Effective Condition Number and its Applications *

Zi-Cai Li
Department of Applied Mathematics and
Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan 80424
and Department of Applied Mathematics
Chung Hua University, Hsin-Chu, Taiwan
E-mail: zcli@math.nsysu.edu.tw
Hung-Tsai Huang
Department of Applied Mathematics
I-Shou University, Taiwan 840
E-mail: huanght@isu.edu.tw
Jeng-Tzong Chen
Department of Harbor and River Engineering
National Taiwan Ocean University
Keeling, Taiwan 80424
E-mail: jtchen@mail.ntou.edu.tw
Yimin Wei†
School of Mathematical Sciences, Fudan University,
Shanghai 200433, P.R. China
and Key Laboratory of Mathematics for Nonlinear Sciences (Fudan University),
Ministry of Education, P.R. China.

Abstract

Consider the over-determined system $\mathbf{Fx} = \mathbf{b}$ where $\mathbf{F} \in \mathcal{R}^{m\times n}, m \geq n$ and rank $(\mathbf{F}) = r \leq n$, the effective condition number is defined by Cond_eff $= \frac{\|\mathbf{b}\|}{\sigma_r \|\mathbf{x}\|}$, where the singular values of $\mathbf{F}$ are given as $\sigma_{\max} = \sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$ and $\sigma_{r+1} = ... = \sigma_n = 0$. For the general perturbed system $(\mathbf{A}+\Delta\mathbf{A})(\mathbf{x}+\Delta\mathbf{x}) = \mathbf{b}+\Delta\mathbf{b}$ involving both $\mathbf{\Delta A}$ and $\mathbf{\Delta b}$, the new error bounds pertinent to Cond_eff are derived. Next, we apply the effective condition number to the solutions of Motz's problem by the collocation Trefftz methods (CTM). Motz's problem is the benchmark of singularity problems. We choose the general particular solutions $v_L = \sum_{k=0}^{L} d_k (\frac{r}{R_p})^{k+\frac{1}{2}} \cos(k+\frac{1}{2})\theta$ with a radius parameter R_p.

*Partial results of this paper were represented at the Minisymposium on Collocation and Trefftz Method for the 7th World Congress on Computational Mechanics, Los Angeles, California, July 16-22, 2006.

†Correspondence author (Y. Wei). E-mail: ymwei@fudan.edu.cn and yimin.wei@gmail.com. This author was supported by supported by the National Natural Science Foundation of China under grant 10871051, Doctoral Program of the Ministry of Education under grant 20090071110003, Shanghai Science and Technology Committee under grant 08511501703, Shanghai Municipal Education Committee (Dawn Project) and 973 Program Project (No. 2010CB327900).

The CTM is used to seek the coefficients D_i and d_i by satisfying the boundary conditions only. Based on the new effective condition number, the optimal parameter $R_p = 1$ is found. which is completely in accordance with the numerical results. However, if based on the traditional condition number Cond, the optimal choice of R_p is misleading. Under the optimal choice $R_p = 1$, the Cond grows exponentially as L increases, but Cond_eff is only linear. The smaller effective condition number explains well the very accurate solutions obtained. The error analysis in [14, 15] and the stability analysis in this paper grant the CTM to become the most efficient and competent boundary method.

Key words. Stability analysis, condition number, effective condition number, radius parameter, particular solutions, collocation Trefftz method, singularity problem, Motz's problem.

AMS(MOS) Subject classification, 65N10, 65N30.

1 Introduction

Consider the over-determined system

$$\mathbf{F}\mathbf{x} = \mathbf{b}, \tag{1}$$

where the matrix $\mathbf{F} \in \mathcal{R}^{m\times n}$ and $m \geq n$ with full column rank, e.g., $rank(\mathbf{F}) = n$. The traditional condition number in the 2-norm is defined by [5, 6, 28],

$$\text{Cond} = \frac{\sigma_{\max}}{\sigma_{\min}}, \tag{2}$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are the maximal and the minimal singular values, respectively. The Cond is often too large, to mislead the true stability of the numerical solutions obtained. Hence, we propose the following effective condition number for better stability analysis in [11, 12],

$$\text{Cond\_eff} = \frac{\|\mathbf{b}\|}{\sigma_{\min}\|\mathbf{x}\|}. \tag{3}$$

The effective condition number was first used in Rice [20], and then studied in [3, 4]. Recently, we develop the effective condition number in [11, 12], and apply it to the symmetric and positive definite matrix $\mathbf{F} \in \mathcal{R}^{n\times n}$ from the finite difference method. In this paper, we will apply the effective condition number for over-determined systems from the spectral and Trefftz methods. Let the rank $(\mathbf{F}) = r \leq n$. for (1) and the perturbed system $\mathbf{F}(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}$, there exists the bound in [11, 12],

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \text{Cond\_eff} \times \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{4}$$

Moreover, for (1) and the general perturbed system $(\mathbf{A}+\Delta\mathbf{A})(\mathbf{x}+\Delta\mathbf{x}) = \mathbf{b}+\Delta\mathbf{b}$, where $\mathbf{A}(= \mathbf{F}) \in \mathcal{R}^{n\times n}$ is nonsingular, the errors from the perturbation of both matrix $\mathbf{F}$ and all vector $\mathbf{b}$ are given by ([1, 8, 6])

$$\frac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|} \leq \frac{1}{1-\delta} \times \left\{ \text{Cond} \times \frac{\|\mathbf{\Delta A}\|}{\|\mathbf{A}\|} + \text{Cond} \times \frac{\|\mathbf{\Delta b}\|}{\|\mathbf{b}\|} \right\}, \tag{5}$$

where $\delta = \frac{\|\mathbf{A}\|}{\sigma_n} < 1$. The following errors are derived in [12],

$$\frac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|} \leq \frac{1}{1-\delta} \times \left\{ \text{Cond} \times \frac{\|\mathbf{\Delta A}\|}{\|\mathbf{A}\|} + \text{Cond\_eff} \times \frac{\|\mathbf{\Delta b}\|}{\|\mathbf{b}\|} \right\}. \tag{6}$$

The above bounds are valid for full column rank only; in the next section, new error bounds with rank deficiency will be explored. For numerical partial differential equations (PDEs), since the discretization errors are usually much larger than the errors resulting from solution methods, Cond_eff in (6) is dominate. Hence, we may use the effective condition number for stability analysis. In this paper, we will apply the effective condition number for the Trefftz solutions of Motz's problem, and seek the optimal choice of a parameter used in the particular solutions. This paper also illustrates that the Cond_eff is more advantageous than Cond for stability analysis.

The collocation Trefftz method (CTM) has been proved to most effective method among several boundary method in [14, 15]. However, only the error analysis is made, but no stability exists so far. This paper is devoted to the stability of the CTM, based on the effective condition number. Small effective condition number explains well the high accuracy of the CTM solutions, and strengthens the CTM. In contrast, the huge Cond is misleading.

In [17, 18], Liu tried to use the domain's characteristic length (e.g., R_p in (7)) for basis functions, and the accuracy and stability can be improved for the Trefftz method (TM). So he called his new approaches as the modified TM. Motz's problem is the benchmark of singularity, and it has been used as the popular model for testing numerical partial differential equations, see [10]. For Motz's problem, choose the admissible functions as in [17],

$$v_L^* = \sum_{k=0}^{L} d_k (\frac{r}{R_p})^{k+\frac{1}{2}} \cos(k+\frac{1}{2})\theta, \tag{7}$$

where d_k are the coefficients to be sought, and R_p is the radius parameter. On the other hand, the basic particular solutions are

$$v_L = \sum_{k=0}^{L} D_k r^{k+\frac{1}{2}} \cos(k+\frac{1}{2})\theta, \tag{8}$$

with the coefficients D_i. Since the convergence radius $r = 2$ of (8) is proved in [21], Eq. (8) has been used for Motz's problem by many researchers, see [9, 10, 13, 14, 15, 16, 21].

Obviously, Eq. (8) is a special case with $R_p = 1$ of (7). In [17], Liu found the better solutions at $R_p = 1.71$ in (7) than (8), based on condition number only. In this paper, based on the effective condition number, we will give a comprehensive study for R_p used in (7) for Motz's problem by TM and CTM.

For the spectral methods and TM, choosing a suitable parameter (e.g., R_p) may be helpful for better accuracy and stability, but not for Motz's solutions. In preconditioner, such a technique is well known for better stability. Hence, we should not consider the TM using (7) as the modified TM, because choosing good basis functions is one of requirements to apply TM.

In this paper, both analysis and computation are carried, to confirm that the basis particular solutions are optimal (i.e., $R_p = 1$) for Motz's problem by CTM. Next, for $R_p = 1$ we prove that Cond_eff $= \mathcal{O}(L)$ and Cond $= \mathcal{O}\left(L^{\frac{3}{2}}(\sqrt{2})^L\right)$. The Cond grows exponentially as L increases, but Cond_eff is only linear. The smaller effective condition number explains well the very accurate solutions obtained. The error analysis in [14, 15] and the stability analysis in this paper grant the CTM to become the most efficient and competent boundary method.

This paper is organized as follows. In Section 2, for over-determined systems the effective condition number Cond_eff is defined, and the error bounds pertinent to Cond_eff are derived. In Section 3, the collocation Trefftz method (CTM) is used for Motz's problem, and the general particular solutions (7) are chosen. In Section 4, the bounds of Cond_eff with the parameter R_p are derived, and the optimal radius parameter $R_p = 1$ is found. In Section 5, the stability for CTM with $R_p = 1$ is discussed in detail. In Section 6, numerical experiments are carried out, and in Section 7, a few remarks are made.

2 Effective Condition Number

For solving the over-determined system of linear algebraic equations, the traditional condition number was given in Wilkinson [28], and then discussed in the monographs by Stewart [23, Chapter 3.3] and Higham [7, Chapter 7]. The condition number is used to provide the bounds of relative errors from the perturbation of both $\mathbf{F}$ and $\mathbf{b}$. However, in practical applications, we only deal with a certain vectors $\mathbf{b}$, and the true relative errors may be smaller, or even much smaller than the worst Cond indicates. Such a case was studied in Chan and Faulser [3] and Christiansen and Hansen [4], and called the effective

condition number. However, the effective condition number was first proposed in Rice [20] in 1981, but the natural condition number was called. Below, we will explore the computational formulas to evaluate the effective condition number.

Consider the over-determined system

$$\mathbf{F}\mathbf{x} = \mathbf{b}, \tag{9}$$

where the matrix $\mathbf{F} \in \mathcal{R}^{m\times n}$ and $m \geq n$. When there exists a perturbation of $\mathbf{F}$ and $\mathbf{b}$, we have

$$\mathbf{F}(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}, \tag{10}$$

$$(\mathbf{F} + \Delta\mathbf{F})(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}. \tag{11}$$

Since the exact solutions in (9) – (11) may not exist, the solutions are considered as the least squares solutions: To seek $\mathbf{x}$ and $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\Delta}\mathbf{x}$ such that

$$\min_{\mathbf{x}\in\mathcal{R}^n} \|\mathbf{F}\mathbf{x} - \mathbf{b}\|, \tag{12}$$

$$\min_{\mathbf{x}\in\mathcal{R}^n} \|\mathbf{F}\tilde{\mathbf{x}} - (\mathbf{b} + \Delta\mathbf{b})\|, \tag{13}$$

$$\min_{\mathbf{x}\in\mathcal{R}^n} \|(\mathbf{F} + \Delta\mathbf{F})\tilde{\mathbf{x}} - (\mathbf{b} + \Delta\mathbf{b})\|. \tag{14}$$

First, for simplicity, we suppose the full column rank of $\mathbf{F}$ is n, and then extend the case for rank $r \leq n$. Let matrix $\mathbf{F}$ be decomposed by the singular value decomposition

$$\mathbf{F} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T, \tag{15}$$

where matrices $\mathbf{U} \in \mathcal{R}^{m\times m}$ and $\mathbf{V} \in \mathcal{R}^{n\times n}$ are orthogonal, and matrix $\boldsymbol{\Sigma} \in \mathcal{R}^{m\times n}$ is diagonal with the positive singular values σ_i in a descending order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$. The traditional condition number in the 2-norm is defined by Golub and Van Loan [6, p.223],

$$\text{Cond} = \frac{\sigma_1}{\sigma_n} = \frac{\sigma_{\max}}{\sigma_{\min}}, \tag{16}$$

where $\sigma_{\max} = \sigma_1$ and $\sigma_{\min} = \sigma_n$.

Let us consider (10). Denote $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_m)$ and $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_n)$. We have the expansions

$$\mathbf{b} = \sum_{i=1}^{m} \beta_i \mathbf{u}_i, \quad \Delta\mathbf{b} = \sum_{i=1}^{m} \alpha_i \mathbf{u}_i,$$

where the expansion coefficients are

$$\beta_i = \mathbf{u}_i^T\mathbf{b}, \quad \alpha_i = \mathbf{u}_i^T\Delta\mathbf{b}. \tag{17}$$

Hence, we have

$$\|\mathbf{b}\| = \sqrt{\sum_{i=1}^{m} \beta_i^2}, \quad \|\Delta\mathbf{b}\| = \sqrt{\sum_{i=1}^{m} \alpha_i^2}.$$

Denote the pseudo-inverse matrix $\boldsymbol{\Sigma}^+ \in \mathcal{R}^{n\times m}$ of $\boldsymbol{\Sigma}$ to be diagonal with the entries $\dfrac{1}{\sigma_i}$, see [6, 26]. Hence, the pseudo-inverse matrix of $\mathbf{F}$ is given by $\mathbf{F}^+ = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^T$, and the least squares solution is expressed by

$$\mathbf{x} = \mathbf{F}^+\mathbf{b} = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^T\mathbf{b}. \tag{18}$$

Also from (9) and (10), $\Delta\mathbf{x} = \mathbf{F}^{+}\Delta\mathbf{b} = \mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{T}\Delta\mathbf{b}$. Since $\mathbf{U}$ is orthogonal, we obtain

$$\|\mathbf{x}\| = \|\mathbf{\Sigma}^{+}\mathbf{U}^{T}\mathbf{b}\| = \sqrt{\sum_{i=1}^{n}\frac{\beta_i^2}{\sigma_i^2}}, \tag{19}$$

and [1]

$$\|\Delta\mathbf{x}\| = \|\mathbf{\Sigma}^{+}\mathbf{U}^{T}\Delta\mathbf{b}\| = \sqrt{\sum_{i=1}^{n}\frac{\alpha_i^2}{\sigma_i^2}} \le \frac{1}{\sigma_n}\sqrt{\sum_{i=1}^{n}\alpha_i^2} \le \frac{1}{\sigma_n}\sqrt{\sum_{i=1}^{m}\alpha_i^2} = \frac{\|\Delta\mathbf{b}\|}{\sigma_n}. \tag{20}$$

Hence, we obtain

$$\begin{aligned}\frac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|} &\le \frac{\|\Delta\mathbf{b}\|}{\sigma_n\|\mathbf{x}\|} = \frac{\|\Delta\mathbf{b}\|}{\sigma_n} \times \frac{1}{\sqrt{\sum_{i=1}^{n}\frac{\beta_i^2}{\sigma_i^2}}} \\ &= \text{Cond\_eff} \times \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|},\end{aligned} \tag{21}$$

where

$$\text{Cond\_eff} = \frac{\|\mathbf{b}\|}{\sigma_n\|\mathbf{x}\|} = \frac{\|\mathbf{b}\|}{\sigma_n\sqrt{(\frac{\beta_1}{\sigma_1})^2 + \cdots + (\frac{\beta_n}{\sigma_n})^2}}. \tag{22}$$

Note that when vector $\mathbf{b}$ (i.e., $\mathbf{x}$) is just parallel to the eigenvector $\mathbf{u}_1$, i.e.,

$$\beta_2 = \beta_3 = \cdots = \beta_n = 0, \tag{23}$$

we have $\|\mathbf{b}\| = |\beta_1|$ and Cond_eff $= \frac{\sigma_1}{\sigma_n}$ from (22) leads to the traditional Cond in (16). However, the cases in (23) may not happen for the practical vector $\mathbf{b}$. Hence, the effective condition number may provide a better upper bound of relative errors of the obtained $\mathbf{x}$.

We may extend the above effective condition number for rank deficiency. Suppose $rank(\mathbf{F}) = r \le n$. The singular values are denoted by

$$\sigma_1 \ge \sigma_2 \cdots \ge \sigma_r > 0, \quad \sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0. \tag{24}$$

The traditional condition number, Cond $= \frac{\sigma_1}{\sigma_r}$, is defined by Van Loan [25], and the effective condition number (22) is modified as

$$\text{Cond\_eff} = \frac{\|\mathbf{b}\|}{\sigma_r\sqrt{(\frac{\beta_1}{\sigma_1})^2 + \cdots + (\frac{\beta_r}{\sigma_r})^2}}. \tag{25}$$

On the other hand, when the matrix $\mathbf{F}$ is positive definite and symmetric, the effective condition numbers of this paper are all valid if letting $\sigma_i = \lambda_i$, where λ_i are the eigenvalues of $\mathbf{F}$, see [11, 12].

Below, we consider (11) with $rank(\mathbf{F}) = r \le n$ by the perturbation theory of matrix analysis. First from Wedin [27], Stewart [22], Wang, Wei and Qiao [26], and Sun [24], we have the following lemma.

[1] In practical computation, the worst cases as in (20) may or may not happen. Then in some times, we have $\|\Delta\mathbf{x}\| < \frac{1}{\sigma_n}\|\Delta\mathbf{b}\|$ which may also give a lower bound of $\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|}$ than that in (21).

Lemma 2.1 *Let matrices* $\mathbf{F}, \mathbf{\Delta F} \in \mathcal{R}^{m\times n}, (m \geq n)$ *with* $rank(\mathbf{F}) = rank(\mathbf{F} + \mathbf{\Delta F}) = r \leq n \leq m$ *and denote* $\delta = \|\mathbf{F}^\dagger\| \, \|\mathbf{\Delta F}\| < 1$, *where* $\mathbf{F}^\dagger$ *is the pseudo-inverse matrix of* $\mathbf{F}$. *Then there exist the bounds,*

$$\|(\mathbf{F} + \Delta\mathbf{F})^\dagger\| \leq \frac{\|\mathbf{F}^\dagger\|}{1 - \|\mathbf{F}^\dagger\| \, \|\Delta\mathbf{F}\|} = \frac{\|\mathbf{F}^\dagger\|}{1-\delta}, \tag{26}$$

$$\|(\mathbf{F} + \Delta\mathbf{F})^\dagger - \mathbf{F}^\dagger\| \leq \mu\|(\mathbf{F} + \Delta\mathbf{F})^\dagger\| \, \|\mathbf{F}^\dagger\| \, \|\Delta\mathbf{F}\| \leq \mu\delta\frac{\|\mathbf{F}^\dagger\|}{1-\delta}, \tag{27}$$

where the constant $\mu = \frac{1+\sqrt{5}}{2}$ *if* $r < n \leq m$, $\mu = \sqrt{2}$ *if* $r = n < m$, *and* $\mu = 1$ *if* $r = n = m$.

Theorem 2.1 *Let matrices* $\mathbf{F}, \Delta\mathbf{F} \in \mathcal{R}^{m\times n}, m \geq n$ *with* $rank(\mathbf{F}) = rank(\mathbf{F} + \Delta\mathbf{F}) = r \leq n \leq m$ *and denote* $\delta = \|\mathbf{F}^\dagger\| \, \|\Delta\mathbf{F}\| < 1$. *Then we have*

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \mathit{Cond\_eff} \times \frac{1}{1-\delta}\left[\mu\delta + \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\right], \tag{28}$$

where Cond_eff is defined in (25), and the constant $\mu = \frac{1+\sqrt{5}}{2}$ *if* $r < n \leq m$, $\mu = \sqrt{2}$ *if* $r = n < m$, *and* $\mu = 1$ *if* $r = n = m$.

Proof. Since $\mathbf{x} = \mathbf{F}^\dagger\mathbf{b}$ and $\mathbf{x} + \Delta\mathbf{x} = (\mathbf{F} + \Delta\mathbf{F})^\dagger(\mathbf{b} + \Delta\mathbf{b})$, we have

$$\begin{aligned} \Delta\mathbf{x} &= (\mathbf{F} + \Delta\mathbf{F})^\dagger(\mathbf{b} + \Delta\mathbf{b}) - \mathbf{F}^\dagger\mathbf{b} \\ &= [(\mathbf{F} + \Delta\mathbf{F})^\dagger - \mathbf{F}^\dagger]\mathbf{b} + (\mathbf{F} + \Delta\mathbf{F})^\dagger\Delta\mathbf{b}, \end{aligned} \tag{29}$$

and then

$$\begin{aligned} &\|\Delta\mathbf{x}\| = \|[(\mathbf{F} + \Delta\mathbf{F})^\dagger - \mathbf{F}^\dagger]\mathbf{b} + (\mathbf{F} + \Delta\mathbf{F})^\dagger\Delta\mathbf{b}\| \\ &\leq \|(\mathbf{F} + \Delta\mathbf{F})^\dagger - \mathbf{F}^\dagger\| \, \|\mathbf{b}\| + \|(\mathbf{F} + \Delta\mathbf{F})^\dagger\| \, \|\Delta\mathbf{b}\|. \end{aligned} \tag{30}$$

It follows from Lemma 2.1 that

$$\begin{aligned} \|\Delta\mathbf{x}\| &\leq \mu\frac{\delta}{1-\delta}\|\mathbf{F}^\dagger\| \, \|\mathbf{b}\| + \frac{\|\mathbf{F}^\dagger\|}{1-\delta}\|\Delta\mathbf{b}\| \\ &\leq \mu\frac{\delta}{1-\delta} \times \frac{\|\mathbf{b}\|}{\sigma_r} + \frac{1}{1-\delta} \times \frac{\|\Delta\mathbf{b}\|}{\sigma_r} \\ &= \frac{\|\mathbf{b}\|}{\sigma_r} \times \frac{1}{1-\delta}\left[\mu\delta + \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\right], \end{aligned} \tag{31}$$

by noting that $\|\mathbf{F}^\dagger\| = \frac{1}{\sigma_r}$. The desired result (28) is obtained from (31) and the proof is completed. ∎

When $rank(\mathbf{F}) = n < m$, we have from (28)

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \text{Cond\_eff} \times \frac{1}{1-\delta}\left[\sqrt{2}\delta + \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\right]. \tag{32}$$

When $m = n$ and $rank(\mathbf{F}) = n$,

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \text{Cond\_eff} \times \frac{1}{1-\delta}\left[\delta + \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\right], \tag{33}$$

to give (6) in [12]. Note that the error bounds in (28) with the rank deficiency of $\mathbf{F}$ are given, while those in (32) and (33) are valid only for the full column rank of $\mathbf{F}$. This is a development of effective condition number from [12].

3 Collocation Trefftz Method for Motz's Problem

The spectral method and the Trefftz method using the particular solutions of PDEs can provide the extremely accurate solution, while the traditional Cond are often huge. Since Motz's problem is a benchmark of singularity problems, it has been used as a test model for many numerical methods (see [10]). In Lu et al. [16], the leading coefficient of the Motz's solution by the CTM can have 17 significant digits, while Cond $= \mathcal{O}(10^6)$. Such a puzzle can be clarified well by small effective condition number given in this paper.

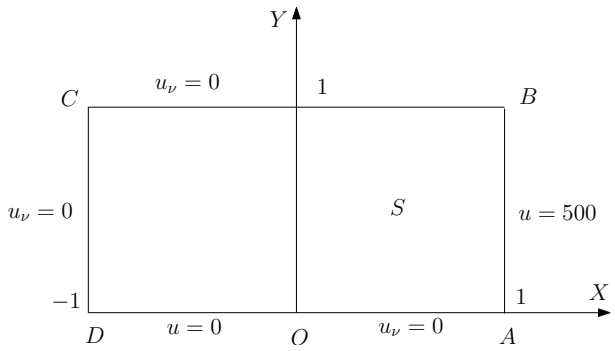

Figure 1: Motz's problem.

Consider Motz's problem (see Figure 1)

$$\begin{aligned} &\triangle u = 0 \ \text{ in } \ S, \\ &u = 0 \ \text{ on } \ \overline{OD}, \quad u = 500 \ \text{ on } \ \overline{AB}, \\ &u_\nu = 0 \ \text{ on } \ \overline{BC} \cup \overline{CD} \cup \overline{OA}, \end{aligned} \tag{34}$$

where $S = \{(x, y) | -1 < x < 1, 0 < y < 1\}$, and $u_\nu = \frac{\partial u}{\partial \nu}$ is the outward normal derivative to ∂S. To solve (34), we may use the collocation Trefftz method (CTM) involving integration approximation. Choose the admissible solutions as

$$u_L = \sum_{i=0}^{L} d_i (\frac{r}{R_p})^{i+\frac{1}{2}} \cos(i + \frac{1}{2})\theta \quad \text{ in } \quad S, \tag{35}$$

where d_i are the unknown coefficients to be sought, and R_p is the bounded radius parameter satisfying

$$R_p \geq r_0 > 0. \tag{36}$$

In our previous study, we always choose (8). Obviously, when $R_p = 1$, Eq. (35) (i.e., (7)) leads to (8), and there exists the relations between the coefficients d_i and D_i,

$$D_i = d_i (\frac{1}{R_p})^{i+\frac{1}{2}}, \quad i = 0, 1, ... \tag{37}$$

Since the expansions (35) satisfy the Laplace equation and the boundary conditions at $y = 0$ already, the coefficients d_i should be chosen to satisfy the rest of the boundary conditions,

$$u\Big|_{\overline{AB}} = 500, \quad u_\nu\Big|_{\overline{BC}} = 0, \quad u_\nu\Big|_{\overline{CD}} = 0, \tag{38}$$

as best as possible, where $\overline{AB}$, $\overline{BC}$ and $\overline{CD}$ are shown in Figure 1. Denote the energy

$$I(v) = \int_{\overline{AB}} (v - 500)^2 d\ell + w^2 \int_{\overline{BC} \cup \overline{CD}} v_\nu^2 d\ell, \tag{39}$$

where w is a positive weight. A good choice of the weight $w = \frac{1}{L+1}$ can be found in [13]. Also denote by V_L the set of the functions (35). The TM reads: To seek $u_L \in V_L$ such that

$$I(u_L) = \min_{v \in V_L} I(v). \tag{40}$$

The equation (40) leads to the linear algebraic system

$$\mathbf{A}\mathbf{x} = \mathbf{b}^*, \tag{41}$$

where $\mathbf{x} \in \mathcal{R}^{L+1}$ is the unknown vector consisting of coefficients d_i, $(i = 0, 1, \cdots, L)$, and $\mathbf{b}^* \in \mathcal{R}^{L+1}$ is the known vector resulting from the boundary condition $u\Big|_{\overline{AB}} = 500$ in (38), and the stiffness matrix, $\mathbf{A} \in \mathcal{R}^{(L+1)\times(L+1)}$, is symmetric and positive definite, but not sparse. By the Gaussian elimination without pivoting in [6], the coefficients d_i (i.e., $\mathbf{x}$) can be obtained. Once the coefficients d_i are known, the errors on $\overline{AB} \cup \overline{BC} \cup \overline{CD}$

$$\|\epsilon\|_B = \|u - u_L\|_B = \left[\int_{\overline{AB}} (500 - u_L)^2 d\ell + w^2 \int_{\overline{BC} \cup \overline{CD}} (u_L)_\nu^2 d\ell \right]^{\frac{1}{2}} \tag{42}$$

are computable. For the TM involving numerical quadrature, we may seek $\tilde{u}_L \in V_L$ such that

$$\tilde{I}(\tilde{u}_L) = \min_{v \in V_L} \tilde{I}(v), \tag{43}$$

where

$$\tilde{I}(v) = \widetilde{\int}_{\overline{AB}} (v - 500)^2 d\ell + w^2 \widetilde{\int}_{\overline{BC} \cup \overline{CD}} v_\nu^2 d\ell. \tag{44}$$

The minimization of $\tilde{I}(v)$ also leads to a linear system as (41). This is a direct implementation to the TM involving numerical integration, called the normal method (NM).

Now, we turn to the collocation Trefftz method (CTM). Suppose that the simplest central rule is chosen. The equations (38) can be performed at the boundary collocation nodes,

$$\sqrt{h} u_L(P_i) = \sqrt{h} 500 \qquad \text{for} \quad P_i \in \overline{AB}, \tag{45}$$

$$w\sqrt{h} \frac{\partial u_L}{\partial y}(P_i^*) = 0 \qquad \text{for} \quad P_i^* \in \overline{BC}, \tag{46}$$

$$w\sqrt{h} \frac{\partial u_L}{\partial x}(Q_i) = 0 \qquad \text{for} \quad Q_i \in \overline{CD}, \tag{47}$$

where h is the integration meshspacing of uniform subsections. Eqs. (45) – (47) are equivalent to (43) with the central rule, see [14, 15]. When other rules such as the Gaussian rule are chosen, the collocation equations (45) − (47) are modified as

$$\alpha_i \sqrt{h} u_L(P_i) = \alpha_i \sqrt{h} 500 \qquad \text{for} \quad P_i \in \overline{AB}, \tag{48}$$

$$w\alpha_i \sqrt{h} \frac{\partial u_L}{\partial y}(P_i^*) = 0 \qquad \text{for} \quad P_i^* \in \overline{BC}, \tag{49}$$

$$w\alpha_i \sqrt{h} \frac{\partial u_L}{\partial x}(Q_i) = 0 \qquad \text{for} \quad Q_i \in \overline{CD}, \tag{50}$$

where the constants $\alpha_i \asymp \mathcal{O}(1)$ [2]. When $\alpha_i = 1$, the equations (48) − (50) lead to (45) − (47). The equations (48) − (50) can be written in the matrix form

$$\mathbf{F}\mathbf{x} = \mathbf{b}, \tag{51}$$

where $\mathbf{F} \in \mathcal{R}^{m\times(L+1)}$ $(m \geq L+1)$ is the stiffness matrix, $\mathbf{x} \in \mathcal{R}^{L+1}$ is the unknown vector consisting of the coefficients d_i, and $\mathbf{b} \in \mathcal{R}^m$ is the known vector. The equation (51) is the over-determined system, and the equation (41) is its normal equation. Solving (51) directly is more advantageous for better stability, see [14, 15].

4 Bounds of Effective Condition Number

In [16], the error analysis of CTM is made for Motz's problem, to give the exponential convergence rates,

$$\|u - u_L\|_B = \mathcal{O}\left(\left(\frac{1}{\sqrt{2}}\right)^L\right), \quad \|u - u_L\|_{\infty,\overline{AB}} = \mathcal{O}\left(\left(\frac{1}{\sqrt{2}}\right)^L\right), \tag{52}$$

which are independent of R_p. The main concern for choosing the radius parameter R_p is that whether or not the instability may damage the accuracy of the Motz solution under a certain working digits. From our recent study [11, 12], the stability analysis should be made, based on Cond_eff, but not on Cond.

In fact, the matrix $\mathbf{F}$ in (51) is given by $\mathbf{F} = \mathbf{B}\mathbf{P}^{-1}$, where $\mathbf{B} \in \mathcal{R}^{m\times n}$ is the stiffness matrix of the CTM from the basis particular solutions (8), and $\mathbf{P} \in \mathcal{R}^{n\times n}$ is the diagonal matrix given by $\mathbf{P} = Diag\{..., (R_p)^{i-\frac{1}{2}}, ...\}$, $(i = 1, 2, ..., L+1)$, and $n = L+1$. Hence the question aries: How to choose the radius parameter R_p, to reduce of Cond and Cond_eff. Since the Cond is often misleading to the true stability, in this section we focus on Cond_eff in (3) and derive its bounds.

Denote

$$\tilde{I}(v) = \frac{1}{2}(\tilde{\mathbf{A}}\mathbf{x}, \mathbf{x}) = \overline{\|v\|}^2_{0,\overline{AB}} + w^2\overline{\|v_\nu\|}^2_{0,\overline{BC}\cup\overline{CD}}, \tag{53}$$

where $(\mathbf{x}, \mathbf{x}) = \|\mathbf{x}\|^2$, $\tilde{\mathbf{A}} = \mathbf{F}^T\mathbf{F}$, the matrix $\mathbf{F}$ is given in (51), and the notations are

$$\overline{\|v\|}^2_{0,\overline{AB}} = \widetilde{\int}_{\overline{AB}} v^2 d\ell, \quad \overline{\|v_\nu\|}^2_{0,\overline{BC}\cup\overline{CD}} = \widetilde{\int}_{\overline{BC}\cup\overline{CD}} v_\nu^2 d\ell. \tag{54}$$

Suppose that the integration rules are chosen such that, to satisfy the following equivalence relations,

$$\overline{\|v\|}_{0,\overline{AB}} \asymp \mathcal{O}(\|v\|_{0,\overline{AB}}), \quad \overline{\|v_\nu\|}_{0,\overline{BC}\cup\overline{CD}} \asymp \mathcal{O}(\|v_\nu\|_{0,\overline{BC}\cup\overline{CD}}). \tag{55}$$

An analysis in [16] shows that the integration rules for (55) are not severe, even the simplest central rule may grant them. Hence, we have

$$\frac{1}{2}(\tilde{\mathbf{A}}\mathbf{x}, \mathbf{x}) = \tilde{I}(v) \asymp I(v) = \frac{1}{2}(\mathbf{A}\mathbf{x}, \mathbf{x}), \tag{56}$$

to give the equivalence relations:

$$\sigma_{\max}(\mathbf{F}) \asymp \sqrt{\lambda_{\max}(\mathbf{A})}, \quad \sigma_{\min}(\mathbf{F}) \asymp \sqrt{\lambda_{\min}(\mathbf{A})}, \tag{57}$$

[2] The notation $a \asymp b$ or $a \asymp \mathcal{O}(b), b > 0$ means that there exist two positive constants C_1 and C_2 such that $C_1 b \leq |a| \leq C_2 b,\ b > 0$.

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximal and the minimal eigenvalues of $\mathbf{A}$, respectively, defined by

$$\lambda_{\max} = \max_{\mathbf{x}\neq 0} \frac{(\mathbf{Ax}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}, \quad \lambda_{\min} = \min_{\mathbf{x}\neq 0} \frac{(\mathbf{Ax}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}. \tag{58}$$

In (58), the notations are given by

$$(\mathbf{x}, \mathbf{x}) = \|\mathbf{x}\|^2 = \sum_{i=0}^{L} d_i^2, \tag{59}$$

$$\begin{aligned} I(v) &= \frac{1}{2}(\mathbf{Ax}, \mathbf{x}) = \int_{\overline{AB}} v^2 d\ell + w^2 \int_{\overline{BC}\cup\overline{CD}} v_\nu^2 d\ell \\ &= \|v\|^2_{0,\overline{AB}} + w^2 \|v_\nu\|^2_{0,\overline{BC}\cup\overline{CD}}, \end{aligned} \tag{60}$$

where $v \in V_L$, and V_L is the set of (35). In the following, C and c_0 are two constants independent of L and R_p, but their values may be different in different places.

We have the following lemma.

Lemma 4.1 *Suppose that for $v \in V_L$ there exists a positive constant $\mu > 0$ such that*

$$\|v\|_{1,\overline{AB}} \leq CL^\mu \|v\|_{0,\overline{AB}}. \tag{61}$$

Then for (51) of the CTM for Motz's problem, there exist the lower bounds,

$$\sigma_{\min} = \sigma_{\min}(\mathbf{F}) \geq c_0 \frac{1}{\sqrt{R_p}} \min\{L^{-\mu}, w\}, \quad \textit{for} \quad R_p \leq 1, \tag{62}$$

$$\sigma_{\min} = \sigma_{\min}(\mathbf{F}) \geq c_0 \left(\frac{1}{R_p}\right)^{L+\frac{1}{2}} \min\{L^{-\mu}, w\}, \quad \textit{for} \quad R_p \geq 1. \tag{63}$$

Proof : We have from (61) and Babuska and Aziz [2, p. 21],

$$\|v\|_{\frac{1}{2},\overline{AB}} \leq C\|v\|_{1,\overline{AB}} \leq CL^\mu \|v\|_{0,\overline{AB}}. \tag{64}$$

Also since $\Delta v = 0$ for $v \in V_L$, we have from [2],

$$\|v_\nu\|_{-\frac{1}{2},\overline{BC}\cup\overline{CD}} \leq C\|v_\nu\|_{0,\overline{BC}\cup\overline{CD}}. \tag{65}$$

In (64) and (65), the semi-norms and the negative norms in the Sobolev space are defined by, respectively

$$\|v\|_{\frac{1}{2},\Gamma} = \left\{ \|v\|^2_{0,\Gamma} + \int_\Gamma \int_\Gamma \frac{(v(P) - v(Q))^2}{(P-Q)^2} d\ell(P) d\ell(Q) \right\}^{\frac{1}{2}},$$

$$\|u\|_{-\frac{1}{2},\Gamma} = \sup_{v\neq 0} \frac{|\int_\Gamma uv d\ell|}{\|v\|_{\frac{1}{2},\Gamma}}.$$

Hence, from (60), (64) and (65), there exists a constant $\bar{c}_0 > 0$ independent of L such that

$$\begin{aligned} I(v) &\geq \bar{c}_0 \left\{ L^{-2\mu} \|v\|^2_{\frac{1}{2},\overline{AB}} + w^2 \|v_\nu\|^2_{-\frac{1}{2},\overline{BC}\cup\overline{CD}} \right\} \\ &\geq \bar{c}_0 \min\{L^{-2\mu}, w^2\} \cdot \{\|v\|^2_{\frac{1}{2},\overline{AB}} + \|v_\nu\|^2_{-\frac{1}{2},\overline{BC}\cup\overline{CD}}\}. \end{aligned} \tag{66}$$

On the other hand, since $\Delta v = 0$ for $v \in V_L$, we have from Oden and Reddy [19, p.189],

$$\|v\|^2_{1,S} \le C\{\|v\|^2_{\frac{1}{2},\overline{AB}} + \|v_\nu\|^2_{-\frac{1}{2},\overline{BC}\cup\overline{CD}}\}. \tag{67}$$

Combining (66) and (67) yields

$$I(v) \ge c_0 \min\{L^{-2\mu}, w^2\}\|v\|^2_{1,S}. \tag{68}$$

Denote the semi-disk with the radius ρ,

$$S_\rho = \{(r,\theta)|0 \le r \le \rho,\ 0 \le \theta \le \pi\}.$$

Since $S_\rho|_{\rho=1} \subset S$, we have

$$\|v\|_{1,S} \ge \|v\|_{1,S_\rho|_{\rho=1}} \ge c_0|v|_{1,S_\rho|_{\rho=1}}. \tag{69}$$

From the Green formula,

$$|v|^2_{1,S_\rho} = \iint_{S_\rho} \left\{ \left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 \right\} ds = \int_{\ell_\rho} v_\nu v d\ell, \tag{70}$$

where $\ell_\rho = \{(r,\theta)|r = \rho, 0 \le \theta \le \pi\}$ is the semi-circle. By calculus, from the orthogonality of $\cos(i+\frac{1}{2})\theta$ we obtain from (35)

$$\begin{aligned}\int_{\ell_\rho} v_\nu v d\ell &= \int_0^\pi \frac{1}{R_p}\left\{\sum_{i=0}^L d_i(i+\frac{1}{2})(\frac{\rho}{R_p})^{i-\frac{1}{2}}\cos(i+\frac{1}{2})\theta\right\}\cdot\left\{\sum_{i=0}^L d_i(\frac{\rho}{R_p})^{i+\frac{1}{2}}\cos(i+\frac{1}{2})\theta\right\}\rho d\theta \\ &= \sum_{i=0}^L d_i^2(i+\frac{1}{2})(\frac{\rho}{R_p})^{2i+1}\int_0^\pi \cos^2(i+\frac{1}{2})\theta d\theta = \frac{\pi}{2}\sum_{i=0}^L d_i^2(i+\frac{1}{2})(\frac{\rho}{R_p})^{2i+1}.\end{aligned} \tag{71}$$

Let $\rho = 1$, and consider two cases: $R_p \le 1$ and $R_p \ge 1$. First when $R_p \le 1$, for $\rho = 1$ we have

$$\int_{\ell_\rho} v_\nu v d\ell \ge \frac{\pi}{2}\frac{1}{R_p}\sum_{i=0}^L d_i^2(i+\frac{1}{2}) \ge \frac{\pi}{4}\frac{1}{R_p}\sum_{i=0}^L d_i^2. \tag{72}$$

Combining (68) $-$ (70) and (72) yields

$$I(v) \ge c_0 \min\{L^{-2\mu}, w^2\}\frac{1}{R_p}\sum_{i=0}^L d_i^2, \tag{73}$$

and then

$$\lambda_{\min}(\mathbf{A}) = \min_{\mathbf{x}\neq 0}\frac{2I(v)}{(\mathbf{x},\mathbf{x})} \ge c_0\frac{1}{R_p}\min\{L^{-2\mu}, w^2\}. \tag{74}$$

Next, when $R_p \ge 1$, for $\rho = 1$ we have

$$\int_{\ell_\rho} v_\nu v d\ell = \frac{\pi}{2}\sum_{i=0}^L d_i^2(i+\frac{1}{2})(\frac{1}{R_p})^{2i+1} \ge c_0(\frac{1}{R_p})^{2L+1}\sum_{i=0}^L d_i^2, \tag{75}$$

to give

$$I(v) \geq c_0(\frac{1}{R_p})^{2L+1} \min\{L^{-2\mu}, w^2\} \sum_{i=0}^{L} d_i^2. \tag{76}$$

Hence we have

$$\lambda_{\min}(\mathbf{A}) = \min_{\mathbf{x}\neq 0} \frac{2I(v)}{(\mathbf{x},\mathbf{x})} \geq c_0(\frac{1}{R_p})^{2L+1} \min\{L^{-2\mu}, w^2\}. \tag{77}$$

The desire results (62) and (63) follow from (57), (74) and (77). This completes the proof of Lemma 4.1. ∎

Theorem 4.1 *Let (61) hold. Then for (51) of the CTM for Motz's problem, there exist the bounds for the effective condition number:*

$$\textit{Cond\_eff} \leq C \max\{L^{\mu}, w^{-1}\}, \quad \textit{if} \;\; R_p \leq 1, \tag{78}$$

$$\textit{Cond\_eff} \leq C(R_p)^L \max\{L^{\mu}, w^{-1}\}, \quad \textit{if} \;\; R_p \geq 1. \tag{79}$$

Proof : By noting (48), the vector $\mathbf{b}$ has the components

$$\mathbf{b}^T = \{\cdots, 500\alpha_i\sqrt{h}, \cdots\}. \tag{80}$$

Since $h \leq C\frac{1}{m}$, where m is the dimension of $\mathbf{b}$, we have

$$\|\mathbf{b}\| = \sqrt{\sum_i (500\alpha_i\sqrt{h})^2} = 500\sqrt{\sum_i \alpha_i^2 h} \leq 500 \max_i |\alpha_i| \sqrt{mh} \leq C, \tag{81}$$

where we have used that $\alpha_i \asymp \mathcal{O}(1)$. Also since $d_0 = D_0\sqrt{R_p}$ from (37), and since the true coefficient D_0 is known (see Table 2), we have

$$\|\mathbf{x}\| = \sqrt{\sum_{i=0}^{L} d_i^2} \geq d_0 = D_0\sqrt{R_p} \geq 400\sqrt{R_p}. \tag{82}$$

Hence, we have from Lemma 4.1, (81) and (82),

$$\text{Cond\_eff} = \frac{\|\mathbf{b}\|}{\sigma_{\min}\|\mathbf{x}\|} \leq C \max\{L^{\mu}, w^{-1}\}, \quad \text{if} \;\; R_p \leq 1, \tag{83}$$

$$\text{Cond\_eff} \leq C(R_p)^L \max\{L^{\mu}, w^{-1}\}, \quad \text{if} \;\; R_p \geq 1. \tag{84}$$

This is the desired results (78) and (79), and completes the proof of Theorem 4.1. ∎

In computation, we choose $w = \frac{1}{L}$, and for the sectorial S we can prove that $\mu = 1$, see [13]. Hence, we have the following corollary from Theorem 4.1 and Lemma 4.1.

Corollary 4.1 *Let (61) hold. Also assume $\mu = 1$ and choose $w = \frac{1}{L}$. Then for (51) (i.e., the CTM for Motz's problem), there exist the bounds:*

$$\textit{Cond\_eff} \leq CL, \quad \sigma_{\min} \geq c_0 \frac{1}{\sqrt{R_p}} L^{-1}, \qquad \textit{if} \;\; R_p \leq 1, \tag{85}$$

$$\textit{Cond\_eff} \leq CL(R_p)^L, \quad \sigma_{\min} \geq c_0 L^{-1} \left(\frac{1}{R_p}\right)^{L+\frac{1}{2}}, \quad \textit{if} \;\; R_p \geq 1. \tag{86}$$

From Theorem 4.1 and Corollary 4.1, we find the optimal case at $R_p \leq 1$ for small Cond_eff. Then we may choose $R_p = 1$. Since the errors (52) retain the same for different R_p, we conclude theoretically that the basis particular solutions (8) (i.e., $R_p = 1$) is optimal for Motz's problem by TM and CTM. Note that this conclusion is against [17].

5 Stability for CTM of $R_p = 1$

Based on the above analysis, we should choose $R_p = 1$. In this section we also derive the bound of Cond for comparison. We have the following lemma.

Lemma 5.1 *Suppose that for $v \in V_L$, there exists a positive constant $\mu > 0$ such that*

$$\|v_\nu\|_{0,\overline{BC}\cup\overline{CD}} \leq CL^\mu \|v\|_{1,S}. \tag{87}$$

Then for (51) of the CTM for Motz's problem, when $R_p = 1$ there exists the upper bound,

$$\sigma_{\max} = \sigma_{\max}(\mathbf{F}) \leq C(1 + wL^\mu)\sqrt{L}(\sqrt{2})^L. \tag{88}$$

Proof : From (87) and the embedding theorem,

$$\|v\|_{0,\overline{AB}} \leq C\|v\|_{1,S}, \tag{89}$$

we obtain from (60)

$$I(v) = \|v\|^2_{0,\overline{AB}} + w^2\|v_\nu\|^2_{0,\overline{BC}\cup\overline{CD}} \leq C(1 + w^2L^{2\mu})\|v\|^2_{1,S}. \tag{90}$$

Since $v|_{y=0\wedge -1<x<0} = 0$ for $v \in V_L$, there exists the bound from the Poincare inequality,

$$\|v\|_{1,S} \leq C|v|_{1,S}, \tag{91}$$

where $|v|_{1,S}$ is the semi-norm of v on S. Hence, we have from (90) and (91),

$$I(v) \leq C(1 + w^2L^{2\mu})|v|^2_{1,S}. \tag{92}$$

Moreover, since $S \subset S_\rho|_{\rho=\sqrt{2}}$, we have from (71) with $R_p = 1$ and $d_i = D_i$,

$$\begin{aligned} |v|^2_{1,S} &\leq |v|^2_{1,S_{\sqrt{2}}} = \int_{\ell_{\sqrt{2}}} v_\nu v d\ell \\ &= \frac{\pi}{2}\sum_{i=0}^{L} D_i^2(i + \frac{1}{2})\rho^{2i+1} \leq \frac{\pi}{2}(L + \frac{1}{2})(\sqrt{2})^{2L+1}\sum_{i=0}^{L} D_i^2. \end{aligned} \tag{93}$$

Combining (92) and (93) yields

$$I(v) \leq C(1 + w^2L^{2\mu})L(\sqrt{2})^{2L}\sum_{i=0}^{L} D_i^2. \tag{94}$$

Hence the maximal eigenvalue $\lambda_{\max}(\mathbf{A})$ has the following bound,

$$\lambda_{\max}(\mathbf{A}) = \max_{\mathbf{x}\neq 0}\frac{2I(v)}{\|\mathbf{x}\|^2} \leq C(1 + w^2L^{2\mu}) \cdot L(\sqrt{2})^{2L}. \tag{95}$$

The desired result (88) follows from (57), and this completes the proof of Lemma 5.1. ∎

Based on Lemmas 4.1 and 5.1, we have the following theorem.

Theorem 5.1 *Let (61) and (87) hold. Then for (51) of the CTM for Motz's problem, when $R_p = 1$ there exists the bound for the traditional condition number:*

$$Cond \leq C(1 + wL^{\mu})\sqrt{L}\max\{L^{\mu}, w^{-1}\} \cdot (\sqrt{2})^{L}. \tag{96}$$

In computation, we choose $w = \frac{1}{L}$, and for the sectorial S we can prove that $\mu = 1$, see [13]. Hence we have the following corollary from Lemmas 4.1 and 5.1.

Corollary 5.1 *Let (61) and (87) hold. Also assume $\mu = 1$ and choose $w = \frac{1}{L}$. Then for (51) of the CTM for Motz's problem, when $R_p = 1$ there exist the bounds:*

$$\sigma_{\min} \geq c_0 L^{-1}, \tag{97}$$

$$\sigma_{\max} \leq C\sqrt{L}(\sqrt{2})^{L}, \tag{98}$$

$$Cond\_eff \leq CL, \tag{99}$$

$$Cond \leq CL^{\frac{3}{2}}(\sqrt{2})^{L}. \tag{100}$$

Corollary 5.1 indicates clearly that for the highly accurate solutions of Motz's problem by the CTM, the small Cond_eff is the correct criterion of numerical stability, but the huge Cond is misleading.

6 Numerical Experiments

6.1 Choice of R_p

In order to see the effects of R_p in (7) on the errors and stability, new numerical experiments are carried out. We use the Gaussian rule with six nodes, and let M denote the number of integration nodes along $\overline{AB}$. Hence $m = 6M$. First, we choose $R_p = 1$, i.e., the basic particular solutions (8). The errors and condition numbers are listed in Table 1, where $\epsilon = u - u_L$. When $L = 34$ the Motz's solution by the CTM is given by the coefficients D_i in Table 2. This solution is the best in accuracy, stability and complexity of algorithms under double precision computation, compared with other TMs in [14]. Note that all computation in this paper is completed by the Fortran programs under double decision. From Table 1, we can see the numerical asymptotes,

$$\|u - u_L\|_B = \mathcal{O}\left((0.55)^L\right), \quad \|u - u_L\|_{\infty,\overline{AB}} = \mathcal{O}\left((0.56)^L\right), \tag{101}$$

which are consistent with (52).

Next we choose different radius parameters $R_p \in [0.8, 2.5]$. Once the coefficients d_i are obtained by the CTM, the original coefficients D_i are obtained from (37). We list the computed d_i and D_i, the errors, Cond and Cond_eff in Table 3. From Table 3, we can draw the following conclusions:

(1) The errors $\|u - u_N\|_B = 0.493(-8)$ and $\|u - u_N\|_{\infty,\overline{AB}} = 0.520(-8)$ are exactly the same for different R_p used. This result also coincides with (52).

(2) When the basic particular solutions (8) (i.e., $R_p = 1$) are used, the leading coefficient D_0 is the most accurate, because its error is less than the rounding error $\tau = \frac{1}{2} \times 10^{-7}$ of double decision.

(3) There exists a minimum of Cond at $R_p = 1.4 \approx \sqrt{2}$, which is much smaller than the Cond at $R_p = 1$; this result is consistent with [17]. However, the stability based on Cond is misleading, see Corollary 5.1.

(4) The effective condition number Cond_eff = 30.2 at $R_p = 1$ is very small. This explains well the highly accurate Motz solution in Table 2, see Section 6.2.

For $R_p = 0.8, 1, 1.2, 1.7$, the errors of D_0, and condition numbers are listed in Tables 4 – 6. Note that all values of $\|\epsilon\|_B$ and $\|\epsilon\|_{\infty,\overline{AB}}$ are the same for different R_p. From Tables 1 and 4 - 6, we can find the following asymptotes:

$$\text{Cond\_eff} = \mathcal{O}(L), \quad \sigma_{\min} = \mathcal{O}(L^{-1}), \quad \text{for} \quad R_p = 1, \tag{102}$$

$$\text{Cond\_eff} = \mathcal{O}(L), \quad \sigma_{\min} = \mathcal{O}(L^{-1}), \quad \text{for} \quad R_p = 0.8, \tag{103}$$

$$\text{Cond\_eff} = \mathcal{O}\left((1.1)^L\right), \quad \sigma_{\min} = \mathcal{O}\left((0.9)^L\right), \quad \text{for} \quad R_p = 1.2, \tag{104}$$

$$\text{Cond\_eff} = \mathcal{O}\left((1.5)^L\right), \quad \sigma_{\min} = \mathcal{O}\left((0.67)^L\right), \quad \text{for} \quad R_p = 1.7. \tag{105}$$

Eqs. (102) and (103) agree with (85) very well, but Eqs. (104) and (105) have a better performance than (86):

$$\text{Cond\_eff} = \mathcal{O}\left((1.2)^L\right), \quad \sigma_{\min} = \mathcal{O}\left(\left(\frac{1}{1.2}\right)^L\right) = \mathcal{O}\left((0.83)^L\right), \quad \text{for} \quad R_p = 1.2, \tag{106}$$

$$\text{Cond\_eff} = \mathcal{O}\left((1.7)^L\right), \quad \sigma_{\min} = \mathcal{O}\left(\left(\frac{1}{1.7}\right)^L\right) = \mathcal{O}\left((0.59)^L\right), \quad \text{for} \quad R_p = 1.7. \tag{107}$$

From the above analysis and computation, we conclude that the basic particular solutions (8) (i.e., $R_p = 1$ in (7)) are optimal for Motz's problems by the CTM. From Table 1 we can see that

$$\text{Cond} = 0.676(6), \quad \text{Cond\_eff} = 30.2, \tag{108}$$

$$\frac{|\Delta D_0|}{D_0} = 0. \tag{109}$$

Eq. (109) implies that the computed D_0 by the CTM is extremely accurate, in the sense that the error is less that the rounding error of computer.

6.2 Extreme Accuracy of D_0

To estimate the relative errors of the leading coefficients D_0, D_1 and D_2, we can have the following proposition.

Proposition 6.1 *Suppose that the leading coefficients $D_i, (i = 0, 1, 2)$, are dominant in $\mathbf{x}$ such that*

$$|D_i| \geq \bar{\alpha}_i \|\mathbf{x}\|, \quad (i = 0, 1, 2), \tag{110}$$

where $\bar{\alpha}_i \geq \bar{\alpha} > 0$. Then there exists the bound

$$\frac{|\Delta D_i|}{|D_i|} \leq \frac{1}{\bar{\alpha}} \times \text{Cond\_eff} \times \frac{\|\Delta \mathbf{b}\|}{\|\mathbf{b}\|}, \tag{111}$$

where $\Delta D_i = D_i - D_i^*$, and D_i^* and D_i are the true and the approximate coefficients respectively.

Proof : From (110) we have

$$\frac{|\Delta D_i|}{|D_i|} \leq \frac{\sqrt{\sum_{i=1}^{L} \Delta D_i^2}}{|D_i|} \leq \frac{\|\Delta \mathbf{x}\|}{\bar{\alpha}_i \|\mathbf{x}\|}. \tag{112}$$

Also from (21) we have

$$\frac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|} \leq \text{ Cond\_eff} \cdot \frac{\|\mathbf{\Delta b}\|}{\|\mathbf{b}\|}. \tag{113}$$

Combining (112) and (113) yields the desired result (111). ∎

We choose $R_p = 1$. From (108), Cond_eff = 30.2 may explain the highly accurate solution in Table 2 with $L = 34$, and Proposition 6.1 indicates that the D_0 has 16 significant digits, provided that $\frac{\|\mathbf{\Delta b}\|}{\|\mathbf{b}\|}$ is just the rounding error. This is the most cases. Occasionally, D_0 has 17 significant digits due to cancelation of rounding errors, or to $\|\Delta\mathbf{x}\| < \frac{1}{\sigma_n}\|\Delta\mathbf{b}\|$, see (20). Moreover, the singular values σ_i and the coefficients β_i are listed in Table 7. From Table 8, we can see the empirical rates

$$\sigma_{\min} \asymp \mathcal{O}(L^{-1}), \quad \text{Cond\_eff} \asymp \mathcal{O}(L), \tag{114}$$

$$\sigma_{\max} \asymp \mathcal{O}\left(\frac{(\sqrt{2})^L}{\sqrt{L}}\right), \quad \text{Cond} \asymp \mathcal{O}\left(\sqrt{L}(\sqrt{2})^L\right). \tag{115}$$

The equation (114) verifies (97) and (99) very well, but the equations (115) have a better performance with a factor $\mathcal{O}(\frac{1}{L})$, than those in (98) and (100).

Moreover, we may compute the true condition number, defined by

$$\text{Cond\_true} = \frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \times \frac{\|\mathbf{b}\|}{\|\Delta\mathbf{b}\|}. \tag{116}$$

We use the true coefficients in [10], and compute $\mathbf{Fx}$ as the $\mathbf{b}$ on the right hand. By solving $\mathbf{Fx} = \mathbf{b}$, the approximate solution $\tilde{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$ is obtained. Then we obtain $\Delta\mathbf{x} = \tilde{\mathbf{x}} - \mathbf{x}$ and $\Delta\mathbf{b} = \mathbf{F}\tilde{\mathbf{x}} - \mathbf{b}$. Based on $\mathbf{x}$, $\Delta\mathbf{x}$, $\mathbf{b}$, $\Delta\mathbf{b}$, the Cond_true in (116) is obtained, and listed in Table 9. It is interesting to note that Cond_true ≈ 1, and the bounds of the traditional Cond are too large and misleading. In contrary, the new Cond_eff $\asymp \mathcal{O}(L)$ is much close to Cond_true. From the viewpoint of Cond, there is a severe instability of the CTM for Motz's problem, but from the viewpoint of Cond_eff, its stability is very well. This is a significant contribution of the new effective condition number, not only to the CTM, but also to numerical partial differential equations.

7 Concluding Remarks

To end this paper, let us make a few final remarks.

1. For solving the over-determined system (1) the traditional condition number (2) in the 2-norm is defined for all $\mathbf{b}$ and $\Delta\mathbf{b}$. In this paper, by following Chan and Fouler [3] and Rice [20] for the given vector $\mathbf{b}$, we define the new effective condition numbers, to provide a better upper bound of the solution errors from the rounding perturbation. In Section 2, the error bounds pertinent to the effective condition number are derived in (28), which can be applied to all kinds of numerical methods for linear algebraic equations, numerical differential equations and numerical integral equations.

2. We apply the effective condition numbers for the CTM for Motz's problem in [16], where the highly accurate solutions are obtained with the exponential convergence rates. In this paper, we focus on the stability analysis, and drive the bounds, Cond_eff $= \mathcal{O}(L)$ and Cond$= \mathcal{O}\left(L^{\frac{3}{2}}(\sqrt{2})^L)\right)$, where L is the number of the singular particular functions used. The Cond_eff $= \mathcal{O}(L)$ explains well the highly accurate

solutions in [16]; while the huge traditional Cond is misleading. The results of effective condition number for the cracked beam problem in [16] are also similar; details are omitted. The TM is a popular method of boundary methods, and its study has become a very active subject in the last two decades. A review of its recent progress is given in Li et al. [14, 15], where only the error analysis is made. This paper is the first time to provide the stability analysis of CTM. It is due to the error analysis in [14, 15] and the stability analysis in this paper that the CTM becomes the most efficient and competent boundary method.

3. Different radius parameter R_p may have an influence on the errors of the leading coefficient D_0, but not on the errors $\|\epsilon\|_B$ and $\|\epsilon\|_{\infty,\overline{AB}}$. Since their error bounds in (52) are independent of R_p, the better choice on R_p is relevant only to stability and the error of D_0. The bounds of Cond_eff are derived in Section 3 for Motz's problem by the CTM. Moreover, the computed results and the theoretical bounds of Cond_eff are consistent with each other.

4. Based on Cond_eff and the error of D_0, we conclude that the basic particular solutions (8) are optimal among (7) (see Table 3). The Motz solution in Table 2 with $L = 34$ and $R_p = 1$ is the highly accurate and stable solution under the double precision. In particular, the leading coefficient D_0 is exact, in the sense that the error of D_0 is less than the rounding error of computer. These conclusions are against to those made in [17], purely based on Cond.

5. In summary, the stability based on Cond_eff is a new development (see [11, 12]), and a new interesting application is given in this paper. The effective condition number may provide a new trend of stability analysis for numerical linear algebraic equations and numerical partial differential equations.

References

[1] K. E. Atkinson, **An Introduction to Numerical Analysis** (2nd Ed.), John Wiley & Sons, New York, 1989.

[2] I. Babuska and A. K. Aziz, Survey lectures in the mathematical foundations of the finite elements, in The Mathematical Foundation of the Finite Element Methods with Applications to Partial Differential Equations, Ed. by A. K. Aziz, pp. 3-369, Academic Press, New York and London, 1973.

[3] T. F. Chan and D. E. Foulser, Effectively well-conditioned linear systems, SIAM J. Stat. Comput., 9 (1988), pp. 963-969.

[4] S. Christiansen and P. C. Hansen, The effective condition number applied to error analysis of certain boundary collocation methods, J. Comput. Appl. Math., 54 (1994), pp. 15-36.

[5] H. Diao and Y. Wei, On Frobenius normwise condition numbers for Moore-Penrose inverse and linear least-squares problems, Numer. Linear Algebra Appl., 14 (2007), pp. 603-610.

[6] G. H. Golub and C. F. van Loan, **Matrix Computations** (2nd Edition), The Johns Hopkins, Baltimore and London, 1989.

[7] Higam, N., **Accuracy and Stabiltiy of Numerical Algorithms**, 2nd Edition, SIAM, Philadelphia, 2002.

[8] Horn, R. A. and Johnson, C. R., **Matrix Analysis**, Cambridge University Press, 1990.

[9] E. Kite, and N. Kamiya, Trefftz method: An overview, Adv. Eng. Software, 24(1995), pp. 3-12.

[10] Z. C. Li, **Combined Methods for Elliptic Equations with Singularities, Interfaces and Infinities**, Kluwer Academic Publishers, Dordrecht, Boston, 1998.

[11] Z. C. Li, C. S. Chien and H. T. Huang, Effective condition number for Finite difference method, J. Comp. and Appl. Math., 198 (2007), pp. 208-235.

[12] Z.C. Li and H.T. Huang, *Effective condition number for numerical partial differential equations*, Numerical Linear Algebra with Applications, 15 (2008), pp. 575-594.

[13] Z. C. Li, R. Mathon and P. Sermer, Boundary methods for solving elliptic problem with singularities and interfaces, SIAM J. Numer. Anal., 24 (1987), pp. 487-498.

[14] Z. C. Li, T. T. Lu, H. T. Huang and A. H.-D. Cheng, Trefftz, collocation and other boundary methods, A comparison, Numer. Meth. PDE, 23 (2007), pp. 93-144.

[15] Z. C. Li, T. T. Lu, H. Y. H and A. H.-D. Cheng, **Trefftz and Collocation methods**, WIT Publisher, Southsampton, January 2008.

[16] T. T. Lu, H. Y. Hu and Z. C. Li, Highly accurate solutions of Motz's and the cracked beam problems, Engineering Analysis with Boundary Elements, 28 (2004), pp. 1387-1403.

[17] C. S. Liu, *A highly accurate solver for the mixed-boundary potential problem and singular problem in arbitrary plane domain*, CMES, 20(2)(2007), pp. 111-122.

[18] C. S. Liu, *A modified Trefftz method for two-dimensional Laplace equation considering the domain characteristic length*, CMES, 21(1)(2007), pp. 53-65.

[19] J. T. Oden and J. N. Reddy, **An Introduction to the Mathematical Theory of Finite Elements**, John Wiley & Sons, New York, 1976.

[20] J. R. Rice, **Matrix Computations and Mathematical Software**, McGraw-Hill Book Company, New York, 1981.

[21] J. B. Rosser and N. Paramichael, *A power series solution of a harmonic mixed boundary value problem*, MRC, Technical report, University of Wisconsin, 1975.

[22] G. W. Stewart, On the perturbation od pseudo-inverse, projections and linear least squares problems, SIAM Review, 19 (1977), pp. 634-662.

[23] Stewart, G., **Matrix Algorithms I: Basic Decompositions**, SIAM, Philadelphia, 1998.

[24] J. G. Sun, **Perturbation Analysis of Matrix** (in Chinese), 2nd edition, Science Press, Beijing, 2001.

[25] C. F. van Loan, Generating the singular value decomposition, SIAM J. Numer. Anal., 13 (1976), pp. 76-83.

[26] G. Wang, Y. Wei and S. Qiao, **Generalized Inverses: Theory and Computations**, Science Press, Beijing, 2004.

[27] P. A. Wedin, Perturbation theory for pseudo-inverses, BIT, 13 (1973), pp. 217-232.

[28] J. H. Wilkinson, **The Algebraic Eigenvalue Problem**, Clarendon Press, Oxford, p. 191, 1965.

List of Tables

L	$\|\epsilon\|_B$	$\|\epsilon\|_{\infty,\overline{AB}}$	$\|\frac{\triangle D_0}{D_0}\|$	$\sigma_{\max}$	$\sigma_{\min}$	Cond	Cond_eff
10	0.146(-1)	0.108(-1)	0.698(-6)	7.06	0.704(-1)	95.2	9.49
14	0.986(-3)	0.623(-3)	0.620(-8)	23.9	0.543(-1)	440	12.9
18	0.780(-4)	0.580(-4)	0.640(-10)	84.4	0.429(-1)	0.197(4)	16.4
22	0.655(-5)	0.550(-5)	0.671(-12)	306	0.354(-1)	0.864(4)	19.8
26	0.578(-6)	0.531(-6)	0.765(-14)	0.113(4)	0.302(-1)	0.374(5)	23.3
30	0.527(-7)	0.522(-7)	0.142(-15)	0.420(4)	0.263(-1)	0.160(6)	26.7
34	0.493(-8)	0.520(-8)	0^*	0.158(5)	0.233(-1)	0.679(6)	30.2

Table 1: Error norms, condition number and errors of leading coefficients from the CTM for Motz's problem for $M = 30$ along $\overline{AB}$ and $R_p = 1.0$, where 0^* denotes the error less than the computer rounding errors in double precision.

i	All digits	Sig. digits	Num. of Sig. digits
0	401.162453745234416	401.16245374523442	17
1	87.6559201950879299	87.6559201950879	15
2	17.2379150794467897	17.2379150794468	15
3	-8.0712152596987790	-8.07121525970	12
4	1.44027271702238968	1.44027271702	12
5	0.331054885920006037	0.33105488592	12
6	0.275437344507860671	0.27543734451	11
7	-0.869329945041107943(-1)	-0.869329945(-1)	9
8	0.336048784027428854(-1)	0.336048784(-1)	9
9	0.153843744594011413(-1)	0.153843745(-1)	9
10	0.730230164737157971(-2)	0.7302302(-2)	7
11	-0.318411361654662899(-2)	-0.3184114(-2)	7
12	0.12206458615497473 6(-2)	0.1220646(-2)	7
13	0.530965295822850803(-3)	0.530965(-3)	6
14	0.271512022889081647(-3)	0.271512(-3)	6
15	-0.120045043773287966(-3)	-0.12005(-3)	5
16	0.505389241414919585(-4)	0.5054(-4)	4
17	0.231662561135488172(-4)	0.2317(-4)	4
18	0.115348467265589439(-4)	0.11535(-4)	5
19	-0.529323807785491411(-5)	-0.529(-5)	3
20	0.228975882995988624(-5)	0.229(-5)	3
21	0.106239406374917051(-5)	0.106(-5)	3
22	0.530725263258556923(-6)	0.531(-6)	3
23	-0.245074785537844696(-6)	-0.25(-6)	2
24	0.108644983229739802(-6)	0.11(-6)	2
25	0.510347415146524412(-7)	0.5(-7)	1
26	0.254050384217598898(-7)	0.3(-7)	1
27	-0.110464929421918792(-7)	-0.1(-7)	1
28	0.493426255784041972(-8)	/	0
29	0.232829745036186828(-8)	/	0
30	0.115208023942516515(-8)	/	0
31	-0.345561696019388690(-9)	/	0
32	0.153086899837533823(-9)	/	0
33	0.722770554189099639(-10)	/	0
34	0.352933005315648864(-10)	/	0

Table 2: The leading coefficients D_i from the CTM for Motz's problem as $L = 34, R_p = 1$ and $M = 30$ along $\overline{AB}$.

R_p	d_0	D_0	$\lvert\frac{\triangle D_0}{D_0}\rvert$	$\sigma_{\max}$	$\sigma_{\min}$	Cond	Cond_eff
0.8	358.810606637983767	401.162453745234473	0.142(-15)	0.331(8)	0.294(-1)	113(10)	26.9
1.0	401.162453745234416	401.162453745234416	0^*	0.158(5)	0.233(-1)	0.679(6)	30.2
1.2	439.451450280775305	401.162453745234359	0.142(-15)	35.2	0.150(-1)	0.239(4)	42.4
1.4	474.661816468163579	401.162453745234700	0.708(-15)	0.734	0.337(-3)	0.218(4)	0.172(4)
1.7	523.051846666585220	401.162453745234018	0.992(-15)	0.523	0.123(-5)	0.424(6)	0.415(6)
2.0	567.329382801379779	401.162453745234302	0.283(-15)	0.463	0.690(-8)	0.671(8)	0.658(8)
2.5	634.293532788443258	401.162453745234700	0.708(-15)	0.400	0.377(-11)	0.106(12)	0.176(11)

Table 3: Errors of D_0, condition numbers, with the numerical d_0 and D_0 from the CTM for Motz's problem for $L = 34$ and $M = 30$, where $\|\epsilon\|_B = 0.493(-8)$ and $\|\epsilon\|_{\infty,\overline{AB}} = 0.520(-8)$, and 0^* denotes the error less than the computer rounding errors in double precision.

L	$\lvert\frac{\triangle D_0}{D_0}\rvert$	$\sigma_{\max}$	$\sigma_{\min}$	Cond	Cond_eff
10	0.698(-6)	70.2	0.933(-1)	75.4	8.49
14	0.620(-8)	580	0.685(-1)	0.846(4)	11.6
18	0.640(-10)	0.499(4)	0.541(-1)	0.923(5)	14.6
22	0.670(-12)	0.441(5)	0.447(-1)	0.987(6)	17.7
26	0.666(-14)	0.397(6)	0.381(-1)	0.104(8)	20.8
30	0.425(-15)	0.361(7)	0.332(-1)	0.109(9)	23.9
34	0.142(-15)	0.331(8)	0.294(-1)	0.113(10)	2.69

Table 4: Errors of D_0 and condition numbers from the CTM for Motz's problem for $R_p = 0.8$ and $M = 30$, where $\|\epsilon\|_B$ and $\|\epsilon\|_{\infty,\overline{AB}}$ are the same as those in Table 1.

L	$\lvert\frac{\triangle D_0}{D_0}\rvert$	$\sigma_{\max}$	$\sigma_{\min}$	Cond	Cond_eff
10	0.698(-8)	1.37	0.597(-1)	23.0	11.6
14	0.620(-8)	2.19	0.438(-1)	49.9	14.5
18	0.640(-10)	3.65	0.346(-1)	106	18.3
22	0.671(-12)	6.30	0.286(-1)	220	22.2
26	0.581(-14)	11.1	0.243(-1)	455	26.1
30	0.283(-15)	19.8	0.212(-1)	934	29.2
34	0.142(-15)	35.7	0.150(-1)	0.239(4)	42.4

Table 5: Errors of D_0, condition numbers from the CTM for Motz's problem for $R_p = 1.2$ and $M = 30$, where $\|\epsilon\|_B$ and $\|\epsilon\|_{\infty,\overline{AB}}$ are the same as those in Table 1.

L	$\lvert\frac{\triangle D_0}{D_0}\rvert$	$\sigma_{\max}$	$\sigma_{\min}$	Cond	Cond_eff
10	0.698(-8)	0.523	0.125(-1)	42.0	41.1
14	0.620(-8)	0.523	0.279(-2)	188	183
18	0.640(-10)	0.523	0.608(-3)	860	841
22	0.672(-12)	0.523	0.131(-3)	0.400(4)	0.391(4)
26	0.723(-14)	0.523	0.278(-4)	0.188(5)	0.184(5)
30	0.142(-15)	0.523	0.587(-5)	0.891(5)	0.871(5)
34	0.992(-15)	0.523	0.123(-5)	0424(6)	0.415(6)

Table 6: Errors of D_0, condition numbers from the CTM for Motz's problem for $R_p = 1.7$ and $M = 30$, where $\|\epsilon\|_B$ and $\|\epsilon\|_{\infty,\overline{AB}}$ are the same as those in Table 1.

i	σ_i	β_i	i	σ_i	β_i
0	.158(5)	.420(2)	18	.156(1)	-.375(2)
1	.121(5)	.610(2)	19	.126(1)	-.602(1)
2	.846(4)	.133(2)	20	.113(1)	-.101(3)
3	.595(4)	.274(1)	21	.974	.130(2)
4	.558(3)	.584(2)	22	.827	-.117(3)
5	.386(3)	.246(2)	23	.720	.144(3)
6	.269(3)	.278(1)	24	.677	-.747(2)
7	.195(3)	-.177(2)	25	.560	.295(2)
8	.513(2)	-.591(2)	26	.463	.243(2)
9	.345(2)	-.544(1)	27	.368	-.180(2)
10	.249(2)	.146(2)	28	.305	-.136(2)
11	.189(2)	-.320(2)	29	.249	.134(2)
12	.931(1)	-.554(2)	30	.188	-.120(2)
13	.619(1)	.231(2)	31	.141	-.898(1)
14	.470(1)	-.421(2)	32	.102	-.908(1)
15	.381(1)	-.416(2)	33	.556(-1)	.815(1)
16	.267(1)	-.306(2)	34	.233(-1)	-.440(1)
17	.202(1)	.777(2)			

Table 7: The singular values σ_i and the coefficients β_i for matrix $\mathbf{F}$ from the CTM solution in Table 2, where the Cond = 0.676(6) and Cond_eff = 30.2.

k	0	1	2	3
$L_k = 10 + 8^k$	10	18	26	34
$\sigma_{\max}^{(k)}$	7.06	84.4	0.113(4)	0.158(5)
$\frac{\sigma_{\max}^{(k)}}{\sigma_{\max}^{(k-1)}}$		12.0	13.4	14.0
$\frac{(\sqrt{2})^8}{\sqrt{\frac{L_k}{L_{k-1}}}}$		12.0	13.3	14.0
$\sigma_{\min}^{(k)}$	0.740(-1)	0.429(-1)	0.320(-1)	0.233(-1)
$\frac{\sigma_{\min}^{(k-1)}}{\sigma_{\min}^{(k)}}$		1.72	1.42	1.29
$\frac{L_k}{L_{k-1}}$		1.80	1.44	1.30
Cond	95.5	0.197(4)	0.375(5)	0.679(6)
Cond_eff	9.50	16.4	23.3	30.2

Table 8: The maximal and the minimal singular values and their empirical asymptotes by the CTM method with $R_p = 1$.

L	10	18	26	34
$\overline{\|\varepsilon\|_0}$	0.508(-12)	0.447(-13)	0.924(-13)	0.130(-12)
Cond	95.5	0.197(4)	0.374(5)	0.679(6)
Cond_eff	9.50	16.4	23.3	30.2
Cond_EE	20.5	35.6	50.7	65.7
$\frac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|}$	0.222(-14)	0.144(-15)	0.288(-15)	0.442(-15)
$\frac{\|\mathbf{\Delta b}\|}{\|\mathbf{b}\|}$	0.176(-14)	0.155(-15)	0.320(-15)	0.451(-15)
Cond_true	1.26	0.932	0.900	0.979

Table 9: Errors, condition numbers, effective condition numbers, and true condition numbers by the CTM method with $R_p = 1$.