A}^T\tilde{b} + \tilde{c})$. Then, $\tilde{x}$ solves

$$(A^TA + A^TE + E^TA)\tilde{x} = (A^Tb + E^Tb + A^Tf + c + g).$$

Recalling that for A of full column rank, $A^\dagger = (A^TA)^{-1}A^T$ and $(A^TA)^{-1} = A^\dagger(A^\dagger)^T$, we have $\delta x = (A^TA)^{-1}E^Tr - A^\dagger Ex + A^\dagger f + A^\dagger(A^\dagger)^T g$. We conclude that

$$F'(A, b, c)(E, f, g) = (A^TA)^{-1}E^Tr - A^\dagger Ex + A^\dagger f + A^\dagger(A^\dagger)^T g$$

for all $E \in \mathbb{R}^{m\times n}, f \in \mathbb{R}^m, g \in \mathbb{R}^n$, and $r = b - Ax$. We then deduce the following property.

LEMMA 3.2. *The conditioning of problem* (ENE) *with Euclidean norm on the solution and Frobenius norm (parameterized by α, β, γ) on the data, is given by*

$$\|F'(A,b,c)\| = \|[((r^T \otimes (A^TA)^{-1})L_T - x^T \otimes A^\dagger)/\alpha, A^\dagger/\beta, (A^TA)^{-1}/\gamma]\|, \tag{3.2}$$

where L_T is a permutation matrix consisting of ones in positions $(n(k-1)+l, m(l-1)+k)$ with $l = 1, \ldots, n$ and $k = 1, \ldots, m$ and of zeros elsewhere [16, 20].

The following theorem gives an explicit and computable formula for the structured condition number.

THEOREM 3.3. *The absolute condition number of problem* (ENE) *with Euclidean norm on the solution and Frobenius norm (parameterized by α, β, γ) on the data, is $\sqrt{\|\bar{M}\|}$ with $\bar{M} \in \mathbb{R}^{n\times n}$ given by*

$$\bar{M} = \left(\frac{1}{\gamma^2} + \frac{\|r\|^2}{\alpha^2}\right)(A^TA)^{-2} + \left(\frac{1}{\beta^2} + \frac{\|x\|^2}{\alpha^2}\right)(A^TA)^{-1} - \frac{2}{\alpha^2}\,\mathrm{sym}(B) \tag{3.3}$$

with $B = A^\dagger r x^T (A^TA)^{-1}$, $\mathrm{sym}(B) = \frac{1}{2}(B + B^T)$, and x the exact solution of (ENE).

Proof. Let us define $M = [((r^T \otimes (A^TA)^{-1})L_T - x^T \otimes A^\dagger)/\alpha, A^\dagger/\beta, (A^TA)^{-1}/\gamma]$. We recall that

$$\|F'(A,b,c)\| = \|M\| = \|M^T\| := \sup_{y\neq 0} \frac{\|M^T y\|}{\|y\|}. \tag{3.4}$$

Let us consider

$$\begin{aligned} y^T F'(A,b,c)(E,f,g) &= y^T (A^TA)^{-1} E^T r - y^T A^\dagger E x + y^T A^\dagger f + y^T A^\dagger (A^\dagger)^T g \\ &= r^T E (A^TA)^{-1} y - y^T A^\dagger E x + y^T A^\dagger f + y^T A^\dagger (A^\dagger)^T g. \end{aligned}$$

We recall that $E = \sum_{i=1}^n \sum_{j=1}^m e_i^T E e_j$ for e_i, e_j, the ith and jth vectors of the canonical basis. Then, we can rewrite the expression as

$$[\mathrm{vec}(S)/\alpha, y^T A^\dagger/\beta, y^T A^\dagger (A^\dagger)^T/\gamma] \cdot [\alpha \mathrm{vec}(E), \beta f, \gamma g]^T := w^T [\alpha \mathrm{vec}(E), \beta f, \gamma g]^T,$$

introducing the matrix S such that

$$S_{i,j} = r^T e_i e_j^T (A^TA)^{-1} y - y^T A^\dagger e_i e_j^T x.$$

It follows that $w = M^T y$. We are then interested in the norm of w. We can compute the squared norm of $\mathrm{vec}(S)$ as

$$\begin{aligned} \|\mathrm{vec}(S)\|^2 &= \sum_{i=1}^n \sum_{j=1}^m (r^T e_i e_j^T (A^TA)^{-1} y - y^T A^\dagger e_i e_j^T x)^2 \\ &= \|r\|^2 \|(A^TA)^{-1} y\|^2 + \|x\|^2 \|(A^\dagger)^T y\|^2 \\ &\quad - y^T (A^\dagger r x^T (A^TA)^{-1} + (A^TA)^{-1} x r^T (A^\dagger)^T) y. \end{aligned}$$

Then,

$$\begin{aligned} \|w\|^2 &= y^T \bar{M} y, \\ \bar{M} &:= \left(\frac{1}{\gamma^2} + \frac{\|r\|^2}{\alpha^2}\right)(A^TA)^{-2} + \left(\frac{1}{\beta^2} + \frac{\|x\|^2}{\alpha^2}\right) A^\dagger (A^\dagger)^T - \frac{1}{\alpha^2}(B + B^T), \\ B &:= A^\dagger r x^T (A^TA)^{-1}. \end{aligned}$$

From (3.4),

$$\|F'(A,b,c)\| = \|M\| = \sup_{y\neq 0} \frac{\|M^T y\|}{\|y\|} = \sup_{y\neq 0} \frac{\sqrt{y^T \bar{M} y}}{\|y\|} = \sqrt{\|\bar{M}\|}. \qquad \square$$

We remind the analogous result for the least squares case. If we define

$$F_{LS} : \mathbb{R}^{m\times n} \times \mathbb{R}^m \to \mathbb{R}^n,$$
$$(A,b) \longmapsto F_{LS}(A,b) = A^\dagger b,$$

the absolute condition number for least squares (or structured conditioning of the normal equations) is [20]

$$\|F'(A,b)\| = \|A^\dagger\| \sqrt{\frac{1}{\beta^2} + \frac{\|x\|^2 + \|A^\dagger\|^2 \|r\|^2}{\alpha^2}}.$$

The term $\|A^\dagger\|^2$ appears here multiplied by the norm of the residual. This is interpreted as saying that the sensitivity of the problem depends on $\kappa(A)$ for small or zero residual problems and on $\kappa^2(A)$ for all other problems [4, 23]. If $c = 0$ and $\gamma \to \infty$, the known result for least squares problems is recovered (note that in this case $B = 0$ as $A^T r = 0$).

Let us assume that $(\alpha, \beta, \gamma) = (1, 1, 1)$. We define the structured relative condition number

$$\kappa_S = \frac{\sqrt{\|\bar{M}\|}\, \|A,b,c\|_F}{\|x\|} = \frac{\|M\|\, \|A,b,c\|_F}{\|x\|}, \tag{3.5}$$

where we recall that $M = [((r^T \otimes (A^T A)^{-1}) L_T - x^T \otimes A^\dagger), A^\dagger, (A^T A)^{-1}]$. We can get more insight on this condition number. First we remark that it depends on x, b, c, which is not the case for the standard condition number. Depending on the values of such parameters, it can then vary in a wide range, that we can bound. We define $M_1 = (r^T \otimes (A^T A)^{-1}) L_T - x^T \otimes A^\dagger$. It holds that

$$\max\{\|M_1\|, \|A^\dagger\|, \|(A^T A)^{-1}\|\} \leq \|M\| \leq \sqrt{\|M_1\|^2 + \|A^\dagger\|^2 + \|(A^T A)^{-1}\|^2}.$$

$\|M_1\|$ can be bounded repeating the proof of Corollary 2.2 in [17], which uses the properties of the Kronecker product:

$$\Big| \|r\|\|A^\dagger\| - \|x\| \Big|\, \|A^\dagger\| \leq \|M_1\| \leq (\|r\|\|A^\dagger\| + \|x\|)\, \|A^\dagger\|.$$

Let us assume that x is the right singular vector associated with $\sigma_{\min}$, that $b = 0$, and that $\|A^\dagger\| < 1$. In this case $\|Ax\| = \sigma_{\min}$, $\|M_1\| \leq 2\|A^\dagger\|$, and $\|M\| \leq \sqrt{6}\|A^\dagger\|$, so that

$$\kappa_S \leq \sqrt{6}\|A^\dagger\| \sqrt{\|A\|_F^2 + \|c\|^2} \leq \sqrt{6(n+1)}\kappa(A).$$

In the case we choose x as the right singular vector associated with $\sigma_{\max}$ and $b = 0$ we obtain $\|M_1\| \geq \|A^\dagger\|(\kappa(A) - 1)$, $\|c\| = \|A\|^2$ and we can conclude that

$$\kappa_S \geq \|A^\dagger\|(\kappa(A) - 1) \sqrt{\|A\|_F^2 + \|c\|^2} \geq (\kappa(A)^2 - \kappa(A)) \sqrt{1 + \|A\|^2}.$$

Then, we deduce that in some cases κ_S can be as large as a quantity of order $\kappa(A)^2$, while in others it can be as low as $\kappa(A)$. Analogous results can be established if b is in the direction of the left singular vector associated with $\sigma_{\min}$ ($\sigma_{\max}$) and its norm is close to $\|A^\dagger\|$ ($\|A\|$). We will show in section 5 that both cases are often encountered in practice.

3.2. Backward error analysis. In this section, we address the computation of the backward error by considering the following problem. Suppose $\tilde{x}$ is a perturbed solution to (ENE). Find the smallest perturbation (E, f, g) of (A, b, c) such that $\tilde{x}$ exactly solves

$$(A+E)^T(A+E)x = (A+E)^T(b+f) + (c+g).$$

That is, given

$$\begin{aligned}\mathcal{G} := \{(E,f,g), E \in \mathbb{R}^{m\times n}, f \in \mathbb{R}^m, g \in \mathbb{R}^n : \\ (A+E)^T(A+E)\tilde{x} = (A+E)^T(b+f) + (c+g)\},\end{aligned}$$

we want to compute the quantity

$$\eta(\tilde{x}, \theta_1, \theta_2) = \min_{(E,f,g)\in\mathcal{G}} \|(E, \theta_1 f, \theta_2 g)\|_F^2 := \min_{(E,f,g)\in\mathcal{G}} \|E\|_F^2 + \theta_1^2\|f\|^2 + \theta_2^2\|g\|^2 \tag{3.6}$$

with θ_1, θ_2 positive parameters [20, 29].

We provide an explicit representation of the set of admissible perturbations on the matrix (Theorem 3.4) and a linearization estimate for $\eta(\tilde{x}, \theta_1, \theta_2)$ (Lemma 3.5).

Given $v \in \mathbb{R}^m$, we define

$$v^\dagger = \begin{cases} v^T/\|v\|^2 & \text{if } v \neq 0, \\ 0 & \text{if } v = 0. \end{cases}$$

Note the following properties that are used later:

$$(I_m - vv^\dagger)v = 0, \qquad\qquad vv^\dagger v = v.$$

Considering just the perturbations of A we next give an explicit representation of the set of admissible perturbations.

THEOREM 3.4. *Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $c, \tilde{x} \in \mathbb{R}^n$, and assume that $\tilde{x} \neq 0$. Let $\tilde{r} = b - A\tilde{x}$ and define two sets $\mathcal{E}, \mathcal{M}$ by*

$$\begin{aligned}\mathcal{E} &= \{E \in \mathbb{R}^{m\times n} : \ (A+E)^T(b - (A+E)\tilde{x}) = -c\}, \\ \mathcal{M} &= \{v\left(\alpha c^T - v^\dagger A\right) + (I_m - vv^\dagger)(\tilde{r}\tilde{x}^\dagger + Z(I_n - \tilde{x}\tilde{x}^\dagger)) : \\ &\qquad v \in \mathbb{R}^m,\ Z \in \mathbb{R}^{m\times n}, \alpha \in \mathbb{R}\ s.t.\ \alpha\|v\|^2(v^\dagger b - \alpha c^T\tilde{x}) = -1\}.\end{aligned}$$

Then $\mathcal{E} = \mathcal{M}$.

Proof. The proof is inspired by that of [29, Theorem 1.1]. First, we prove $\mathcal{E} \subseteq \mathcal{M}$, so we assume $E \in \mathcal{E}$. We begin by noting the identity, for each v and $\tilde{x}$

$$E = (I_m - vv^\dagger)E\tilde{x}\tilde{x}^\dagger + vv^\dagger E + (I_m - vv^\dagger)E(I_n - \tilde{x}\tilde{x}^\dagger). \tag{3.7}$$

We choose $v = \tilde{r} - E\tilde{x}$. Then $E\tilde{x} = \tilde{r} - v$ and

$$(I_m - vv^\dagger)E\tilde{x} = (I_m - vv^\dagger)\tilde{r}. \tag{3.8}$$

Moreover,

$$-c = (A+E)^T(b-(A+E)\tilde{x}) = (A+E)^T(\tilde{r}-E\tilde{x}) = (A+E)^T v. \tag{3.9}$$

From (3.9), $v^\dagger E = -\frac{c^T}{\|v\|^2} - v^\dagger A$. Hence, from relations (3.7)–(3.9),

$$E = (I_m - vv^\dagger)\tilde{r}\tilde{x}^\dagger - v\left(\frac{c^T}{\|v\|^2} + v^\dagger A\right) + (I_m - vv^\dagger)E(I_n - \tilde{x}\tilde{x}^\dagger).$$

Then $E \in \mathcal{M}$ with $v = \tilde{r} - E\tilde{x}$, $\alpha = -\frac{1}{\|v\|^2}$, and $Z = E$, as

$$\begin{aligned}\alpha\|v\|^2(v^\dagger b - \alpha c^T\tilde{x}) &= -\frac{1}{\|v\|^2}(v^T b + c^T\tilde{x}) = -\frac{1}{\|v\|^2}(v^T b - v^T(A+E)\tilde{x}) \\ &= -\frac{1}{\|v\|^2}v^T(\tilde{r}-E\tilde{x}) = -1,\end{aligned}$$

where the second equality follows from (3.9).

Conversely, let $E \in \mathcal{M}$. Then,

$$E\tilde{x} = \alpha v c^T\tilde{x} - vv^\dagger A\tilde{x} + \tilde{r} - vv^\dagger\tilde{r} = \alpha v c^T\tilde{x} - vv^\dagger b + \tilde{r}, \tag{3.10}$$

$$E^T v = \alpha\|v\|^2 c - A^T v \tag{3.11}$$

and, hence,

$$\begin{aligned}(A+E)^T(b-(A+E)\tilde{x}) &= (A+E)^T(\tilde{r}-E\tilde{x}) = (A+E)^T(vv^\dagger b - \alpha v c^T\tilde{x}) \\ &= (A+E)^T v(v^\dagger b - \alpha c^T\tilde{x}) = \alpha\|v\|^2 c(v^\dagger b - \alpha c^T\tilde{x}) = -c,\end{aligned}$$

where the second equality follows from (3.10), the fourth from (3.11), and the last from the constraint in $\mathcal{M}$. We conclude that $E \in \mathcal{E}$. ∎

Let us remark that if $c = 0$ we recover the known result for least squares problems given in [29, Theorem 1.1]. Note also that the parameterization of the set of perturbations $\mathcal{E}$ is similar to that obtained for equality constrained least squares problems in [10], even if there is no constraint in [10].

Because of the constraint in $\mathcal{M}$, it is rather difficult to find an analytical formula for $\eta(\tilde{x},\theta_1,\theta_2)$. It is, however, easy to find a linearization estimate for $\eta(\tilde{x},\theta_1,\theta_2)$ with θ_1,θ_2 strictly positive, i.e., given

$$h(A,b,c,x) = A^T(b-Ax) + c,$$

we can find (E,f,g) such that

$$\bar{\eta}(\tilde{x},\theta_1,\theta_2) = \min\|(E,\theta_1 f,\theta_2 g)\|_F \quad \text{s.t.} \tag{3.12a}$$

$$h(A,b,c,\tilde{x}) + [J_A, \theta_1^{-1}J_b, \theta_2^{-1}J_c]\begin{bmatrix}\mathrm{vec}(E)\\ \theta_1 f\\ \theta_2 g\end{bmatrix} = 0, \tag{3.12b}$$

where J_A, J_b, J_c are the Jacobian matrices of h with respect to $\mathrm{vec}(A)$, b, c [9].

LEMMA 3.5. *Let $\eta(\tilde{x},\theta_1,\theta_2)$ be defined as in (3.6), $\bar{\eta}(\tilde{x},\theta_1,\theta_2)$ be defined as in (3.12), and $\tilde{r} = b - A\tilde{x}$. Then the linearized backward error satisfies*

$$\bar{\eta}(\tilde{x},\theta_1,\theta_2) = \left\|\begin{bmatrix}\mathrm{vec}(E)\\ \theta_1 f\\ \theta_2 g\end{bmatrix}\right\| = \|J^\dagger h(A,b,c,\tilde{x})\|$$

with $J := [I_n \otimes \tilde{r}^T - A^T(\tilde{x} \otimes I_m), \theta_1^{-1}A^T, \theta_2^{-1}I_n]$. *Moreover, assume that* $\tilde{r} \neq 0$. *If* $4\eta_1\|J^\dagger\|\eta(\tilde{x}, \theta_1, \theta_2) \leq 1$, *then*

$$\frac{2}{1+\sqrt{2}}\ \bar{\eta}(\tilde{x}, \theta_1, \theta_2) \leq \eta(\tilde{x}, \theta_1, \theta_2) \leq 2\ \bar{\eta}(\tilde{x}, \theta_1, \theta_2),$$

where $\eta_1 = \sqrt{\theta_1^{-2} + \theta_2^{-2} + \|\tilde{x}\|^2}$.

Proof. The first assertion follows from [24, section 2]. Simply adding the term corresponding to c in the linearization leads to

$$[J_A, \theta_1^{-1}J_b, \theta_2^{-1}J_c] = [I_n \otimes \tilde{r}^T - A^T(\tilde{x} \otimes I_m), \theta_1^{-1}A^T, \theta_2^{-1}I_n] = J$$

from which the result follows. The second result can be obtained by repeating the arguments of [24, Corollary 2]. ∎

The linearized estimate is usually called an asymptotic estimate, as it becomes exact in the limit for $\tilde{x}$ that tends to the exact solution x [24]. It also has the advantage of being easily computable.

3.3. First-order approximation for the forward error. The formula we have derived in Theorem 3.3 for the structured condition number of (ENE) can be used to provide a first-order estimate Δ_S of the forward error for a solution $\hat{x}$ obtained by a method that does not form the matrix A^TA explicitly. We define this estimate as the product of the relative condition number (3.5) and the relative linearized estimate $\bar{\eta}_r(\hat{x}) := \bar{\eta}(\hat{x}, 1, 1)/\|(A, b, c)\|_F$ of the backward error in Lemma 3.5:

$$\Delta_S := \frac{\sqrt{\|\bar{M}\|}\|(A, b, c)\|_F}{\|\hat{x}\|}\bar{\eta}_r(\hat{x}). \tag{3.13}$$

We show in subsection 5.2.2 that the proposed estimate accurately predicts the forward error in the numerical simulations for the method we propose in section 4 and that it is more accurate than the classical bounds in the cases in which $\kappa_S \sim \kappa(A)$.

In this section we have considered theoretical questions related to the solution of (ENE). In the following, we consider computing a solution in finite-precision arithmetic.

4. A stable variant of CGLS for (ENE). There are several mathematically equivalent implementations of the standard CG method when it is applied to the normal equations. It is well known that they are not equivalent from a numerical point of view and some of them are not stable [12, 27]. In particular Björck, Elfving, and Strakos [5] compare the achievable accuracy in finite precision of different implementations, and show that the most stable implementation is the one that is often referred to as CGLS, which is due to Hestenes and Stiefel [22, section 10], which we report in Algorithm 4.1.

To a large extent, instability is due to the explicit use of vectors of the form A^TAp_k [27]. In the stable implementations of CGLS, forming matrix A^TA explicitly is avoided and matrix-vector products are thus computed from the action of A or A^T on a vector; intermediate vectors of the form Ap_k are used, so that $p_k^TA^TAp_k$ is computed as $\|Ap_k\|^2$. Another crucial difference lays in the fact that in CGLS the residual $r_k = b - Ax_k$ is recurred instead of the full residual of the normal equations $s_k = A^T(b - Ax_k)$. This avoids propagation of the initial error introduced in s_0 by

Algorithm 4.1 CGLS method for $A^TAx = A^Tb$ [22].

Input: A, b, x_0.
Define $r_0 = b - Ax_0$, $s_0 = A^Tr_0$, $p_1 = s_0$.
for $k = 1, 2, \ldots$ **do**
  $t_k = Ap_k$
  $\alpha_k = \|s_{k-1}\|^2/\|t_k\|^2$
  $x_k = x_{k-1} + \alpha_k p_k$
  $r_k = r_{k-1} - \alpha_k t_k$
  $s_k = A^Tr_k$
  $\beta_k = \|s_k\|^2/\|s_{k-1}\|^2$
  $p_{k+1} = s_k + \beta_k p_k$
end for

the computation A^Tb, as s_k is recomputed at each iteration as A^Tr_k, and r_k is not affected by a significant initial error [5].

Writing (ENE) as

$$A^Tr + c = 0, \quad r = b - Ax,$$

suggests that the CGLS method can also be adapted to provide a stable solution method for the case $c \neq 0$. Also in this case we can avoid the operations that are expected to have the same effect as in the case $c = 0$, and this rewriting allows us to design a new method that we name CGLSc, in which the residual $r_k = b - Ax_k$ is recurred, and the full residual is recovered as $s_k = A^Tr_k + c$. The method is described in Algorithm 4.2.

Algorithm 4.2 CGLSc method for $A^TAx = A^Tb + c$.

Input: A, b, x_0
Define $r_0 = b - Ax_0$, $s_0 = A^Tr_0 + c$, $p_1 = s_0$.
for $k = 1, 2, \ldots$ **do**
  $t_k = Ap_k$
  $\alpha_k = \|s_{k-1}\|^2/\|t_k\|^2$
  $x_k = x_{k-1} + \alpha_k p_k$
  $r_k = r_{k-1} - \alpha_k t_k$
  $s_k = A^Tr_k + c$
  $\beta_k = \|s_k\|^2/\|s_{k-1}\|^2$
  $p_{k+1} = r_k + \beta_k p_k$
end for

The next result is proved in [5], which applies to least squares problems and is an extension of the corresponding result in Greenbaum [21], valid for $Ax = b$ with A square and invertible, where the author studies the finite-precision implementation of the class of iterative methods considered in Lemma 4.1.

LEMMA 4.1. *Let $A \in \mathbb{R}^{m\times n}$ have rank n. Consider an iterative method to solve the least squares problem $\min_x \|Ax - b\|^2$, in which each step updates the approximate solution x_k and the residual r_k of the system $Ax = b$ using*

$$x_{k+1} = x_k + \alpha_k\ p_k,$$
$$r_{k+1} = r_k - \alpha_k\ Ap_k,$$

where $\alpha_k \in \mathbb{R}$ and $p_k \in \mathbb{R}^n$. The difference between the true residual $b - Ax_k$ and the

recursively computed residual r_k satisfies

$$\frac{\|b - Ax_k - r_k\|}{\|A\|\|x\|} \leq \epsilon\ O(k) \left(1 + \Theta_k + \frac{\|r\|}{\|A\|\|x\|}\right) \tag{4.1}$$

with ϵ the machine precision, $r = b - Ax$, and $\Theta_k = \max_{j \leq k} \|x_j\|/\|x\|$.

Lemma 4.1 is used in [5] to deduce a bound on the forward error for such methods:

$$\frac{\|x - x_k\|}{\|x\|} \leq \kappa(A)\ \epsilon\ O(k) \left(3 + \frac{\|r\|}{\|A\|\|x\|}\right) + \kappa(A) \frac{\|r - r_k\|}{\|A\|\|x\|}. \tag{4.2}$$

If it can be shown that there is $c_1 > 0$ such that the computed recursive residual r_k satisfies

$$\frac{\|r - r_k\|}{\|A\|\|x\|} \leq c_1 \epsilon + O(\epsilon^2), \quad k \geq S, \tag{4.3}$$

then (4.2) gives an upper bound on the accuracy attainable in the computed approximation. Here S denotes the number of iterations needed to reach a steady state, i.e., a state in which the iterates do not change substantially from one iteration to the next [5]. In this case we deduce that the method might compute more accurate solutions than a backward stable method [5].

Remark 4.2. Lemma 4.1 also applies to CGLSc, as r_k is recurred. We can then deduce that, if condition (4.3) holds, the bound (4.2) is also valid and CGLSc will provide numerical solutions to (ENE) in a stable way. As pointed out in [5], proving condition (4.3) is not an easy task. Nevertheless, we have found from extensive numerical experimentations that this condition is satisfied numerically; see section 5.

5. Numerical experiments. We numerically validate the performance of the method presented in section 4. As we are interested in the possibility of seeking an approximate solution, we consider standard iterative methods as reference methods: CG (Algorithm 5.1) and MINRES [26] (which is applied to (1.1)), but to provide a fair comparison we also consider standard direct methods. We show that CGLSc performs better than CG and MINRES and that it can compare with direct methods in terms of solution accuracy. We also evaluate the first-order estimate of the forward error based on the relative condition numbers derived in subsection 3.3.

5.1. Problem definition and methodology. All the numerical methods have been implemented in MATLAB. For CG[1] and MINRES,[2] MATLAB codes available online have been employed.

For CG, the computation of A^TA is avoided and products are computed as $A^T(Av)$; cf. Algorithm 5.1. We consider (ENE), where $A \in \mathbb{R}^{m \times n}$ has been obtained as $A = U\Sigma V^T$ and U, V have been selected as orthogonal matrices generated with the MATLAB commands[3] gallery('orthog',m,j), gallery('orthog',n,j) for different choices of $j = 1, \ldots, 6$. We consider two choices for the diagonal elements of Σ for $i = 1, \ldots, n$:

- C1 : $\Sigma_{ii} = a^{-i}$ for $a > 0$,
- C2 : $\Sigma_{ii} = \sigma_i$, where $\sigma \in \mathbb{R}^n$ is generated with the linspace MATLAB command, i.e., $\sigma =$ linspace(dw, up, n) with dw, up being strictly positive real values.

The values of a and dw, up are specified for each test; see Table 1.

[1] https://people.sc.fsu.edu/~jburkardt/m_src/kelley/kelley.html
[2] http://stanford.edu/group/SOL/software/minres/
[3] https://www.mathworks.com/help/matlab/ref/gallery.html

Algorithm 5.1 CG method for $A^TAx = A^Tb + c$.

Input: A, b, c, x_0.
Define $s_0 = A^Tb + c - A^T(Ax_0)$, $p_1 = s_0$.
for $k = 1, 2, \ldots$ **do**
  $\alpha_k = \|s_{k-1}\|^2 / \|Ap_k\|^2$
  $x_k = x_{k-1} + \alpha_k p_k$
  $s_k = s_{k-1} - \alpha_k A^T(Ap_k)$
  $\beta_k = \|s_k\|^2 / \|s_{k-1}\|^2$
  $p_{k+1} = s_k + \beta_k p_k$
end for

TABLE 1

Description of the synthetic test problems considered in Table 2. The free parameters in choices C1 and C2 are specified. In particular, in the second column if a single scalar is given then it is the value of a (choice C1); if a couple is given these are the values of dw *and* up *(choice C2). In the tests,* $c = \bar{c} \cdot \mathsf{rand}(\mathsf{n}, 1)$ *for the chosen constant* $\bar{c}$ *given in the last column.*

Pb.	a /(dw, up)	$\bar{c}$
1	0.5	-10^{-3}
2	1.5	10
3	2	10^{-2}
4	2.5	10^{-7}
5	3	2
6	$(10^{-3}, 10^2)$	-1
7	$(10, 10^4)$	10^{-1}
8	$(10^{-6}, 10^{-2})$	10^{-7}
9	$(10^{-1}, 10^3)$	10^2
10	$(10^{-3}, 10^4)$	-10^{-2}

TABLE 2

Comparison of the computed forward error and its estimates for the synthetic test problems listed in Table 1. We report the condition number $\kappa(A)$*, the structured condition number* κ_S*, and for both CG and CGLSc the computed forward error (FE)* $\|x - \hat{x}\|/\|x\|$*, the standard bound* Δ_C*, and the structured estimate* Δ_S*.*

			CG			CGLSc		
Pb.	$\kappa(A)$	κ_S	FE	Δ_C	Δ_S	FE	Δ_C	Δ_S
1	$9 \cdot 10^2$	$1 \cdot 10^6$	$2 \cdot 10^{-11}$	$1 \cdot 10^{-10}$	$6 \cdot 10^{-11}$	$5 \cdot 10^{-13}$	$2 \cdot 10^{-10}$	$1 \cdot 10^{-11}$
2	$2 \cdot 10^3$	$4 \cdot 10^3$	$2 \cdot 10^{-12}$	$4 \cdot 10^{-10}$	$3 \cdot 10^{-12}$	$7 \cdot 10^{-15}$	$3 \cdot 10^{-10}$	$3 \cdot 10^{-13}$
3	$5 \cdot 10^5$	$6 \cdot 10^5$	$1 \cdot 10^{-7}$	$5 \cdot 10^{-5}$	$1 \cdot 10^{-7}$	$1 \cdot 10^{-12}$	$3 \cdot 10^{-5}$	$5 \cdot 10^{-11}$
4	$4 \cdot 10^7$	$4 \cdot 10^7$	$7 \cdot 10^{-6}$	$9 \cdot 10^{-2}$	$2 \cdot 10^{-5}$	$4 \cdot 10^{-11}$	$6 \cdot 10^{-2}$	$4 \cdot 10^{-9}$
5	$1 \cdot 10^9$	$5 \cdot 10^8$	$2 \cdot 10^{-1}$	$1 \cdot 10^2$	$1 \cdot 10^{-1}$	$3 \cdot 10^{-8}$	$7 \cdot 10^2$	$3 \cdot 10^{-7}$
6	$1 \cdot 10^5$	$3 \cdot 10^{10}$	$3 \cdot 10^{-7}$	$3 \cdot 10^{-6}$	$7 \cdot 10^{-7}$	$2 \cdot 10^{-8}$	$3 \cdot 10^{-6}$	$1 \cdot 10^{-7}$
7	$1 \cdot 10^4$	$5 \cdot 10^5$	$6 \cdot 10^{-9}$	$3 \cdot 10^{-8}$	$6 \cdot 10^{-9}$	$6 \cdot 10^{-13}$	$2 \cdot 10^{-8}$	$2 \cdot 10^{-12}$
8	$1 \cdot 10^4$	$8 \cdot 10^9$	$2 \cdot 10^{-9}$	$1 \cdot 10^{-8}$	$1 \cdot 10^{-8}$	$9 \cdot 10^{-10}$	$8 \cdot 10^{-8}$	$7 \cdot 10^{-8}$
9	$1 \cdot 10^4$	$3 \cdot 10^7$	$5 \cdot 10^{-9}$	$1 \cdot 10^{-8}$	$5 \cdot 10^{-9}$	$5 \cdot 10^{-11}$	$2 \cdot 10^{-8}$	$1 \cdot 10^{-10}$
10	$1 \cdot 10^7$	$3 \cdot 10^{10}$	$2 \cdot 10^{-4}$	$2 \cdot 10^{-2}$	$2 \cdot 10^{-4}$	$3 \cdot 10^{-8}$	$3 \cdot 10^{-2}$	$1 \cdot 10^{-7}$

The numerical tests are intended to show specific properties of CGLSc. We therefore consider matrices of relatively small dimensions ($m = 40$ and $n = 20$), in order to avoid too ill-conditioned problems [5]. For all the performance profiles [11] reported in the following, we consider a set of 55 matrices of slightly larger dimension ($m = 100$, $n = 50$), which we later call $\mathcal{P}$, to also test the robustness of the methods (but we do not focus on issues related to large-scale problems). This set is composed of selected matrices from the gallery MATLAB command (those with condition number lower than 10^{10}), and synthetic matrices corresponding to both choices C1 and C2 (of size

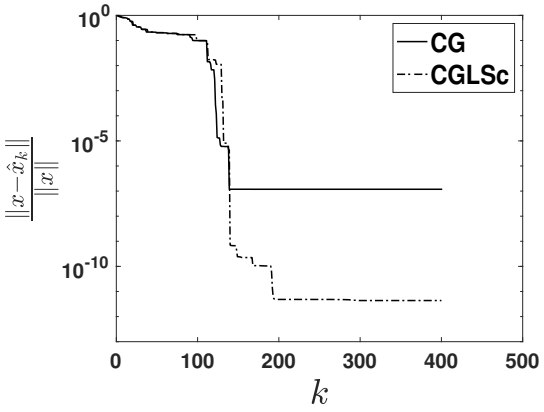

FIG. 1. $\kappa(A) = 10^5$. *Relative error* $\|x-\hat{x}_k\|/\|x\|$ *between the exact solution* x *and the computed solution* $\hat{x}_k$ *at iteration* k *for CG and CGLSc.*

100×50 rather than 40×20). The condition number of the matrices is between 1 and 10^{10}. The optimality measure considered is the relative solution accuracy $\|x-\hat{x}\|/\|x\|$ with x the exact solution (chosen to be $x = [n-1, n-2, \ldots, 0]^T$) and $\hat{x}$ the computed numerical solution. In the tests, c is chosen as a random vector of the form $c = \bar{c}\ \mathsf{rand}(\mathsf{n}, 1)$ for various choices of $\bar{c} \in \mathbb{R}$ (cf. Table 1), and $b = Ax - A^{\dagger^T} c$. A simulation is considered unsuccessful if the relative solution accuracy exceeds 10^{-2}.

5.2. Comparison of CGLS*c* with iterative methods. In this section, we compare the performance of CGLS*c* with the reference iterative methods. We first show its improved performance over CG on selected problems.

5.2.1. Solution accuracy. We first consider the numerical experiment corresponding to choice C1 with $a = 0.5$, $c = 10^{-1}\,\mathsf{rand}(\mathsf{n}, 1)$, and $\kappa(A) = 10^5$. CG and CGLS*c* are compared in Figure 1, where we report the relative error $\|x - \hat{x}_k\|/\|x\|$ versus the number of iterations k. CG achieves an accuracy of order 10^{-7}, while the error in the solution computed by CGLS*c* is $\kappa(A)$ times smaller.

We now consider a second numerical experiment based on the choice C2 with $\mathsf{up} = 0.5$, $\mathsf{dw} = 10^{-8}$, $c = 10^{-14}\,\mathsf{rand}(\mathsf{n}, 1)$, and $\kappa(A) = 5 \times 10^7$, respectively. The gap in the results is even larger: CGLS*c* finds an accurate solution, while CG does not solve the problem at all; see Figure 2.

We finally compare the two iterative methods plus MINRES on the synthetic set of matrices $\mathcal{P}$, where c is chosen as $\bar{c}\ \mathsf{rand}(\mathsf{n}, 1)$ for different values of $\bar{c}$. We report in Figure 3 the performance profile corresponding to these simulations. MINRES has been applied to (1.1) for the two choices $\xi = 1$ and $\xi = \xi^* = \frac{1}{\sqrt{2}}\sigma_{\min}(A)$ [3]. It is evident that a good scaling in (1.1) is beneficial for MINRES, but we also need to take into account that the computation of the optimal scaling parameter may be expensive. Clearly, CGLS*c* performs much better than CG and both versions of MINRES.

5.2.2. Forward error bounds. In this section, we wish to compare the *classical analysis* for which the condition number of (ENE) is $\kappa(A)^2$ and the backward error for a computed solution $\hat{x}$ is $\|A^TA\hat{x} - A^Tb - c\|/(\|A\|^2\|\hat{x}\|)$, with the *structured analysis* for which the relative condition number is given by κ_S in (3.5) and the relative (linearized) backward error is $\bar{\eta}_r(\hat{x}) := \bar{\eta}(\hat{x}, 1, 1)/\|(A, b, c)\|_F$, which is given in

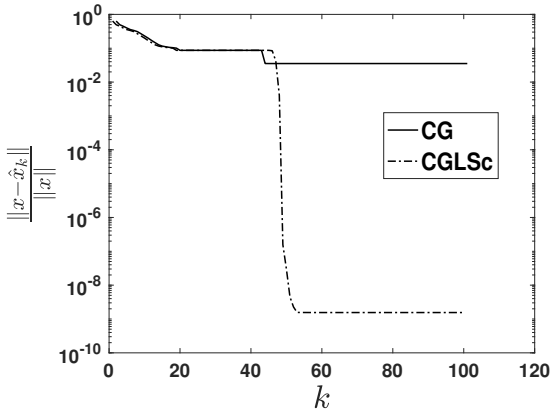

FIG. 2. $\kappa(A) = 5 \times 10^7$. *Relative error* $\|x - \hat{x}_k\|/\|x\|$ *between the exact solution* x *and the computed solution* $\hat{x}_k$ *at iteration* k *for CG and CGLSc.*

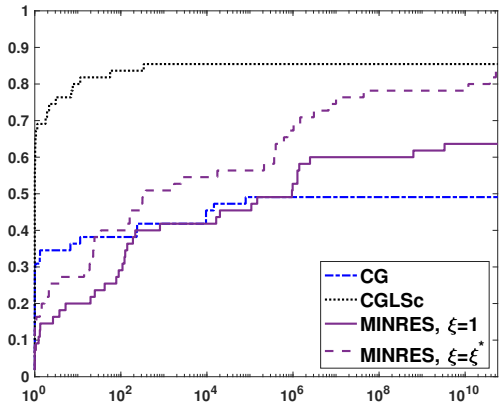

FIG. 3. *Performance profile in logarithmic scale of CG, CGLSc, and MINRES on the synthetic set of matrices* $\mathcal{P}$. *MINRES is applied to* (1.1) *with* $\xi = 1$ *and* $\xi = \xi^*$. *The optimality measure considered is the relative solution accuracy* $\|x-\hat{x}\|/\|x\|$ *with* x *the exact solution and* $\hat{x}$ *the computed numerical solution.*

Lemma 3.5. For each analysis, the forward error $\|x - \hat{x}\|/\|x\|$ is predicted by the product of the condition number times the corresponding backward error, so that the classical bound will be $\Delta_C = \kappa(A)^2\|A^TA\hat{x} - A^Tb - c\|/(\|A\|^2\|\hat{x}\|)$, and the structured one is Δ_S in (3.13).

In Table 2 we compare the two bounds with the forward error for different problems of the form C1 and C2, for which we report the condition number of A and the relative structured condition number κ_S. For all tests, the proposed first-order estimate provides an accurate upper bound for the forward error. On the contrary, the classical analysis gives a rather satisfactory precision when $\kappa_S \sim \kappa(A)^2$, but it is too pessimistic in predicting the error in the case $\kappa_S \sim \kappa(A)$. In such cases the structured bound is much sharper than the standard bound. These results also show again that in general CGLS*c* performs better than CG.

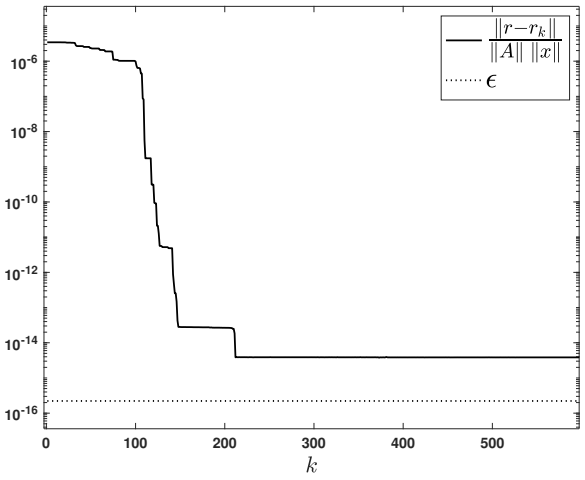

FIG. 4. *CGLSc, choice C*1 *with* $a = 2$ *and* $c = \mathsf{rand}(\mathsf{n}, 1)$. $\|r - r_k\|/(\|A\|\|x\|)$ *with* $r = b - Ax$ *and* r_k *the recurred residual defined in Algorithm* 4.2 *versus iteration index* k *with machine precision* $\epsilon \approx 10^{-16}$.

5.2.3. Norm of the residual. In Figure 4, we consider for CGLS*c* the quantity $\|r - r_k\|/(\|A\|\|x\|)$ that appears in (4.3) (with $r = b - Ax$ and r_k defined in Algorithm 4.2). We can deduce that the bound (4.2) holds for k large enough, as (4.3) is satisfied with $c_1 = O(10)$; see Remark 4.2. This test corresponds to choice C1 with $a = 2$ and $c = \mathsf{rand}(\mathsf{n}, 1)$, but the same behavior is observed in many other simulations.

5.3. Final comparison with direct methods. Motivated by practical applications outlined in section 2, we have mainly focused on the design of iterative methods. However, it is also natural to consider direct methods for the solution of problem (ENE) that are known to be backward stable.

We consider two different methods:

- QR, which solves (1.1) with $\xi = 1$, employing the QR factorization of $[A, b]$ [2, 6], as described in Theorem 5.1 below.
- AUG, which solves the augmented system (1.1) with $\xi = \xi^*$ using an $\mathrm{LBL}^{\mathrm{T}}$ factorization [3]. We implement this method using the ldl MATLAB command.

THEOREM 5.1 (Theorem 1.3.3 [4]). *Let* $A \in \mathbb{R}^{m\times n}$, $m \geq n$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$. *Assume that* $\mathrm{rank}(A) = n$ *and let*

$$[A, b] = Q \begin{bmatrix} R & d_1 \\ 0 & d_2 \end{bmatrix}.$$

For any $\xi \neq 0$, *the solution to* (1.1) *can be computed from*

$$R^T z = -c, \quad Rx = (d_1 - z), \quad r = Q \begin{bmatrix} z \\ d_2 \end{bmatrix}.$$

We remark that the full orthogonal factor Q is required to compute r. Even if the QR factorization can be performed efficiently with Householder transformations,

TABLE 3

Summary of direct and iterative methods. We report the label used for the method, the formulation of the problem the method is applied to, and a brief description of each method.

Label	Formulation	Description
QR	$\begin{bmatrix} I_m & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ -c \end{bmatrix}$	QR factorization of $[A, b]$, Theorem 5.1
AUG	$\begin{bmatrix} \xi I_m & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \xi^{-1} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ -\xi^{-1} c \end{bmatrix}$	LBL$^{\mathrm{T}}$, $\xi = \xi^* = \sigma_{\min}(A)/\sqrt{2}$
CG	$A^T A x = A^T b + c$	CG, Algorithm 5.1
CGLSc	$A^T A x = A^T b + c$	Modified CGLS, Algorithm 4.2
MINRES	$\begin{bmatrix} \xi I_m & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \xi^{-1} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ -\xi^{-1} c \end{bmatrix}$	Minimum residual method, $\xi = 1$, $\xi = \xi^*$

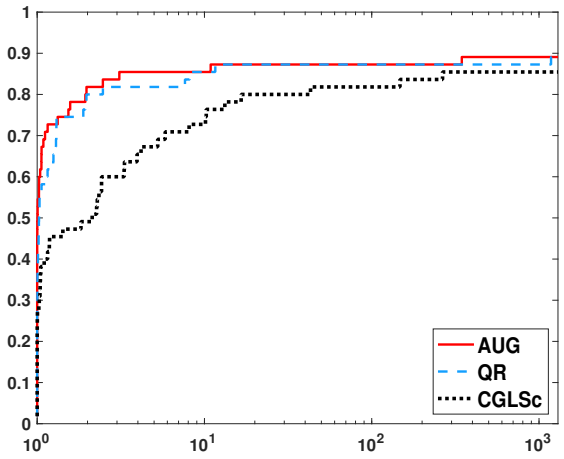

FIG. 5. *Performance profile in logarithmic scale on the synthetic set of matrices $\mathcal{P}$ considering CGLSc and the direct methods in Table 3. The optimality measure considered is the relative solution accuracy $\|x - \hat{x}\|/\|x\|$ with x the exact solution and $\hat{x}$ the computed numerical solution.*

numerical experiments do not show a significant difference in the results using $r = b - Ax$.

All considered methods are summarized in Table 3. In Figure 5 we compare CGLSc with the direct methods on the same set of matrices used for the performance profile of Figure 3. Even if AUG and QR remain slightly more robust and more efficient, the performance of the proposed iterative method is really close to that of the two direct backward stable methods. The performance profile shows indeed that in about 70% of the problems the difference in the solution accuracy is at most one digit and that in all cases it is at most three digits.

6. Conclusions. We considered both theoretical and practical aspects related to the solution of linear systems of the form (ENE). First, we studied the structured condition number of the system and proposed a related explicit formula for its computation. Then, we considered the numerical solution of (ENE). We found that the same issues that degrade the performance of the CG method on the normal equations also arise in this setting. This guided us in the development of a robust iterative method for solving (ENE), which has been validated numerically. From the numeri-

cal experiments, we can draw the following conclusions. The proposed method shows better performance than standard iterative methods in terms of solution accuracy. The error bounds proposed, based on structured condition numbers of the problems, are better able to predict forward errors than classical bounds from linear system theory. Finally, the solution accuracy achieved by the proposed method is comparable to that provided by stable direct methods.

Acknowledgments. The authors wish to thank Nick Higham, Theo Mary, and the anonymous referees for the really useful comments and suggestions that helped to improve the current version of the manuscript.

REFERENCES

[1] M. Arioli, M. Baboulin, and S. Gratton, *A partial condition number for linear least squares problems*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 413–433, https://doi.org/10.1137/050643088.

[2] A. Björck, *Iterative refinement of linear least squares solutions* I, BIT, 7 (1967), pp. 257–278, https://doi.org/10.1007/BF01939321.

[3] A. Björck, *Pivoting and stability in the augmented system method*, in Numerical Analysis 1991, Proceedings of the 14th Dundee Conference, D. F. Griffiths and G. A. Watson, eds., Pitman Res. Notes Math. 260, Longman Scientific and Technical, Essex, UK, 1992, pp. 1–16.

[4] A. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.

[5] A. Björck, T. Elfving, and Z. Strakos, *Stability of conjugate gradient and Lanczos methods for linear least squares problems*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 720–736, https://doi.org/10.1137/S089547989631202X.

[6] A. Björck and C. C. Paige, *Loss and recapture of orthogonality in the Modified Gram–Schmidt algorithm*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 176–190, https://doi.org/10.1137/0613015.

[7] H. Calandra, S. Gratton, E. Riccietti, and X. Vasseur, *On High-Order Multilevel Optimization Strategies*, preprint, https://arxiv.org/abs/1904.04692 (2019).

[8] H. Calandra, S. Gratton, E. Riccietti, and X. Vasseur, *On a multilevel Levenberg-Marquardt method for the training of artificial neural networks and its application to the solution of partial differential equations*, Optim. Methods Softw., 2020, https://doi.org/10.1080/10556788.2020.1775828.

[9] X. W. Chang and D. Titley-Peloquin, *Backward perturbation analysis for scaled total least-squares problems*, Numer. Linear Algebra Appl., 16 (2009), pp. 627–648, https://doi.org/10.1002/nla.640.

[10] A. J. Cox and N. J. Higham, *Backward error bounds for constrained least squares problems*, BIT, 39 (1999), pp. 210–227, https://doi.org/10.1023/A:1022385611904.

[11] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213, https://doi.org/10.1007/s101070100263.

[12] T. Elfving, *On the Conjugate Gradient Method for Solving Linear Least Squares Problems*, Technical report, Department of Mathematics, Linköping University, Linköping, Sweden, 1978.

[13] R. Estrin, M. P. Friedlander, D. Orban, and M. A. Saunders, *Implementing a smooth exact penalty function for general constrained nonlinear optimization*, SIAM J. Sci. Comput., 42 (2020), pp. A1836–A1859, https://doi.org/10.1137/19M1255069.

[14] R. Estrin, D. Orban, and M. A. Saunders, *LNLQ: An iterative method for least-norm problems with an error minimization property*, SIAM J. Matrix Anal. Appl., 40 (2019), pp. 1102–1124, https://doi.org/10.1137/18M1194948.

[15] R. Fletcher, *A class of methods for nonlinear programming:* III. *Rates of convergence*, in Numerical Methods for Non-linear Optimization, Academic, London, 1972, pp. 371–381.

[16] V. Frayssé, S. Gratton, and V. Toumazou, *Structured backward error and condition number for linear systems of the type $A^*Ax = b$*, BIT, 40 (2000), pp. 74–83, https://doi.org/10.1023/A:1022366318322.

[17] V. Frayssé, S. Gratton, and V. Toumazou, *Structured Backward Error and Condition Number for Linear Systems of the Type $A^*Ax = b$*, Technical report CERFACS TR/PA/99/05, Toulouse, France, 2000.

[18] A. J. Geurts, *A contribution to the theory of condition*, Numer. Math., 39 (1982), pp. 85–96, https://doi.org/10.1007/bf01399313.
[19] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., The Johns Hopkins University Press, Baltimore, MD, 2012.
[20] S. Gratton, *On the condition number of linear least squares problems in a weighted Frobenius norm*, BIT, 36 (1996), pp. 523–530, https://doi.org/10.1007/BF01731931.
[21] A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551, https://doi.org/10.1137/S0895479895284944.
[22] M. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Natl. Bur. Standards, 49 (1952), pp. 409–436.
[23] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002, https://doi.org/10.1137/1.9780898718027.
[24] X. G. Liu and N. Zhao, *Linearization estimates of the backward errors for least squares problems*, Numer. Linear Algebra Appl., 19 (2012), pp. 954–969, https://doi.org/10.1002/nla.827.
[25] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, 2006, https://doi.org/10.1007/978-0-387-40065-5.
[26] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629, https://doi.org/10.1137/0712047.
[27] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71, https://doi.org/10.1145/355984.355989.
[28] J. R. Rice, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310, https://doi.org/10.1137/0703023.
[29] B. Walden, R. Karlson, and J. G. Sun, *Optimal backward perturbation bounds for the linear least squares problem*, Numer. Linear Algebra Appl., 2 (1995), pp. 271–286, https://doi.org/10.1002/nla.1680020308.