Second order cone programming for frictional contact mechanics using interior point algorithm

Vincent Acary, Paul Armand, Hoang Minh Nguyen, Maksym Shpakovych

Second order cone programming for frictional contact mechanics using interior point algorithm

Vincent Acary* Paul Armand† Hoang Minh Nguyen‡ Maksym Shpakovych§

December 27, 2022

Abstract

We report experiments of an implementation of a primal-dual interior point algorithm for solving mechanical models of one-sided contact problems with Coulomb friction. The objective is to recover an optimal solution with high precision and as quickly as possible. These developments are part of the design of Siconos[1], an open-source software for modeling and simulating non-smooth dynamical systems. Currently, Siconos uses mainly first order methods for the numerical solution of these systems. These methods are very robust, but suffer from a linear rate of convergence and are therefore too much slow to recover accurate solutions in a reasonable time. As these variational inequalities systems lead to the solution of an optimization problem with second order cone constraints, a natural idea is to apply second order optimization methods to speed up the convergence. We will present in detail a primal-dual interior point algorithm for minimizing a convex quadratic function with second order cone constraints. We will show, with some examples, that well known implementations of this algorithm such as SDPT3, do not provide solutions satisfactorily in terms of computation time and accuracy. The major difficulty in implementing this type of algorithm comes from the fact that at each iteration of the algorithm, a change of variable, called a scaling, must be performed to guarantee the non-singularity of the linear system to be solved, as well as to recover a symmetric system. While this scaling strategy is very nice from a theoretical point of view, it leads to a huge deterioration of the conditioning of the linear system when approaching the optimal solution and therefore to all the numerical difficulties that result from it. We will detail the numerical algebra that we have developed in our implementation, in order to overcome these problems of numerical instability. We will also present the solution of the models resulting from the problems with rolling friction, for which the cone of constraints is no longer self-dual like the Lorentz cone.

1 Introduction

Contact problems with dry Coulomb friction are present in many design and validation processes in mechanical engineering. As soon as several objects are involved, the question of computing contact

*INRIA Rhône-Alpes; e-mail: vincent.acary@inria.fr
†Laboratoire XLIM, Université de Limoges; e-mail: paul.armand@unilim.fr
‡INRIA Rhône-Alpes; e-mail: hoang-minh.nguyen@inria.fr
§INRIA Bordeaux; e-mail: maksym.shpakovych@inria.fr
[1]https://nonsmooth.gricad-pages.univ-grenoble-alpes.fr/siconos

forces arises. Examples include multi-body systems and mechanisms [42, 14], robotic systems and grasping problems [13, 15, 29, 16], deformable solid mechanics [30, 48, 43], and granular materials [44, 19]. A usual model of one-sided contact is to consider a set of inequality constraints on the configuration parameters of the mechanical problem (positions, rotations) associated with forces which must themselves be positive. Coulomb's friction, which governs how objects slide in relation to each other, imposes a constraint on the contact forces which must remain within the Lorentz cone. These laws lead us to write second order cone constraints on the contact forces. By introducing a non-linear change of variables of the contact relative velocities that defines the so-called modified velocities, the contact laws can be written as a complementarity condition on second order cones between the modified velocities and the contact forces.

In solid and structural dynamics, the discrete mechanical model is usually supplemented by equations of motion that relate the velocities of the system to applied and contact forces. Following space and time discretization, the resulting system can be put in the form of a nonlinear Second Order Cone Complementarity Problem SOCCP [5, 7, 6]. Section 2 details the problem formulation when the discrete equation of motions are assumed to be linear.

To solve these SOCCP problems numerically, a large number of methods have been used in the computational contact mechanics community. Many of these methods are in fact adaptations of well-known mathematical programming methods for solving variational inequalities and complementarity problems. Some of these methods have been jointly developed by optimization specialists. There are two inherent difficulties to the frictional contact mechanics problems. The first difficulty is related to the non-monotone character of the complementarity problem. The identification of an optimization problem for which the complementarity problem would be the optimality conditions is difficult, and leads to non-convex optimization problems. The second difficulty comes from the fact that the constraints are not of full rank and may be of very low rank with respect to the number of constraints in the case of rigid multi-body systems. In [3], a review of the main literature is made and methods are compared with performance profiles. When the second-order constraints are of full rank, second-order methods such as semi-smooth Newton methods, e.g., Alart-Curnier's method [8], are generally robust and accurate. If not, as it is usual in multi-body systems made of rigid parts (robots, granular material), then the first–order methods such as projected successive over-relaxation gradient methods are robust, but slow and with a limited accuracy in practice. Second order techniques fail to solve the problem due to robustness issues. As far as we know, there is no second order method with high accuracy able to solve the frictional contact problems with redundant constraints.

One of the objectives of this work is to propose a second-order method, based on an interior point method that is accurate, robust and efficient for problems where the constraints are rank deficient, but, in the first step, on a convex optimization problem. Some applications of interior point methods have already been attempted for contact problems in the literature [32, 31, 34, 36]. In the precursor work of [32], the contact problem with Tresca friction (purely quadratic cylindrical constraints) is solved with Mehrotra's algorithm. The work in [31] is the most advanced for the case of conic constraints. The problem considered is the relaxed convex problem as in this work, which is further regularized by adding a constant diagonal term to the Jacobian matrices of the interior point algorithm. To solve large problems, linear systems are solved by iterative methods. The degeneracy of the conditioning as the iterations proceed leads to the use of preconditioners and makes it costly to obtain high accuracy solutions. In [36], a comparison between an interior point method and first order methods is made. The interior point method on second-order cone shows a better convergence rate but without reaching higher accuracies than first order methods.

In this article, we consider a relaxed problem where the actual local velocity is complementary to the reaction force at each contact point. This yields a reformulation of the original problem as a convex second order cone optimization (SOCO) problem. A dedicated interior point method based on the primal-dual algorithm of Mehrotra with the Nesterov-Todd scaling strategy is tailored to solve this problem. In particular, we show on a large bench set of examples that a tight accuracy can be achieved at optimum without regularizing the Jacobian matrices, provided their conditioning is controlled.

The outline and the contribution of the article are as follows. In Section 2, we recall the basics of mechanical models and how the convex SOCO problem is obtained. The corresponding primal-dual pair is formulated in Section 3. General results are given on the non-emptiness and compactness of the solution set under the Slater hypothesis. To further motivate the following developments, preliminary experiments with existing numerical solvers are presented in Section 4. The former is done by reformulating a SOCO problem as a differential nonlinear optimization problem as proposed in [12] and applying a nonlinear optimization solver. The latter are performed with SDPT3 [47], a dedicated solver for semi-definite and second order cone optimization. We will see that none of these solutions are suitable, either in terms of efficiency or accuracy. In Section 5, the properties of the central path are detailed. In particular, under the assumption of strict complementarity, we establish a characterization of the limit point of the central path, the analytic center of the optimal set, which to our knowledge is new in the SOCO context. This property of the algorithm is important from the mechanical point of view, since the selection of dual variables, that corresponds to the reaction forces, is completely controlled by the interior point method. The details of the numerical implementation of the algorithm are given in Section 6. Several alternative equivalent formulations of the linear system to be solved at each iteration are detailed and the comparison of their conditioning over the iterations is illustrated. Our experiments show that the choice of formulation is fundamental for the robustness of the algorithm. In Section 7, the interior point method is extended to the case of rolling friction, where the cone of constraints is no longer a Lorentz cone and is not self-dual.

2 Mechanical models

Let us first introduce the original mechanical models that we are interested in. Let d be the dimension of the space of the mechanical system. Two classes of problems are considered. The frictional contact (FC) problems [3] for which $d = 3$ and the problems with rolling friction (RF) at contact [2] for which $d = 5$. Let $n \in \mathbb{N}$ be the number of contact points and let $m \in \mathbb{N}$ be the number of degrees of freedom. The mechanical system is described by means of three vectors: the global velocity $v \in \mathbb{R}^m$, the velocity $u \in \mathbb{R}^{nd}$ and the reaction $r \in \mathbb{R}^{nd}$. After a discretization in time and space of the dynamical system, the general model to be considered is a conic complementarity problem of the form

$$\begin{aligned} &Mv + f = H^\top r, \\ &Hv + w = u, \\ &\mathcal{K}^* \ni u + \Phi(u) \perp r \in \mathcal{K}, \end{aligned} \tag{1}$$

where $M \in \mathbb{R}^{m\times m}$ is a symmetric and positive-definite matrix, $f \in \mathbb{R}^m$, $H \in \mathbb{R}^{nd\times m}$, $w \in \mathbb{R}^{nd}$ and $\mathcal{K} = \prod_{i=1}^n \mathcal{K}_i$, each $\mathcal{K}_i$ is a cone whose definition depends on the model.

For a FC problem, we have

$$\mathcal{K}_i = \left\{ r_i = (r_{i,\mathrm{N}}, r_{i,\mathrm{T}}^\top)^\top \in \mathbb{R} \times \mathbb{R}^{d-1} : \mu_i r_{i,\mathrm{N}} \geq \|r_{i,\mathrm{T}}\| \right\}.$$

This is the Coulomb's friction second-order cone at the ith contact point. The scalar $r_{i,\mathrm{N}}$ (resp. $u_{i,\mathrm{N}}$) and the vector $r_{i,\mathrm{T}}$ (resp. $u_{i,\mathrm{T}}$) are the normal and tangential components of the reaction (resp. velocity) vector of the ith contact point and $\mu \in \mathbb{R}^n$ is the vector of friction coefficients. The dual cone of $\mathcal{K}$, defined by $\mathcal{K}^* := \{u \in \mathbb{R}^{nd} : u^\top r \geq 0, \text{for all } u \in \mathcal{K}\} = \prod_{i=1}^n \mathcal{K}_i^*$, is given by

$$\mathcal{K}_i^* = \left\{u_i = (u_{i,\mathrm{N}}, u_{i,\mathrm{T}}^\top)^\top \in \mathbb{R} \times \mathbb{R}^{d-1} : u_{i,\mathrm{N}} \geq \mu_i \|u_{i,\mathrm{T}}\|\right\}.$$

The function $\Phi : \mathbb{R}^{nd} \to \mathbb{R}^{nd}$ is defined by n components of the form $\Phi_i(u) = (\mu_i \|u_{i,\mathrm{T}}\|, 0^\top)^\top \in \mathbb{R}^d$.

For a RF problem, the cones are of the form

$$\mathcal{K}_i = \left\{r_i = (r_{i,\mathrm{N}}, r_{i,\mathrm{T}}^\top, r_{i,\mathrm{R}}^\top)^\top \in \mathbb{R} \times \mathbb{R}^{d-1} \times \mathbb{R}^{d-1} : \mu_i r_{i,\mathrm{N}} \geq \|r_{i,\mathrm{T}}\|,\ \mu_{\mathrm{R},i} r_{i,\mathrm{N}} \geq \|r_{i,\mathrm{R}}\|\right\}.$$

The vector $r_{i,\mathrm{R}}$ is the rolling friction reaction moment at contact and $\mu_{\mathrm{R}} \in \mathbb{R}^n$ is the vector of rolling friction coefficients. The dual cone of $\mathcal{K}_i$ is

$$\mathcal{K}_i^* = \left\{u_i = (u_{i,\mathrm{N}}, u_{i,\mathrm{T}}^\top, u_{i,\mathrm{R}}^\top)^\top \in \mathbb{R} \times \mathbb{R}^{d-1} \times \mathbb{R}^{d-1} : u_{i,\mathrm{N}} \geq \mu_i \|u_{i,\mathrm{T}}\| + \mu_{\mathrm{R},i} \|u_{i,\mathrm{R}}\|\right\}.$$

The vector $u_{i,\mathrm{R}}$ is the relative angular velocity at contact. For this model, the components of the function Φ are defined by $\Phi_i(u) = (\mu_i \|u_{i,\mathrm{T}}\| + \mu_{\mathrm{R},i} \|u_{i,\mathrm{R}}\|, 0^\top)^\top \in \mathbb{R}^d$.

By observing that $\Phi(u) = \Phi(u + \Phi(u))$ and making the change of variable $u \leftarrow u + \Phi(u)$, the problem (1) is reformulated as

$$\begin{aligned} &Mv + f = H^\top r, \\ &Hv + w + \Phi(u) = u, \\ &\mathcal{K}^* \ni u \perp r \in \mathcal{K}. \end{aligned} \tag{2}$$

In [7], for the case of a FC problem, an iterative solution of (2) is proposed as follows. The second equation is rewritten as

$$Hv + w + \phi(s) = u \quad \text{and} \quad s_i = \mu_i \|u_{i,\mathrm{T}}\|,\ i \in \{1, \ldots, n\},$$

where $\phi : \mathbb{R}^n \to \mathbb{R}^{nd}$ is defined by n components of the form $(s_i, 0^\top)^\top \in \mathbb{R}^d$. The idea is to solve the nonlinear system as a parametric system of the form

$$\begin{aligned} &Mv + f = H^\top r \\ &Hv + w + \phi(s) = u \\ &\mathcal{K}^* \ni u \perp r \in \mathcal{K}, \end{aligned} \tag{3}$$

where s is periodically updated according to

$$s_i = \mu_i \|u_{i,\mathrm{T}}\|,\ i \in \{1, \ldots, n\}.$$

There is no guarantee of global convergence of this procedure, but the advantage of this formulation is that, for a fixed s, the parametric system corresponds to the first order optimality conditions of the following convex optimization problem:

$$\begin{aligned} \min_{v,u} \quad & \tfrac{1}{2} v^\top M v + f^\top v \\ \text{s.t.} \quad & Hv + w + \phi(s) = u, \\ & u \in \mathcal{K}^*. \end{aligned}$$

This problem belongs to the class of second order cone optimization (SOCO) problems. Our study focuses on the numerical solution of this problem. Obviously, the same parametric strategy can be applied for the solution of an RF problem.

3 Basic properties

To simplify the models, let us remove the friction coefficients by means of a simple change of variables. Let us define the diagonal matrix P_μ whose ith diagonal block is $\mathrm{diag}(1, \mu_i, \mu_i)$ for the FC model and $\mathrm{diag}(1, \mu_i, \mu_i, \mu_{\mathrm{R},i}, \mu_{\mathrm{R},i})$ for the RF model. The parameter s is assumed to be fixed. Let us make the change of variables:

$$u \leftarrow P_\mu u, \quad H \leftarrow P_\mu H, \quad w \leftarrow P_\mu(w + \phi(s)) \quad \text{and} \quad r \leftarrow P_\mu^{-1} r.$$

We keep the same name for the variables, but modify the notation of the components. For the RF model, we define

$$u_i = \begin{pmatrix} u_{i,0} \\ \bar{u}_i \\ \tilde{u}_i \end{pmatrix} = \begin{pmatrix} u_{i,\mathrm{N}} \\ \mu_i u_{i,\mathrm{T}} \\ \mu_{\mathrm{R},i} u_{i,\mathrm{R}} \end{pmatrix} \quad \text{and} \quad r_i = \begin{pmatrix} r_{i,0} \\ \bar{r}_i \\ \tilde{r}_i \end{pmatrix} = \begin{pmatrix} r_{i,\mathrm{N}} \\ \frac{1}{\mu_i} r_{i,\mathrm{T}} \\ \frac{1}{\mu_{\mathrm{R},i}} r_{i,\mathrm{R}} \end{pmatrix}$$

The same notation is used for the FC model, except that the components $\tilde{u}_i$ and $\tilde{r}_i$ does not appear. The system (3) becomes

$$\begin{aligned} &Mv + f = H^\top r, \\ &Hv + w = u, \\ &\mathcal{F}^* \ni u \perp r \in \mathcal{F}, \end{aligned} \tag{4}$$

where $\mathcal{F}$ is the product of cones. The definition of $\mathcal{F}$ depends on the model. For the FC model ($d = 3$) we have

$$\mathcal{F} = \prod_{i=1}^{n} \mathcal{L}_i := \{r_i = (r_{i0}, \bar{r}_i) \in \mathbb{R}^d : r_{i0} \geq \|\bar{r}_i\|\}. \tag{5}$$

This cone is self-dual (i.e., $\mathcal{L}_i^* = \mathcal{L}_i$), therefore

$$\mathcal{F}^* = \prod_{i=1}^{n} \mathcal{L}_i^* := \{u_i = (u_{i0}, \bar{u}_i) \in \mathbb{R}^d : u_{i0} \geq \|\bar{u}_i\|\}. \tag{6}$$

For the RF model ($d = 5$), the cone of constraints is

$$\mathcal{F} = \prod_{i=1}^{n} \mathcal{R}_i := \{r_i = (r_{i0}, \bar{r}_i, \tilde{r}_i) \in \mathbb{R}^d : r_{i0} \geq \max\{\|\bar{r}_i\|, \|\tilde{r}_i\|\}\}. \tag{7}$$

This cone is not self-dual. The dual cone is then

$$\mathcal{F}^* = \prod_{i=1}^{n} \mathcal{R}_i^* := \{u_i = (u_{i0}, \bar{u}_i, \tilde{u}_i) \in \mathbb{R}^d : u_{i0} \geq \|\bar{u}_i\| + \|\tilde{u}_i\|\}. \tag{8}$$

The system (4) can be viewed as the first order optimality conditions of the following second order cone optimization problem (SOCO):

$$\begin{aligned} \min_{v,u} \quad & \tfrac{1}{2} v^\top M v + f^\top v \\ \text{s.t.} \quad & Hv + w = u, \\ & u \in \mathcal{F}^*. \end{aligned} \tag{9}$$

Let $r \in \mathbb{R}^{nd}$ be the dual variable associated to the linear constraint of (9). The Lagrangian function associated to problem (9) is defined by

$$L(v,u,r) = \frac{1}{2}v^\top Mv + f^\top v + r^\top(u - Hv - w).$$

Since L is separable in v and u, convex in v and linear in u, and since $\mathcal{F}^{**} = \mathcal{F}$, the dual function is readily obtained and defined by

$$\inf_{(v,u)\in\mathbb{R}^m\times\mathcal{F}^*} L(v,u,r) = \begin{cases} -\frac{1}{2}r^\top Wr - q^\top r - \frac{1}{2}f^\top M^{-1}f & \text{if } r \in \mathcal{F}, \\ -\infty & \text{otherwise,} \end{cases}$$

where

$$W = HM^{-1}H^\top \quad \text{and} \quad q = w - HM^{-1}f. \tag{10}$$

The dual problem is then

$$\begin{array}{ll} \min\limits_{r} & \frac{1}{2}r^\top Wr + q^\top r \\ \text{s.t.} & r \in \mathcal{F}, \end{array} \tag{11}$$

Let us denote by $\widehat{R}$ the optimal set of this problem. It is a convex set, characterized by the first order optimality conditions:

$$\mathcal{F}^* \ni Wr + q \perp r \in \mathcal{F}. \tag{12}$$

Note that these conditions are equivalent to (4), thanks to the definitions (10) and the equalities

$$v = M^{-1}(H^\top r - f) \quad \text{and} \quad u = Wr + q. \tag{13}$$

By weak duality, the duality gap (i.e., the difference between the value of the primal and the dual function) is nonnegative. Indeed, let (v,u,r) be a primal-dual feasible solution, we have

$$\tfrac{1}{2}v^\top Mv + f^\top v - (-\tfrac{1}{2}r^\top Wv - q^\top r - \tfrac{1}{2}f^\top M^{-1}f) = u^\top r \geq 0,$$

the inequality arising from the fact that $r \in \mathcal{F}$ and $u \in \mathcal{F}^*$. The strong duality and the existence of an optimal solution for (11), we need a constraints qualification assumption.

Assumption 1 (Slater hypothesis)**.** *There exists $v \in \mathbb{R}^m$ such that $Hv + w \in \operatorname{int}(\mathcal{F}^*)$.*

With respect to the reduced problem (11), the Slater hypothesis can be equivalently formulated as follows: for all nonzero $d \in \mathcal{F}$, if $Wd = 0$ then $q^\top d > 0$.

The following result summarizes the link between the solutions of (4) and (11). They are well known and are derived from basic properties of convex optimization. We provide a proof in Appendix A to be complete.

Proposition 1. *Consider the primal-dual pair of SOCO problems* (9)-(11)*, where $M \in \mathbb{R}^{m\times m}$ is symmetric and positive-definite, $f \in \mathbb{R}^m$, $H \in \mathbb{R}^{nd\times m}$, $w \in \mathbb{R}^{nd}$ and $\mathcal{F}$ is the product of second order cones of the form* (5) *or* (7)*.*

(i) Any solution of (4)*, provides a primal-dual optimal solution of* (9)-(11)*.*
(ii) If the problem (9) *is feasible, then it has a unique optimal solution $(\hat{v},\hat{u})$.*
(iii) Assumption 1 is satisfied if and only if the optimal set of (11) *is non-empty and compact.*

4 Numerical experiments with existing softwares

From a numerical point of view, one of the main difficulties in modeling the Lorentz cone, is the non-differentiability of the norm at zero. A SOCO problem can be cast as a differentiable nonlinear optimization problem by using some reformulation tricks. Benson and Vanderbei [12] reported experiments on the solutions SOCO problems with LOQO, their nonlinear optimization solver, and asserted that *"..., a general-purpose solver is quite efficient on problems with linear and second-order cone constraints, especially when a perturbation or reformulation approach is used to handle nonsmoothness issues"*. To make sure that we do not miss a simple and straightforward numerical solution using an existing nonlinear optimization solver, we have tried to solve our problems by using the alternate formulations suggested in [12, §4]. Let us consider the following four alternative formulations of the second order cone constraint $\|\bar{r}\| \leq r_0$:

- Smoothing by perturbation (PER): $(\|\bar{r}\|^2 + \epsilon)^{1/2} \leq r_0$, for some small $\epsilon > 0$ (typically 10^{-8}).
- Smoothing by squaring (SQU): $\|\bar{r}\|^2 \leq r_0^2$ and $r_0 \geq 0$.
- Convexification by exponentiation (EXP): $e^{(\|\bar{r}\|^2 - r_0^2)/2} \leq 1$ and $r_0 \geq 0$.
- Convexification by ratios (RAT): $\frac{\|\bar{r}\|^2}{r_0} \leq r_0$ and $r_0 \geq 0$.

These reformulations have been applied to the solution of the reduced problem (11). The choice of solving the dual problem is justified by the fact that a solution of the primal problem (9) does not provide the optimal reaction vector, because the multiplier associated to a reformulated constraint $\|\bar{u}\| \leq u_0$ is not the dual variable r associated to u. Once an optimal solution r of (11) is computed, the global velocity v and the velocity u are calculated according to (13). We then estimate the quality of the optimal solution, by calculating the duality gap $u^\top r$ and the global error value

$$\mathrm{err}(r, u) = \frac{\|r - \pi(r - u)\|}{\max\{\|r\|, \|u\|\}}, \tag{14}$$

where π is the orthogonal projection onto $\mathcal{L}^n$. This error measure is based on the fact that r is optimal for (11) if and only if $r = \pi(r - u)$ [3].

The model problems were coded in AMPL [25], a modeling language for optimization. The original data files of each problem are in HDF5 format[2]. The data files for the AMPL models were generated by Matlab by using function `hdf5read`. The solver KNITRO [17] is called from AMPL with an optimization tolerance `optol` set to 10^{-10}.

There are seven families of friction problem tests [4]. For these tests we have grouped the two families `Spheres` (200 problems) and `Spheres1mm` (41 problems) into one, so that we consider only six families. For each family, we choose to solve only one problem with the largest number of contact points. The results are reported in Appendix B. For each run, we report the number of iterations, the cpu time, the duality gap and the global error (14). The last line of the tables indicates the results obtained by means of our primal-dual solver GFC3D described thereafter. The stopping criterion used in GFC3D is

$$\mathrm{res}(v, u, r) := \max\left\{ \frac{\|Hv + w - u\|}{\max\{\|Hv\|, \|w\|, \|u\|\}}, \frac{\|Mv + f - H^\top r\|}{\max\{\|Mv\|, \|f\|, \|H^\top r\|\}}, |u^\top r| \right\} \leq \mathrm{tol}, \tag{15}$$

[2] https://github.com/FrictionalContactLibrary/fclib-library/tree/master/Global/siconos

with tol $= 10^{-10}$.

The results reported in Tables 7-9, show that no reformulation gives good performances both in terms of cpu time, duality gap and global error, except the PER reformulation for the solution of the KaplasTower problem. This problem is specific, because at the optimal solution all the conic components of the reaction vector r are in the interior of the Lorentz cone, which leads locally as an unconstrained problem. We also note that for some solutions, the duality gap is not positive, meaning that the conical constraints are not always satisfied. Our conclusion about these experiments is that it is no need to go further in this direction.

We also carried out some experiments with the software SDPT3 [47]. This is an implementation of a primal-dual interior point method, the Mehrotra predictor-corrector algorithm, which we hoped would be effective and robust in solving our problems. We have performed experiments by solving the problems under the formulation (9). Since in SDPT3 the objective function is linear, the problem (9) must be reformulated as

$$\begin{array}{ll} \min\limits_{v,u,t} & t \\ \text{s.t.} & Hv + w = u, \\ & \|L^\top v + L^{-1} f\| \leq t, \\ & u \in \mathcal{F}^*, \end{array} \tag{16}$$

where the matrix L is the Cholesky factor of M. The structure of the problem is quite simple. In addition to the linear constraint, the cone of constraints is made with the product of n three-dimensional Lorentz cones and only one cone of dimension m.

We have performed experiments on the 27 problems of the BoxStacks family. The dimensions of the problems are $31 \leq n \leq 557$ and $m = 450$. The original stopping test of SDPT3 has been modified so that the iterations are stopped when (15) is satisfied. Two tests were carried out. One with the tolerance tol $= 10^{-8}$ and the other with tol $= 10^{-10}$. The results are reported in Table 1. The mean number of iterations is calculated on successful runs. We found that these results are not

tolerance (tol)	1e-8	1e-10
number of failures	3	18
mean number of iterations	47	59

Table 1: Numerical solution of the 27 BoxStacks problems with SDPT3.

robust enough and also that the number of iterations is too large. We tested the code SDPT3 on some other problems in the collection, but found the same behavior as for the BoxStacks problems. The code lacks robustness as soon as the tolerance is small, typically less than 10^{-8}. The number of iterations to satisfy the stopping test becomes very large or the code has difficulties with the solution of the linear system and returns the error message "linsysolve: Schur complement matrix not positive definite" or the evaluation error message "linsysolve: solution contains NaN or inf".

These first numerical experiments motivated us to make our own implementation of the interior point algorithm. Especially since in the conclusion of [47] the authors state that "For problems with second-order (quadratic) cone constraints, experiments indicate that there is room for improvement in SDPT3".

5 Central path related to a frictional contact problem

In this section, we are interested in FC problems. We recall some known results about the convergence of the central path in SOCO and put them in our context. For the sake of completeness we also provide the proofs.

For FC problems, the friction contact cone is equal to the product of n Lorentz cones, i.e., $\mathcal{F} = \mathcal{L}^n$, where $\mathcal{L}$ is the Lorentz cone of dimension d. We recall that this cone is self-dual, i.e., $\mathcal{F}^* = \mathcal{F}$. There is an algebra, the so-called *Euclidean Jordan algebra*, associated to symmetric cones, which allows an almost direct extension of the interior-point algorithms for linear optimization to the case of SOCO [9]. A summary of Euclidean Jordan algebra is given in Appendix C. Throught this, the orthogonality condition in (12) becomes $u \circ r = 0$, see [9, Lemma 15]. This leads to a square system of equations with the additional conical constraints. Under Assumption 1, (r, u) is a primal-dual optimal solution of problem (11) if and only if

$$\begin{array}{lcl} Wr + q & = & u, \\ r \circ u & = & 0, \\ (r, u) & \in & \mathcal{L}^{2n}. \end{array} \tag{17}$$

It is important to keep in mind that the matrix W is not any positive semidefinite matrix, but has a structure given by (10). This structure will be useful for the solution of the linear system done at each iteration. In interior-point methods, a perturbation is introduced in the complementarity equation, so that the system to solve becomes

$$\begin{array}{lcl} Wr + q & = & u, \\ u \circ r & = & 2\mu e, \\ (u, r) & \in & \text{int}(\mathcal{L}^{2n}), \end{array} \tag{18}$$

where e is the unit vector related to the Jordan product and $\mu > 0$ is a parameter progressively driven to zero along the iterations, so that at the end a solution of (17) is found. The parameter μ is called the *barrier parameter*, because (18) can be interpreted as the optimality condition of the barrier problem

$$\min_r \quad \varphi_\mu(r) := \tfrac{1}{2} r^\top W r + q^\top r - \mu \textstyle\sum_{k=1}^n \log\det r_k \tag{19}$$

The first order optimality condition of (19) is $Wr + q - 2\mu r^{-1} = 0$. By introducing the variable $u = 2\mu r^{-1}$, we retrieve (18). This system (18) defines a curve, called the central path. The following result states that under the Slater's hypothesis, the central path is well defined and remains bounded for bounded values of μ. The proof is given in Appendix D

Proposition 2. *Under Assumption 1, for all $\mu > 0$, the perturbed KKT system* (18) *has a unique solution $(r(\mu), u(\mu)) \in$ int$(\mathcal{L}^{2n})$. For all $\bar{\mu} > 0$, the set $\{(r(\mu), u(\mu)) : 0 \le \mu \le \bar{\mu}\}$ is bounded.*

From Proposition 1-((ii)), the optimal solution of (9) is unique, therefore $\lim_{\mu\to 0} u(\mu) = \hat{u}$. For the curve $r(\cdot)$, it can be shown that $\lim_{\mu\to 0} r(\mu) = \hat{r}$, where $\hat{r} \in \text{ri}(\widehat{R})$. Such a solution is called a *maximally complementary optimal solution* of (11). It can be characterized as follows. Regarding to the problem (11), the index set $\{1, \ldots, n\}$ of the Lorentz cones is partitioned into six sets [38]:

$$\mathsf{B} = \{i : \exists r \in \widehat{R}, r_i \in \text{int}(\mathcal{L})\}, \mathsf{N} = \{i : \hat{u}_i \in \text{int}(\mathcal{L})\},$$

$$\mathsf{R} = \{i : \hat{u}_i \in \text{bd}(\mathcal{L})\backslash\{0\} \text{ and } \exists r \in \widehat{R}, r_i \in \text{bd}(\mathcal{L})\backslash\{0\}\},$$

$$\mathsf{T}_0 = \{i \in \mathsf{T} : \forall r \in \widehat{R}, \hat{u}_i = r_i = 0\}, \mathsf{T}_1 = \{i \in \mathsf{T} : \hat{u}_i = 0 \text{ and } \exists r \in \widehat{R}, r_i \in \text{bd}(\mathcal{L})\backslash\{0\}, \},$$

$$\mathsf{T}_2 = \{i \in \mathsf{T} : \hat{u}_i \in \text{bd}(\mathcal{L})\backslash\{0\} \text{ and } \forall r \in \widehat{R}, r_i = 0\}.$$

An optimal solution $r \in \widehat{R}$ is maximally complementary if and only if for all $i \in \mathsf{B}$, $r_i \in \text{int}(\mathcal{L})$ and for all $i \in \mathsf{R} \cup \mathsf{T}_1$, $r_i \in \text{bd}(\mathcal{L})\backslash\{0\}$.

The next result shows the convergence of the central path to a maximally complementary solution. A proof is given in Appendix D for self-completeness.

Proposition 3. *Under Assumptions 1, the central path* $(r(\cdot), u(\cdot))$ *converges to* $(\hat{r}, \hat{u})$*, where* $\hat{r}$ *is a maximally complementary optimal solution of* (11)*.*

In linear optimization, it is known that the central path converges to the analytic center of the optimal set [39]. In semidefinite optimization, it is also known that this result holds under strict complementarity [22, 28]. But in SOCO, as far as we know, we did not find in the literature a characterization of the analytic center, even under the strict complementarity assumption. In order to complete the theory, we provide such a characterization.

Assumption 2 (Strict complementarity)**.** *There exists* $r \in \widehat{R}$ *such that* $\hat{u} + r \in \text{int}(\mathcal{L}^n)$.

This assumption is equivalent to $\mathsf{T}_i = \emptyset$, for $i = 1, 2, 3$. In that case, $(\mathsf{B}, \mathsf{N}, \mathsf{R})$ is a partition of $\{1, \ldots, n\}$. Under this assumption, we can define the analytic center of the optimal set $\widehat{R}$. If $\widehat{R}$ is a singleton, then the analytic center of $\widehat{R}$ is this point. Otherwise, it is the unique optimal solution of the problem

$$\min_{r \in \text{ri}(\widehat{\mathrm{R}})} \psi(r) := -\sum_{i \in \mathsf{B}} \log\det r_i - \sum_{i \in \mathsf{R}} \log r_{i,0}. \tag{20}$$

The next proposition shows that the analytic center is well defined.

Proposition 4. *Under Assumptions 1–2, if* $\widehat{R}$ *is not reduced to a singleton, then the problem* (20) *has a unique optimal solution* $\hat{r} \in \text{ri}(\widehat{\mathrm{R}})$*, characterized by the following property: for all* $r \in \text{ri}(\widehat{\mathrm{R}})$*,*

$$\sum_{i \in \mathsf{B}} r_i^\top \hat{r}_i^{-1} + \frac{1}{2} \sum_{i \in \mathsf{R}} \frac{r_{i,0}}{\hat{r}_{i,0}} \leq |\mathsf{B}| + \frac{1}{2}|\mathsf{R}|. \tag{21}$$

Proof. Let $r \in \text{ri}(\widehat{\mathrm{R}})$. For all $i \in \mathsf{B}$, $\det r_i > 0$, and for all $i \in \mathsf{R}$, $r_{i,0} = \|\bar{r}_i\| > 0$. Since $\text{ri}(\widehat{\mathrm{R}}) \neq \emptyset$, one has $\text{dom}\psi \cap \text{ri}(\widehat{\mathrm{R}}) \neq \emptyset$. To show that problem (20) has at least one solution, it suffices to show that the function $\psi + \delta_{\text{ri}(\widehat{\mathrm{R}})}$ is coercive. This last property is a direct consequence of the compactness of the set $\widehat{R}$.

The uniqueness of the minimum comes from the strict convexity of ψ on $\text{ri}(\widehat{\mathrm{R}})$. Indeed, for a conic component $i \in \mathsf{B}$, we have $\nabla_r^2(-\log\det r)_{r=r_i} = 2Q_{r_i^{-1}}$, which is positive definite for all $r_i \in \text{int}(\mathcal{L})$, see Appendix C. For all conic component $i \in \mathsf{R}$, there exists $h_i \in \mathbb{R}^{d-1}$, with $\|h_i\| = 1$, such that for all $r \in \text{ri}(\widehat{R})$, we have $r_i = r_{i,0}(1, h_i^\top)^\top$. Indeed, suppose that for $i \in \mathsf{R}$, there exist r and r' in $\text{ri}(\widehat{R})$, such that r_i and r'_i are not collinear. By the triangle inequality, it is easy to see that $\frac{1}{2}(r_i + r'_i) \in \text{int}(\mathcal{L})$, and thus $i \in \mathsf{B}$, which would contradict $i \in \mathsf{R}$. Finally, for all $r, r' \in \text{ri}(\widehat{\mathrm{R}})$, if

$r \neq r'$ then there exists $i \in \mathsf{B}$ such that $r_i \neq r'_i$ or there exists $i \in \mathsf{R}$ such that $r_{i,0} \neq r'_{i,0}$. In both cases, we have $\psi(r) \neq \psi(r')$, which implies that ψ is strictly convex on $\mathrm{ri}(\widehat{\mathrm{R}})$.

Since ψ is convex on $\mathrm{ri}(\widehat{\mathrm{R}})$, $\hat{r}$ is optimal if and only if $-\nabla\psi(\hat{r})$ is in the normal cone to $\widehat{R}$ at $\hat{r}$, i.e., for all $r \in \mathrm{ri}(\widehat{\mathrm{R}})$, $\nabla\psi(\hat{r})^\top(r - \hat{r}) \geq 0$. For $r \in \mathrm{ri}(\widehat{\mathrm{R}})$, we have

$$\begin{aligned}\nabla\psi(\hat{r})^\top(r-\hat{r}) &= -2\sum_{i\in\mathsf{B}}(\hat{r}_i^{-1})^\top(r_i-\hat{r}_i) - \sum_{i\in\mathsf{R}}\frac{r_{i,0}-\hat{r}_{i,0}}{\hat{r}_{i,0}} \\ &= -2\sum_{i\in\mathsf{B}} r_i^\top\hat{r}_i^{-1} + 2|\mathsf{B}| - \sum_{i\in\mathsf{R}}\frac{r_{i,0}}{\hat{r}_{i,0}} + |\mathsf{R}|,\end{aligned}$$

from which we deduce that (21) is satisfied. □

Let us state and prove the main result of this section.

Theorem 1. *Under Assumptions 1–2, the central path $r(\cdot)$ converges to the analytic center of $\widehat{R}$.*

Proof. Suppose that $\widehat{R}$ is not reduced to a singleton, otherwise the result is a direct consequence of Proposition 3. Let $r \in \mathrm{ri}(\widehat{\mathrm{R}})$. As in the proof of Proposition 3, for all $\mu > 0$, (45) is satisfied. By using $r(\mu) \circ u(\mu) = 2\mu e$ and the definition of the partition $(\mathsf{B}, \mathsf{N}, \mathsf{R})$, we have

$$\sum_{i\in\mathsf{B}} r_i^\top r_i^{-1}(\mu) + \sum_{i\in\mathsf{N}} \hat{u}_i^\top u_i^{-1}(\mu) + \sum_{i\in\mathsf{R}}(r_i^\top r_i^{-1}(\mu) + \hat{u}_i^\top u_i^{-1}(\mu)) \leq n. \tag{22}$$

For $i \in \mathsf{R}$, by using the Cauchy-Schwarz inequality and the fact that $r_{i,0} = \|\bar{r}_i\|$, we have

$$\begin{aligned} r_i^\top r_i^{-1}(\mu) &= \frac{r_i^\top R r_i(\mu)}{\det r_i(\mu)} \\ &= \frac{r_{i,0}\, r_{i,0}(\mu) - \bar{r}_i^\top\bar{r}_i(\mu)}{r_{i,0}^2(\mu) - \|\bar{r}_i(\mu)\|^2} \\ &\geq \frac{r_{i,0}\, r_{i,0}(\mu) - \|\bar{r}_i\|\|\bar{r}_i(\mu)\|}{(r_{i,0}(\mu) - \|\bar{r}_i(\mu)\|)(r_{i,0}(\mu) + \|\bar{r}_i(\mu)\|)} \\ &= \frac{r_{i,0}}{r_{i,0}(\mu) + \|\bar{r}_i(\mu)\|}. \end{aligned} \tag{23}$$

In the same manner, for all $i \in \mathsf{R}$ we have

$$\hat{u}_i^\top u_i^{-1}(\mu) \geq \frac{\hat{u}_{i,0}}{u_{i,0}(\mu) + \|\bar{u}_i(\mu)\|}. \tag{24}$$

From (22), (23) and (24), we deduce that

$$\sum_{i\in\mathsf{B}} r_i^\top r_i^{-1}(\mu) + \sum_{i\in\mathsf{N}} \hat{u}_i^\top u_i^{-1}(\mu) + \sum_{i\in\mathsf{R}}\Big(\frac{r_{i,0}}{r_{i,0}(\mu) + \|\bar{r}_i(\mu)\|} + \frac{\hat{u}_{i,0}}{u_{i,0}(\mu) + \|\bar{u}_i(\mu)\|}\Big) \leq n.$$

When μ tends to zero, the first sum tends to $\sum_{i\in\mathsf{B}} r_i^\top\hat{r}_i^{-1}$ and the second one to $|\mathsf{N}|$. By using the fact that $\hat{r}_{i,0} = \|\bar{\hat{r}}_i\|$ and $\hat{u}_{i,0} = \|\bar{\hat{u}}_i\|$ for $i \in \mathsf{R}$, each term of the third sum tends to $\frac{r_{i,0}}{2\hat{r}_{i,0}} + \frac{1}{2}$. By taking the limit $\mu \downarrow 0$ and using $n = |\mathsf{B}| + |\mathsf{N}| + |\mathsf{R}|$, we obtain

$$\sum_{i\in\mathsf{B}} r_i^\top\hat{r}_i^{-1} + \frac{1}{2}\sum_{i\in\mathsf{R}}\frac{r_{i,0}}{\hat{r}_{i,0}} \leq |\mathsf{B}| + \frac{1}{2}|\mathsf{R}|,$$

which shows by Proposition 4 that $\hat{r}$ is the optimal solution of (20), the analytic center of $\widehat{R}$. □

6 Numerical solution of the friction contact problems

We have implemented the primal-dual interior point algorithm developed by Tütüncü, Toh and Todd [47] and adapted it to our context. This is an extension of the predictor-corrector algorithm of Mehrotra [35] to the solution of a SOCO problem. The main part of the algorithm is the solution of two linear systems that result from the linearization of the equation (18) at the current iterate $(u, r) \in \text{int}(\mathcal{L}^{2n})$. They only differ on the right-hand side and are of the form:

$$\begin{pmatrix} W & -I \\ U & R \end{pmatrix} \begin{pmatrix} d^r \\ d^u \end{pmatrix} = \begin{pmatrix} -Wr - q + u \\ -u \circ r - [d_a^u \circ d_a^r - 2\sigma\mu e] \end{pmatrix} \tag{25}$$

where $U = \text{Arw}(u)$ and $R = \text{Arw}(r)$. The first direction, denoted (d_a^u, d_a^r) and called the *affine scaling direction*, is the solution of (25) without the square bracketed term in the right-hand side. It then satisfies the linear equation

$$u \circ d_a^r + r \circ d_a^u = -u \circ r.$$

The affine scaling direction is a Newton step on the original optimality system (17). The barrier parameter is set to $\mu = \frac{u^\top r}{n}$. The centralization parameter $\sigma \in (0, 1]$ is fixed by comparing the current value of μ with its expected reduction obtained along the affine step. The second direction is a linear combination of the affine scaling direction and of a corrector step, to keep the iterates near the central path. It then satisfies the following linear equation

$$u \circ d^r + r \circ d^u = -u \circ r - d_a^u \circ d_a^r + 2\sigma \tfrac{u^\top r}{n} e.$$

The next iterate is set to $(u^+, r^+) = (u, r) + \alpha(d^u, d^r)$, where α is the largest value in $(0, 1]$ such that

$$(u^+, r^+) \in (1 - \tau)(u, r) + \mathcal{L}^{2n},$$

for some value $\tau \in (0, 1]$ (typically $\tau = 0.995$). Contrary to common practice, different primal and dual steplengths are not taken, because of the first equation in (25), which is linear and includes both primal and dual variables. Indeed, suppose that the current iterate is such that $Wr + q - u = 0$ and that different steplengths $\alpha \neq \alpha'$ are taken with $d^u \neq 0$. We then have

$$W(r + \alpha d^r) + q - (u + \alpha' d^u) = (\alpha - \alpha')d^u.$$

If $d^u \neq 0$, then at the next iteration the residual of the linear equation is no longer zero.

The algorithm will be well defined if the matrix in (25) is nonsingular at each iteration. Its determinant is equal to $\det(W + R^{-1}U) \det R$. Although the vectors r and u are kept inside the interior of the second order cones, and thus R and U are positive definite, the matrix $W + R^{-1}U$ can be singular. This is because the matrix $R^{-1}U$ is not necessarily symmetric, since in general r and u do not commute. The following example is given by [41, p.143]: If $U = \text{Arw}([1, 0.7, 0.7]^\top)$, $R = \text{Arw}([1, 0.8, 0.5]^\top)$ and $W = \text{diag}([0.3, 1, 0]^\top)$, then $\det(W + R^{-1}U) = 0$.

In addition to this singularity issue, there is the problem of symmetry. In IP algorithms, the matrices U and R are diagonal and so the matrix of (25) can be symmetrized, for example by left-multiplying the second row by $-U^{-1}$. See, e.g., [26] for several symmetrization techniques in interior point methods. The major advantages of a symmetric system are a lower factorization cost and an effective control of the inertia of the factorized matrix. Moreover, very efficient codes such as MA57 [24] or MUMPS [10] can be used for this task.

To overcome these problems of singularity and symmetry, a change of variables, called a *scaling scheme*, is applied in order to retrieve a symmetric nonsingular system. The idea is to make a change of variables leaving invariant the Lorentz cone and such that in the new space the vectors u and r commute. However, this change of variable depends on the current iterate and must be done at each iteration. Let $p \in \operatorname{int} \mathcal{K}$ and let $Q_p \succ 0$ be the corresponding quadratic representation. From [9, Theorem 9], we have $Q_p(\mathcal{L}^n) = \mathcal{L}^n$ and $Q_p(\operatorname{int}(\mathcal{L}^n)) = \operatorname{int}(\mathcal{L}^n)$. Let us consider the change of variables

$$\check{r} = Q_{p^{-1}} r \quad \text{and} \quad \hat{u} = Q_p u.$$

The problem (11) becomes

$$\begin{array}{ll} \min\limits_{\check{r}} & \frac{1}{2}\check{r}^\top Q_p W Q_p \check{r} + (Q_p w)^\top \check{r} \\ \text{s.t.} & \check{r} \in \mathcal{L}^n. \end{array} \tag{26}$$

The corresponding perturbed KKT system is

$$Wr + q = u \quad \text{and} \quad \hat{u} \circ \check{r} = 2\mu e,$$

with $(\hat{u}, \check{r}) \in \operatorname{int} \mathcal{L}^{2n}$ The linearization of these equations leads the following linear system:

$$\begin{pmatrix} W & -I \\ \hat{U} Q_{p^{-1}} & \check{R} Q_p \end{pmatrix} \begin{pmatrix} d^r \\ d^u \end{pmatrix} = \begin{pmatrix} -Wr - q + u \\ -\hat{u} \circ \check{r} - [\hat{d}_a^u \circ \check{d}_a^r - 2\sigma\mu e] \end{pmatrix}, \tag{27}$$

where $\hat{U} = \operatorname{Arw}(\hat{u})$ and $\check{R} = \operatorname{Arw}(\check{r})$. The choice of the vector $p \in \operatorname{int} \mathcal{K}$ is made so that $\hat{u}$ and $\check{r}$ commute, which implies that the matrix of the linear system (27) is nonsingular. Indeed, this matrix is nonsingular if and only if $\det(W + (\check{R}Q_p)^{-1}\hat{U}Q_{p^{-1}}) \neq 0$. Since $\hat{U}$ and $\check{R}$ are positive definite, $\hat{u}$ and $\check{r}$ commute, and $Q_{p^{-1}} = Q_p^{-1}$, we have

$$(\check{R}Q_p)^{-1}\hat{U}Q_{p^{-1}} = Q_{p^{-1}}\check{R}^{-1/2}\hat{U}^{1/2}\hat{U}^{1/2}\check{R}^{-1/2}Q_{p^{-1}},$$

which is symmetric and positive definite.

Several choices for the vector p are possible, see [9]. As mentioned in [47], the most efficient scaling technique for the solution of a SOCO problem, is the one using the Nesterov and Todd (NT) direction [40]:

$$p = \left(Q_{u^{1/2}}(Q_{u^{1/2}} r)^{-1/2}\right)^{-1/2} = \left(Q_{r^{-1/2}}(Q_{r^{1/2}} u)^{1/2}\right)^{-1/2}. \tag{28}$$

The main property of the NT direction is that $\hat{u} = \check{r}$, which implies that

$$\hat{U}^{-1}\check{R} = I.$$

The symmetrization of the system (27) is done by left-multiplying the last row by $-Q_p\hat{U}^{-1}$, leading to a symmetric matrix of the form

$$\begin{pmatrix} W & -I \\ -I & -Q_{p^2} \end{pmatrix}.$$

To take advantage of the sparsity of the matrices M and H, the system (25) is considered in the following equivalent augmented form:

$$\begin{pmatrix} M & -H^\top & 0 \\ -H & 0 & I \\ 0 & U & R \end{pmatrix} \begin{pmatrix} d^v \\ d^r \\ d^u \end{pmatrix} = \begin{pmatrix} -Mv - f + H^\top r \\ u - Hv - w \\ -u \circ r - [d_a^u \circ d_a^r - 2\sigma\mu e] \end{pmatrix}. \tag{29}$$

The system (29) can be interpreted as the linearization of the perturbed KKT conditions of the problem (9). Applying the scaling scheme, the system becomes

$$\begin{pmatrix} M & -H^\top & 0 \\ -H & 0 & I \\ 0 & I & Q_{p^2} \end{pmatrix} \begin{pmatrix} d^v \\ d^r \\ d^u \end{pmatrix} = \begin{pmatrix} -Mv - f + H^\top r \\ Hv + w - u \\ -r - [Q_p(\hat{u}^{-1} \circ (\hat{d}_a^u \circ \check{d}_a^r)) - 2\sigma\mu u^{-1}] \end{pmatrix}. \tag{30}$$

A reduction of (30) can be done by eliminating the variable d^u, while keeping the sparse structure. This leads to the reduced symmetric system

$$\begin{pmatrix} M & -H^\top \\ -H & -Q_{p^{-2}} \end{pmatrix} \begin{pmatrix} d^v \\ d^r \end{pmatrix} = \begin{pmatrix} -Mv - f + H^\top r \\ Hv + w + [Q_{p^{-1}}(\hat{u}^{-1} \circ (\hat{d}_a^u \circ \check{d}_a^r)) - 2\sigma\mu r^{-1}] \end{pmatrix}. \tag{31}$$

The big flaw of the scaling strategy is the ill-conditioning of the matrix Q_{p^2} when the solution pair (u, r) approaches an optimal solution. Indeed, suppose that (u^*, r^*) is a primal-dual optimal solution of (11), which satisfies the strict complementarity condition. Let (u, r) be an interior point iterate near the optimal solution and let p be defined by (28). For $i \in \{1, \ldots, n\}$, three situations can occur [18]:

- $u_i^* \in \text{int}(\mathcal{L})$ and $r_i^* = 0$, then all the eigenvalues of $Q_{p_i^2}$ are of order $\mu := u^\top r$;
- $u_i^* = 0$ and $r_i^* \in \text{int}(\mathcal{L})$, then all the eigenvalues of $Q_{p_i^2}$ are of order $1/\mu$;
- $u_i^* \in \text{bd}(\mathcal{L})$, $r_i^* \in \text{bd}(\mathcal{L})$ and $(u_i^*, r_i^*) \neq (0, 0)$, then the largest eigenvalue of $Q_{p_i^2}$ is of order $1/\mu$ and the smallest is of order μ.

To overcome the difficulties due to ill-conditioning, we propose to solve the linear systems (30) and (31) under the following equivalent form:

$$\begin{pmatrix} M & -\hat{H}^\top & 0 \\ -\hat{H} & 0 & I \\ 0 & I & I \end{pmatrix} \begin{pmatrix} d^v \\ \check{d}^r \\ \hat{d}^u \end{pmatrix} = \begin{pmatrix} Mv - f - H^\top r \\ \hat{u} - \hat{H}v - \hat{w} \\ -\check{r} - [\hat{u}^{-1} \circ (\hat{d}_a^u \circ \check{d}_a^r) - 2\sigma\mu\hat{u}^{-1}] \end{pmatrix} \tag{32}$$

and

$$\begin{pmatrix} M & -\hat{H}^\top \\ -\hat{H} & -I \end{pmatrix} \begin{pmatrix} d^v \\ \check{d}^r \end{pmatrix} = \begin{pmatrix} -Mv - f + H^\top r \\ \hat{H}v + \hat{w} + [\hat{u}^{-1} \circ (\hat{d}_a^u \circ \check{d}_a^r) - 2\sigma\mu\hat{u}^{-1}] \end{pmatrix}, \tag{33}$$

where $\hat{H} = Q_p H$, $\hat{w} = Q_p w$, $\hat{d}^u = Q_p d^u$ and $\check{d}^r = Q_{p^{-1}} d^r$. In our numerical experiments, we also consider the reduced system

$$(\hat{H}M^{-1}\hat{H}^\top + I)\check{d}^r = -\hat{H}M^{-1}(f + H^\top r) - \hat{w} - [\hat{u}^{-1} \circ (\hat{d}_a^u \circ \check{d}_a^r) - 2\sigma\mu\hat{u}^{-1}], \tag{34}$$

for which the matrix is positive definite.

Figure 1 shows the behavior of the condition number of the matrices of the six linear systems (29)-(34) along the iterations of the IP algorithm for two examples. The first example (left figure) has a single contact point: $n = 1$, $m = 3$, $M = I$, $w = 0$, $f = [3, 3, 3, 1, -1, -3, 1, -1, -3]^\top$, $H = [D, 0, -D]$ where $D = \text{diag}(1, 0.1, 0.1)$. Since H is of full rank, the primal-dual solution is unique, the optimal reaction and relative velocity vectors are non-zero and on the boundary of the Lorentz cone. The second example is from the Box_Stacks family. There are $n = 69$ contact points

and $m = 450$ degrees of freedom, $H \in \mathbb{R}^{207\times 450}$ and $\text{rank}(H) = 157$. The optimal solution satisfies the strict complementarity condition and $(|\mathsf{B}|, |\mathsf{N}|, |\mathsf{R}|) = (18, 5, 46)$. The matrix of (29) at the end point of the minimization procedure, is nearly rank deficient, there are 15 singular values less than $\sqrt{\epsilon}$, where ϵ is the epsilon machine. With these two examples, it can be seen the matrices in (32) and (33) remain the least ill-conditioned.

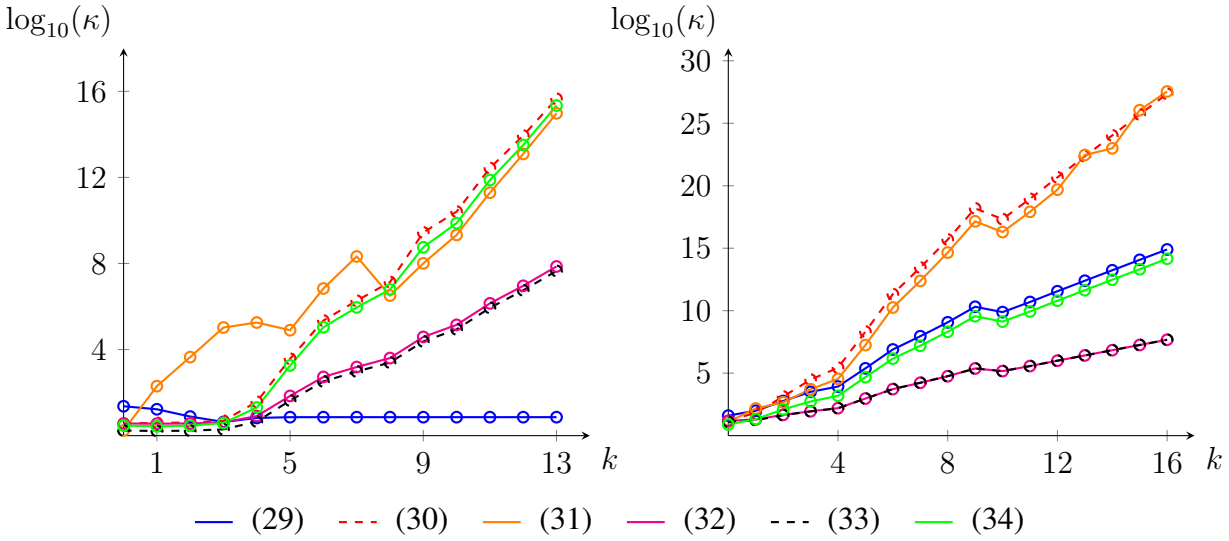

Figure 1: Condition number (κ) of the matrix of the linear systems along the iterations

An advantage of the systems (32) and (33) is that for the whole computations, the quadratic representation matrices are never explicitly built in memory. Only the product of these matrices times a vector must be performed. Indeed, for a pair of vectors $(x, y) \in \mathcal{L}^2$, the product of a vector y by the quadratic representation of x can be done via the formula $Q_x y = 2(x^\top y)x - (\det x)R_d y$. Therefore, the product of a vector by a matrix Q_p, where p is the NT-vector (28), can be done by performing only three products of a quadratic representation matrix by a vector. In the same way, the computation of the inverse or the square root of a vector in the Jordan algebra, is done by using the spectral decomposition of this vector. Moreover, even if the number of cones can be large, the computational cost of a spectral decomposition per cone is very low, because the dimension of a Lorentz cone is only three.

Algorithm 1 One iteration of Mehrotra primal-dual algorithm for solving a FC problem

Parameters: $\eta_1 \in (0,1)$, $\eta_2 \geq 1$, $\eta_3 \geq 1$, $\tau_1 \in (0,1)$, $\tau_2 \in (0, 1-\tau_1)$, tol > 0

if the stopping criterion (15) is satisfied **then** return (v,u,r) as primal-dual solution of (9);
Set $\mu \leftarrow u^\top r/n$;
Compute (d_a^v, d_a^r, d_a^u) solution of (29) (resp. (30)-(34)) without the square bracketed term;
Find the greatest $\alpha_a \in (0,1]$ such that $(u,r) + \alpha_a(d_a^u, d_a^r) \in \mathcal{L}^{2n}$;
Set $\mu_a \leftarrow (u + \alpha_a d_a^u)^\top (r + \alpha_a d_a^r)/n$;
if $\mu > \eta_1$ **then** set $e \leftarrow \max\{1, \eta_2 \alpha_a^2\}$ **else** set $e \leftarrow \eta_3$;
Set $\sigma \leftarrow \min\{1, (\mu_a/\mu)^e\}$;
Compute (d^v, d^r, d^u) solution of (29) (resp. (30)-(34)) including the square bracketed term;
Set $\tau \leftarrow \tau_1 + \alpha_a \tau_2$;
Find the greatest $\alpha \in (0,1]$ such that $(u,r) + \alpha(d^u, d^r) \in (1-\tau)(u,r) + \mathcal{L}^{2n}$;
Set $(v,u,r) \leftarrow (v,u,r) + \alpha(d^v, d^u, d^r)$ and goto 1.

Algorithm 1 details one iteration of the implemented algorithm based on that of SDPT3 [47]. The values of the parameters are fixed to $\eta_1 = 10^{-10}$, $\eta_2 = 3$, $\eta_3 = 1$, $\tau_1 = 0.9$, $\tau_2 = 0.09$ and tol $= 10^{-10}$. The starting point is set as follows: for all $i \in \{1, \ldots, n\}$, $u_i = r_i = (0.1, 0.01, 0.01)^\top$ and $v = M^{-1}(H^\top r + f)$. The experiments were done on 1091 problems of the FCLIB collection [4]. There are seven families of problems, whose dimensions are described in Table 2, where n is the number of three-dimensional cones and m is six times the number of bodies.

Family	# problems	n	m
BoxStacks	28	$[31, 557]$	450
Capsules	200	$[15, 314]$	600
Chute	182	$[3, 3224]$	$[1002, 12672]$
KaplasTower	240	$[48, 899]$	792
PrimitveSoup	200	$[37, 3123]$	6000
Spheres	200	$[344, 4290]$	12000
Spheres1mm	41	$[715, 4213]$	12000

Table 2: Sizes of problems of the FCLIB collection

The linear system (29) is solved by means of a LU factorization, the symmetric ones with a $LDL^\top$ factorization with MA57 [24]. Even for the solution of the positive definite system (34) MA57 is used. Two types of failures are returned during a run:

- Failure 1: the stopping criterion (15) is not satisfied after a maximum of 100 iterations.
- Failure 2: A NaN (Not a Number) is detected during the computation of the new iterate.

Table 3 indicates the number of successes and failures when solving the 1091 problems of the FCLIB collection, with a tolerance fixed to tol $= 10^{-10}$. Each row corresponds to a run of Algorithm 1 with the numerical solution of the indicated linear system. Figure2 shows the corresponding performance profiles [23]. For $\tau \geq 0$, $\rho_s(\tau)$ is the fraction of problems for which the performance of a given version of the algorithm is within a factor 2^τ of the best one. With the system (29) the failures are due to a nearly singular system. In these cases, either the algorithm stalls to a spurious

solution (13 out of 22 cases) or the convergence becomes very slow (9 out of 22 cases). However, it should be noted that for almost 98% of the problems, the "no-scaling" strategy returns an optimal solution. The systems (30) and (31) return a great number of failures of type 2. This is mainly due to the ill-conditioning of the matrix Q_{p^2} when approaching an optimal solution. The reduction of the system worsens the results. Surprisingly, the worse results are with the system (32). A deterioration of the residual of the second linear equation in (32) over the iterations is observed, when the matrix Q_p becomes increasingly ill-conditioned. This leads to a loss of the primal feasibility of the iterates. This is mainly due to the scaling of the linear equation $-Hd^v + d^u = u - Hv - w$. To address this issue, the refinement procedure described in the documentation of MA57 can be used for the solution of (32). We performed a run with a refinement tolerance fixed to the tolerance tol and a maximum of 10 refinement iterations. This leads to only six type 2 failures and no more failure of type 1, but it takes more running time than with (33) as shown by Figure 2. The best performance in terms of robustness is obtained with the system (33). The positive definite system (34) gives a good performance in terms of efficiency, but its robustness is not sufficient, even if refinement is applied.

linear system	# success	# failure 1	# failure 2
(29)	1069	22	0
(30)	947	0	144
(31)	852	5	234
(32)	308	1	782
(32) + refin	1085	0	6
(33)	1091	0	0
(34)	1083	1	7

Table 3: Number of successes and failures when solving the FCLIB problems with tol $= 10^{-10}$

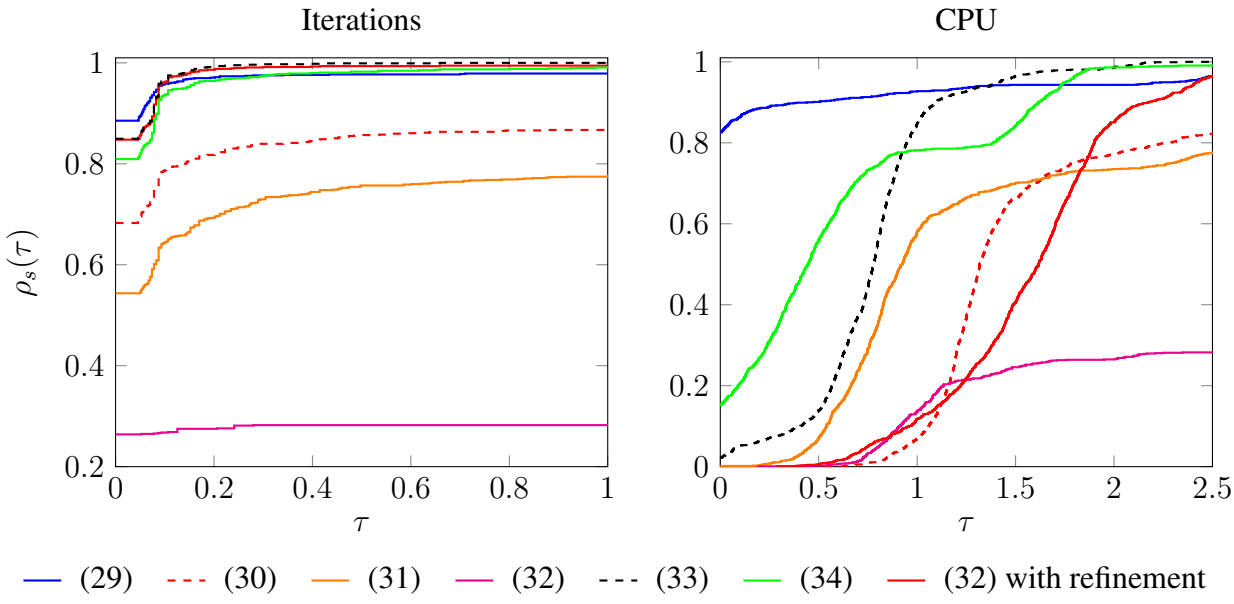

Figure 2: Performance profiles for eight different linear system choices

At last, it must be mentioned that all the computations related to the scaling are performed with the floating-point data type `long double` in C. This data type is also used in the computation of the steplengths. This always results in a better accuracy, although the computing time is slightly higher. Table 4 shows the results of the solution of the FCLIB problems with Algorithm 1, the linear system under the form (33), with a stopping tolerance tol $= 10^{-11}$. The comparison is between the use of the data type `long double` versus `double`. We can see that the number of failures of type 2 is more than twice with the type `double`. For these runs, even in case of a failure, the residual (left-hand term in (15)) is small, meaning that an optimal solution has been found. The column $\max \mathrm{res}$ indicates the maximum value of residual (15) of all the 1091 residuals and shows that the type `long double` allows to obtain a better precision. Even with the data type `double`, all problems are successfully solved for tol $= 10^{-10}$. The last column of this table shows the total computational time to solve all the FCLIB problems with this tolerance. It can be seen that the increase in computational time is less than 10% with `long double`.

	# success	# failure 1	# failure 2	$\max \mathrm{res}$	cpu (sec)
double	1047	5	39	9e-9	616
long double	1069	5	17	6e-11	675

Table 4: Performance comparison `double` versus `long double` when solving the FCLIB problems with system (33) and tol $= 10^{-11}$

7 Rolling friction contact problem

We now consider the solution of (4) in the framework of the RF model ($d = 5$) defined by the *rolling friction cones* (7) and (8). The main difficulty is that an elementary cone $\mathcal{R}_i$ is no more self-dual and therefore not symmetric. There is no Jordan product such that $\mathcal{R}_i$ is a cone of squares with respect to this product. A potential approach is to transform the cone of constraints related to problem (9) into the product of Lorentz cones by means of the following usual trick of introducing artificial variables. For real numbers, $a \geq b + c$ if and only if there exist two real numbers $t \geq b$ and $t' \geq c$ such that $a = t + t'$. By setting $u_{i,0} = \bar{t}_i + \tilde{t}_i$ for all $i \in \{1, \ldots, n\}$, the primal-dual pair of problems (9)-(11) can be rewritten as

$$\begin{array}{ll} \min\limits_{v,z} & \frac{1}{2} v^\top M v + f^\top v \\ \text{s.t.} & Hv + w = Jz, \\ & z \in \mathcal{L}^{2n}, \end{array} \tag{35}$$

and

$$\begin{array}{ll} \min\limits_{r} & \frac{1}{2} r^\top W r + q^\top r \\ \text{s.t.} & J^\top r \in \mathcal{L}^{2n}, \end{array} \tag{36}$$

where $z = (z_1^\top, \ldots, z_n^\top)^\top$ and $z_i = (\bar{t}_i, \bar{u}_i^\top, \tilde{t}_i, \tilde{u}_i^\top)^\top \in \mathbb{R}^{d+1}$, $J \in \mathbb{R}^{nd \times n(d+1)}$ is a block diagonal matrix with n blocks of the form

$$\jmath = \begin{pmatrix} 1 & & 1 & \\ & I & & \\ & & & I \end{pmatrix} \in \mathbb{R}^{d \times (d+1)}.$$

For all $i \in \{1, \ldots, n\}$, we have $\jmath z_i = (\bar{t}_i + \tilde{t}_i, \bar{u}_i^\top, \tilde{u}_i^\top)^\top$ and $\jmath^\top r_i = (r_{i0}, \bar{r}_i^\top, r_{i0}, \tilde{r}_i^\top)^\top$.

The perturbed KKT system associated of problem (36) is then

$$\begin{array}{rcl} Wr + q & = & Jz, \\ z \circ J^\top r & = & 2\mu e, \\ (z, J^\top r) & \in & \text{int}(\mathcal{L}^{4n}). \end{array} \tag{37}$$

This optimality system is associated to the barrier problem

$$\min_r \quad \varphi_\mu(r) := \tfrac{1}{2} r^\top W r + q^\top r - \mu \textstyle\sum_{i=1}^n \log(\det(r_{i,0}, \bar{r}_i) \det(r_{i,0}, \tilde{r}_i)). \tag{38}$$

As for Proposition 2 and by coercivity of the barrier function, under Assumption 1, for all $\mu > 0$ the system (37) has a unique solution such that $(z(\mu), J^\top r(\mu)) \in \text{int}(\mathcal{L}^{4n})$. Although the optimal solution of (35) is not unique, because J is non-injective, it can be shown, like for Proposition 3, that the central path converges to a relative interior point of the primal-dual optimal set. Under the hypothesis of strict complementarity, it can also be shown that the central path $r(\cdot)$ converges to the analytic center of the dual optimal set. For the sake of completeness, we state the result, but without proof in order to lighten the paper.

Assumption 3. *There exists a solution* (z, r) *of* (37) *with* $\mu = 0$*, such that* $z + J^\top r \in$ *int*$(\mathcal{L}^{2n})$.

The strict complementarity hypothesis can be equivalently reformulated relatively to the original problems (9) and (11), thanks to the orthogonality condition (12). Let (z, r) be a solution of (37) with $\mu = 0$. Then $(v, u = Jz)$ is the unique optimal solution of (9). Assumption 3 is satisfied if and only if for all $i \in \{1, \ldots, n\}$, one of the following three assertions is satisfied:

- $r_{i0} > \max\{\|\bar{r}_i\|, \|\tilde{r}_i\|\}$ and $u_i = 0$,
- $u_{i0} > \|\bar{u}_i\| + \|\tilde{u}_i\|$ and $r_i = 0$,
- $r_{i0} = \max\{\|\bar{r}_i\|, \|\tilde{r}_i\|\} > 0$ and $u_{i0} = \|\bar{u}_i\| + \|\tilde{u}_i\| > 0$.

As in Section 5, under the strict complementarity assumption, the index set $\{1, \ldots, n\}$ can be partitioned into two partitions $(\bar{\mathsf{B}}, \bar{\mathsf{N}}, \bar{\mathsf{R}})$ and $(\tilde{\mathsf{B}}, \tilde{\mathsf{N}}, \tilde{\mathsf{R}})$, whose definition is a direct extension of the previous one to the current framework. Let $\widehat{Z}$ and $\widehat{R}$ be the primal and dual optimal sets of problems (35) and (36). The first partition is defined as follows, the second in a similar way.

$$\bar{\mathsf{B}} = \{i : \exists r \in \widehat{R}, (r_{i,0}, \bar{r}_i) \in \text{int}(\mathcal{L})\}, \bar{\mathsf{N}} = \{i : \exists z \in \widehat{Z}, (\bar{t}_i, \bar{u}_i) \in \text{int}(\mathcal{L})\},$$

$$\bar{\mathsf{R}} = \{i : \exists (z, r) \in \widehat{Z} \times \widehat{R}, ((\bar{t}_i, \bar{u}_i), (r_{i,0}, \bar{r}_i)) \in \text{bd}(\mathcal{L}^2) \backslash \{0\}\}.$$

We can then define the analytic center of the optimal set $\widehat{R}$ as follows. If $\widehat{R}$ is reduced to a singleton, then it is this point, otherwise it is the minimum of the problem

$$\min_{r \in \text{ri}(\widehat{\mathsf{R}})} -\sum_{i \in \bar{\mathsf{B}}} \log\det(r_{i,0}, \bar{r}_i) - \sum_{i \in \tilde{\mathsf{B}}} \log\det(r_{i,0}, \tilde{r}_i) - \sum_{i \in \bar{\mathsf{R}}, i \in \tilde{\mathsf{R}}} \log r_{i,0}. \tag{39}$$

The analytic center can be characterized like in Proposition 4, by which it can be shown that Theorem 1 still holds for the rolling friction framework.

Algorithm 1 is modified in order to solve a RF problem. This is described by Algorithm 2.

Algorithm 2 One iteration of Mehrotra primal-dual algorithm for solving a RF problem

Parameters: $\eta_1 \in (0,1)$, $\eta_2 \geq 1$, $\eta_3 \geq 1$, $\tau_1 \in (0,1)$, $\tau_2 \in (0, 1-\tau_1)$, tol > 0

1: **if** (15) with $u = Jz$ is satisfied **then** return (v, z, r) as primal-dual solution of (35);
2: Set $\mu \leftarrow r^\top Jz/n$;
3: Compute (d_a^v, d_a^r, d_a^z) solution of (40) (resp. (41)-(42)) without the square bracketed term;
4: Find the greatest $\alpha_a \in (0,1]$ such that $(z, J^\top r) + \alpha_a(d_a^z, J^\top d_a^r) \in \mathcal{L}^{4n}$;
5: Set $\mu_a \leftarrow (r + \alpha_a d_a^r)^\top J(z + \alpha_a d_a^z)/n$;
6: **if** $\mu > \eta_1$ **then** set $e \leftarrow \max\{1, \eta_2 \alpha_a^2\}$ **else** set $e \leftarrow \eta_3$;
7: Set $\sigma \leftarrow \min\{1, (\mu_a/\mu)^e\}$;
8: Compute (d^v, d^r, d^z) solution of (40) (resp. (41)-(42)) including the square bracketed term;
9: Set $\tau \leftarrow \tau_1 + \alpha_a \tau_2$;
10: Find the greatest $\alpha \in (0,1]$ such that $(z, J^\top r) + \alpha(d^z, J^\top d^r) \in (1-\tau)(u, r) + \mathcal{L}^{4n}$;
11: Set $(v, z, r) \leftarrow (v, z, r) + \alpha(d^v, d^z, d^r)$ and goto 1.

The linear system solved at each iteration is obtained by linearizing the system (37). It is reformulated under the form of the following augmented system

$$\begin{pmatrix} M & -H^\top & 0 \\ -H & 0 & J \\ 0 & ZJ^\top & R \end{pmatrix} \begin{pmatrix} d^v \\ d^r \\ d^z \end{pmatrix} = \begin{pmatrix} -Mv - f + H^\top r \\ Hv + w - Jz \\ -z \circ J^\top r - [d_a^z \circ (J^\top d_a^r) - 2\sigma\mu e] \end{pmatrix}, \tag{40}$$

where $Z = \mathrm{Arw}(z)$, $R = \mathrm{Arw}(J^\top r)$ and $\mu = \frac{r^\top Jz}{n}$. As in Algorithm 1, the affine scaling direction (d_a^v, d_a^r, d_a^z) is the solution of (40) without the square bracketed term in the right-hand side, while the full step (d^v, d^r, d^z) is the solution of the complete system.

The scaling strategy is similar to that described in Section 6. The NT direction p is defined by the formula (28) where u and r are respectively replaced by z and $J^\top r$. The change of variables is done by setting

$$\hat{z} := Q_p z \quad \text{and} \quad \check{y} := Q_{p^{-1}} J^\top r.$$

Recall that $\hat{z} = \check{y}$, which allows to symmetrize the linear system under the form

$$\begin{pmatrix} M & -H^\top & 0 \\ -H & 0 & J \\ 0 & J^\top & Q_{p^2} \end{pmatrix} \begin{pmatrix} d^v \\ d^r \\ d^z \end{pmatrix} = \begin{pmatrix} -Mv - f + H^\top r \\ Hv + w - Jz \\ -J^\top r - [Q_p(\hat{z}^{-1} \circ (\hat{d}_a^z \circ \check{d}_a^y)) - 2\sigma\mu z^{-1}] \end{pmatrix}. \tag{41}$$

Because of the ill-conditioning of the matrix Q_{p^2}, an equivalent form of (41) has been considered:

$$\begin{pmatrix} M & -H^\top & 0 \\ -H & 0 & JQ_{p^{-1}} \\ 0 & Q_{p^{-1}} J^\top & I \end{pmatrix} \begin{pmatrix} d^v \\ d^r \\ \hat{d}^z \end{pmatrix} = \begin{pmatrix} -Mv - f + H^\top r \\ Hv + w - Jz \\ -\check{y} - [\hat{z}^{-1} \circ (\hat{d}_a^z \circ \check{d}_a^y - 2\sigma\mu e)] \end{pmatrix}, \tag{42}$$

where $\hat{d}^z = Q_p d^z$. Figure 3 shows the condition number of the three matrices (40)-(42) along the iterations of the numerical resolution of a RF problem. It can be seen that the system (42) is better conditioned than (41).

We also tried several reductions to a 2×2 form as in (31) or (33), leading to matrices of the form

$$\begin{pmatrix} M & -H^\top \\ -H & -JQ_{p^{-2}} J^\top \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} M & \hat{H}^\top \\ -\hat{H} & -I \end{pmatrix}, \tag{43}$$

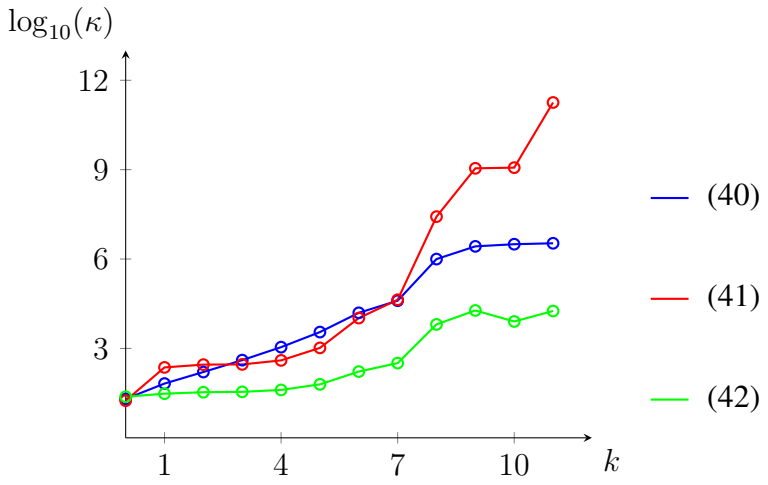

Figure 3: Condition number (κ) of the matrix of the linear systems along the iterations of the Algorithm 2 when solving the problem PrimitiveSoup-ndof-6000-nc-37-0 with $n = 37$ contact points.

where $\hat{H} = P^{-1}H$ and $PP^\top = JQ_{p^{-2}}J^\top$. We also tried several ways to compute the matrix P, by performing a Cholesky factorization or by directly exploiting the structure of a quadratic representation matrix. Despite such a reduction and in contrast to the results obtained with the RF problems, the numerical performance has not been improved. Moreover, the computation of the matrix P increases the overall computational cost, without any real improvement. We also observe the same for the 1×1 system like (34).

The numerical tests were carried out on 526 RF problems of the FCLIB family [4], whose characteristics are in Table 5. The numerical results and performances of Algorithm 2 with the three linear systems described previously, are reported in Table 6 and Figure 4. These results show that, with a tolerance tol $= 10^{-10}$, the choice of system (42) gives the best performance. But the performance gap between the systems with the NT scaling is smaller than those observed for the FC problems. It can also be observed that, as for the RF problems, without scaling the failures are of type 1, while with NT scaling the failures only occur when a NaN is detected. Moreover, in the latter case the maximum value of the residual (15) (right column of Table 6), shows that the stopping point of the algorithm is nearly optimal.

Family	# problems	n	m
Chute	155	$[4, 1372]$	$[768, 6528]$
PrimitiveSoup	171	$[37, 2269]$	6000
SpherePile	200	$[2, 542]$	$[24, 1500]$

Table 5: Sizes of rolling friction problems of the FCLIB collection

linear system	# success	# failure 1	# failure 2	max res
(40)	391	135	0	3e-3
(41)	451	0	75	4e-9
(42)	508	0	18	5e-9

Table 6: Number of successes and failures when solving the RF problems with tol $= 10^{-10}$

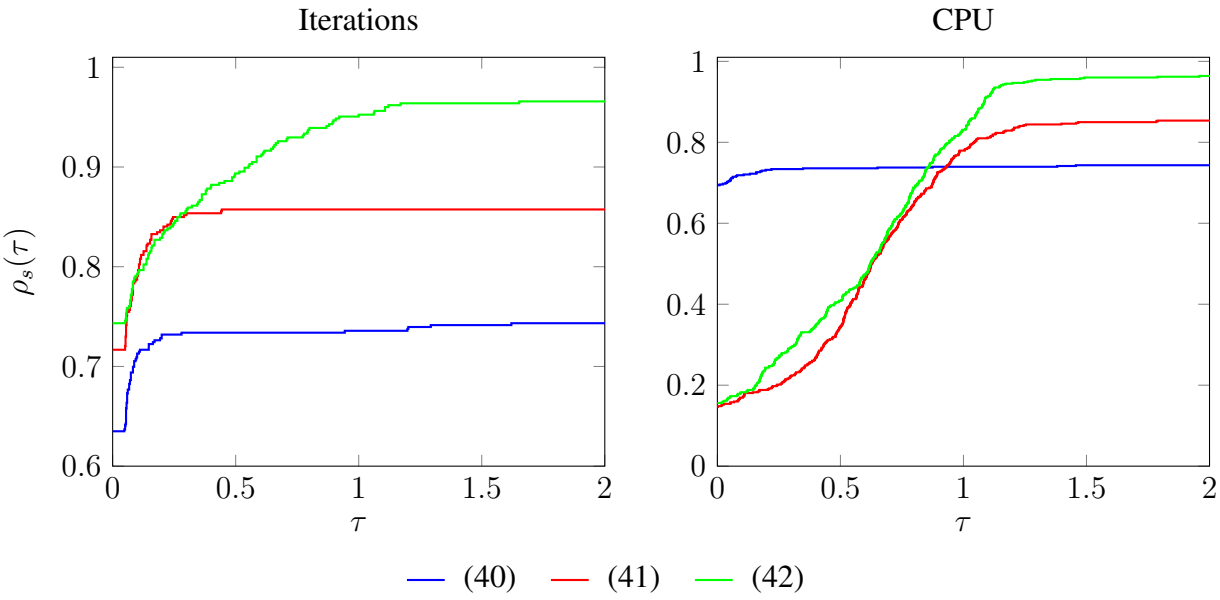

Figure 4: Performance profiles for the three linear systems for the RF model

8 Conclusion

At the beginning of this work, solving the relaxed convex form of friction contact models seemed very simple. But after our preliminary experiences with existing softwares and our first implementation of the primal-dual algorithm of Mehrotra, we were a bit confused by the lack of efficiency and robustness. The Nesterov-Todd scaling strategy is a wonderful theoretical tool, but numerically very painful. As the iterates approach the boundary of the second order cones, the conditioning of the linear system explodes, the iterates get stuck on the boundary and divisions by zero occur, which produces NaN and thus an emergency stop. We have therefore reviewed a large number of equivalent formulations of the linear system and found that the one that gives the best results, or shall we say least bad, is the one in which the quadratic representation matrix that allows scaling is not a direct component of the matrix of the linear system. In addition, particular attention must be paid to the way in which the matrix-vector products are done in order to build the system to be factorized. Unfortunately, for rolling friction problems, we have not been able to find a formulation that is as efficient and robust as for friction cones. The accuracy we have achieved is somewhat a bit lower. Nevertheless, the accuracies and calculation times we have achieved for both models seem to us quite suitable and can be used for real applications. The next step in this research work is to extend the primal-dual

algorithm to solve the original model (1), which is non-smooth and is not the optimality system of an optimization problem.

References

[1] Vincent Acary, Paul Armand, and Hoang Minh Nguyen. High-accuracy computation of rolling friction contact problems. In *2022 9th NAFOSTED Conference on Information and Computer Science (NICS)*, to appear.

[2] Vincent Acary and Franck Bourrier. Coulomb friction with rolling resistance as a cone complementarity problem. *European Journal of Mechanics - A/Solids*, 85:104046, 2021.

[3] Vincent Acary, Maurice Brémond, and Olivier Huber. *On Solving Contact Problems with Coulomb Friction: Formulations and Numerical Comparisons*, pages 375–457. Springer International Publishing, Cham, 2018.

[4] Vincent Acary, Maurice Brémond, Tomasz Koziara, and Franck Pérignon. FCLIB: a collection of discrete 3D Frictional Contact problems. Technical Report RT-0444, INRIA, February 2014.

[5] Vincent Acary and Bernard Brogliato. *Numerical methods for nonsmooth dynamical systems: applications in mechanics and electronics*. Springer Science & Business Media, 2008.

[6] Vincent Acary and Florent Cadoux. Applications of an existence result for the coulomb friction problem. In *Recent Advances in Contact Mechanics*, pages 45–66. Springer, 2013.

[7] Vincent Acary, Florent Cadoux, Claude Lemaréchal, and Jérôme Malick. A formulation of the linear discrete Coulomb friction problem via convex optimization. *ZAMM Z. Angew. Math. Mech.*, 91(2):155–175, 2011.

[8] P. Alart and A. Curnier. A mixed formulation for frictional contact problems prone to Newton like solution method. *"Computer Methods in Applied Mechanics and Engineering*, 92(3):353–375, 1991.

[9] F. Alizadeh and D. Goldfarb. Second-order cone programming. *Math. Program.*, 95(1, Ser. B):3–51, 2003. ISMP 2000, Part 3 (Atlanta, GA).

[10] P.R. Amestoy, I. S. Duff, J. Koster, and J.-Y. L'Excellent. A fully asynchronous multifrontal solver using distributed dynamic scheduling. *SIAM Journal on Matrix Analysis and Applications*, 23(1):15–41, 2001.

[11] Alfred Auslender and Marc Teboulle. *Asymptotic cones and functions in optimization and variational inequalities*. Springer Monographs in Mathematics. Springer-Verlag, New York, 2003.

[12] Hande Y. Benson and Robert J. Vanderbei. Solving problems with semidefinite and related constraints using interior-point methods for nonlinear programming. *Math. Program.*, 95(2, Ser. B):279–302, 2003. Computational semidefinite and second order cone programming: the state of the art.

[13] Antonio Bicchi and Vijay Kumar. Robotic grasping and contact: A review. In *Proceedings 2000 ICRA. Millennium conference. IEEE international conference on robotics and automation. Symposia proceedings (Cat. No. 00CH37065)*, volume 1, pages 348–353. IEEE, 2000.

[14] Bernard Brogliato and B Brogliato. *Nonsmooth mechanics*, volume 3. Springer, 1999.

[15] Martin Buss, Leonid Faybusovich, and John B Moore. Recursive algorithms for real-time grasping force optimization. In *Proceedings of International Conference on Robotics and Automation*, volume 1, pages 682–687. IEEE, 1997.

[16] Martin Buss, Hideki Hashimoto, and John B Moore. Dextrous hand grasping force optimization. *IEEE transactions on robotics and automation*, 12(3):406–418, 1996.

[17] Richard H. Byrd, Jorge Nocedal, and Richard A. Waltz. KNITRO: An integrated package for nonlinear optimization. In *Large-scale nonlinear optimization*, volume 83 of *Nonconvex Optim. Appl.*, pages 35–59. Springer, New York, 2006.

[18] Zhi Cai and Kim-Chuan Toh. Solving second order cone programming via a reduced augmented system approach. *SIAM J. Optim.*, 17(3):711–737, 2006.

[19] Bernard Cambou, Michel Jean, and Farhang Radjaï. *Micromechanics of granular materials*. John Wiley & Sons, 2013.

[20] Timothy A Davis. *Direct methods for sparse linear systems*. SIAM, 2006.

[21] E. de Klerk, C. Roos, and T. Terlaky. Initialization in semidefinite programming via a self-dual skew-symmetric embedding. *Oper. Res. Lett.*, 20(5):213–221, 1997.

[22] E. de Klerk, C. Roos, and T. Terlaky. Infeasible-start semidefinite programming algorithms via self-dual embeddings. In *Topics in semidefinite and interior-point methods (Toronto, ON, 1996)*, volume 18 of *Fields Inst. Commun.*, pages 215–236. Amer. Math. Soc., Providence, RI, 1998.

[23] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Math. Program.*, 91(2, Ser. A):201–213, 2002.

[24] Iain S. Duff. Ma57—a code for the solution of sparse symmetric definite and indefinite systems. *ACM Trans. Math. Software*, 30:118–144, 2004.

[25] David M. Gay. The AMPL modeling language: an aid to formulating and solving optimization problems. In *Numerical analysis and optimization*, volume 134 of *Springer Proc. Math. Stat.*, pages 95–116. Springer, Cham, 2015.

[26] Alexandre Ghannad, Dominique Orban, and Michael A. Saunders. Linear systems arising in interior methods for convex optimization: a symmetric formulation with bounded condition number. *Optimization Methods and Software*, 0(0):1–26, 2021.

[27] M. Halická, E. de Klerk, and C. Roos. On the convergence of the central path in semidefinite optimization. *SIAM J. Optim.*, 12(4):1090–1099, 2002.

[28] M. Halická, E. De Klerk, and C. Roos. Limiting behavior of the central path in semidefinite optimization. *Optim. Methods Softw.*, 20(1):99–113, 2005.

[29] Li Han, Jeffrey C Trinkle, and Zexiang X Li. Grasp analysis as linear matrix inequality problems. *IEEE Transactions on Robotics and Automation*, 16(6):663–674, 2000.

[30] Kenneth L Johnson. Contact mechanics. *Proceedings of the Institution of Mechanical Engineers*, 223(J3):254, 2009.

[31] Jan Kleinert, Bernd Simeon, and Martin Obermayr. An inexact interior point method for the large-scale simulation of granular material. *Computer Methods in Applied Mechanics and Engineering*, 278:567–598, 2014.

[32] Radek Kučera, Jitka Machalová, Horymír Netuka, and Pavel Ženčák. An interior-point algorithm for the minimization arising from 3d contact problems with friction. *Optimization methods and software*, 28(6):1195–1217, 2013.

[33] Olvi L. Mangasarian. *Nonlinear programming*. McGraw-Hill Book Co., New York-London-Sydney, 1969.

[34] Dario Mangoni, Alessandro Tasora, and Rinaldo Garziera. A primal–dual predictor–corrector interior point method for non-smooth contact dynamics. *Computer Methods in Applied Mechanics and Engineering*, 330:351–367, 2018.

[35] Sanjay Mehrotra. On the implementation of a primal-dual interior point method. *SIAM J. Optim.*, 2(4):575–601, 1992.

[36] Daniel Melanz, Luning Fang, Paramsothy Jayakumar, and Dan Negrut. A comparison of numerical methods for solving multibody dynamics problems with frictional contact modeled via differential variational inequalities. *Computer Methods in Applied Mechanics and Engineering*, 320:668–693, 2017.

[37] John Milnor. *Singular points of complex hypersurfaces*. Annals of Mathematics Studies, No. 61. Princeton University Press, Princeton, N.J.; University of Tokyo Press, Tokyo, 1968.

[38] Ali Mohammad-Nezhad and Tamás Terlaky. On the sensitivity of the optimal partition for parametric second-order conic optimization. *Math. Program.*, 189(1-2, Ser. B):491–525, 2021.

[39] Renato D. C. Monteiro and Fangjun Zhou. On the existence and convergence of the central path for convex programming and some duality results. *Comput. Optim. Appl.*, 10(1):51–77, 1998.

[40] Yu. E. Nesterov and M. J. Todd. Self-scaled barriers and interior-point methods for convex programming. *Math. Oper. Res.*, 22(1):1–42, 1997.

[41] Jiming Peng, Cornelis Roos, and Tamás Terlaky. *Self-regularity: a new paradigm for primal-dual interior-point algorithms*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2002.

[42] Friedrich Pfeiffer and Christoph Glocker. *Multibody dynamics with unilateral contacts*. John Wiley & Sons, 1996.

[43] Valentin L Popov et al. *Contact mechanics and friction*. Springer, 2010.

[44] Farhang Radjai and Frédéric Dubois. *Discrete-element modeling of granular materials*. Wiley-Iste, 2011.

[45] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[46] Brad Skarpness and V. A. Sposito. A note on Gordan's theorem over cone domains. *Internat. J. Math. Math. Sci.*, 5(4):809–812, 1982.

[47] R. H. Tütüncü, K. C. Toh, and M. J. Todd. Solving semidefinite-quadratic-linear programs using SDPT3. *Math. Program.*, 95(2, Ser. B):189–217, 2003. Computational semidefinite and second order cone programming: the state of the art.

[48] Peter Wriggers and Tod A Laursen. *Computational contact mechanics*, volume 2. Springer, 2006.

A Proof of Proposition 1

Before giving a proof of Proposition 1, we propose some equivalent formulations of Assumption 1. They are proved thanks to the following known lemma [46, Lemma 2], called Tucker's theorem of the alternative in the case where K is the nonnegative orthant, see [33, p. 29]. The proof can be made by applying a separation theorem of convex sets, see, e.g., [45, Theorem 11.3].

Lemma 1. *Let $K \subset \mathbb{R}^n$ be a closed pointed convex cone, $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times n}$. One and only one of the following statements is true.*

(i) *There exists a non-zero $x \in K$ such that $Ax = 0$ and $Bx \leq 0$.*

(ii) *There exists $(y, z) \in \mathbb{R}^m \times \mathbb{R}^p_+$ such that $A^\top y + B^\top z \in \mathrm{int}(K^*)$.*

The following lemma provides equivalent formulations of Assumption 1.

Lemma 2. *Let $\mathcal{F}$ be the product of second order cones of the form* (5) *or* (7)*. Let $M \in \mathbb{R}^{m\times m}$ be symmetric and positive-definite, $f \in \mathbb{R}^m$, $H \in \mathbb{R}^{nd\times m}$, $w \in \mathbb{R}^{nd}$ and $W = HM^{-1}H^\top$. The following four assertions are equivalent.*

(i) *There exists $v \in \mathbb{R}^m$ such that $Hv + w \in \mathrm{int}(\mathcal{F}^*)$.*
(ii) *There exists $(v, t) \in \mathbb{R}^m \times \mathbb{R}_+$ such that $Hv + tw \in \mathrm{int}(\mathcal{F}^*)$.*
(iii) *There does not exist a nonzero $d \in \mathcal{F}$, such that $H^\top d = 0$ and $w^\top d \leq 0$.*
(iv) *There does not exist a nonzero $d \in \mathcal{F}$, such that $Wd = 0$ and $q^\top d \leq 0$.*

Proof. The implication (i) $\Rightarrow$ (ii) is obvious. The equivalence (ii) $\Leftrightarrow$ (iii) is a direct consequence of the fact that $\mathcal{F}^*$ is a closed pointed convex cone and of Lemma 1. The equivalence (iii) $\Leftrightarrow$ (iv) follows from the definitions of W and q in (10), and of the positive definiteness of M. It remains to prove that (ii) $\Rightarrow$ (i).

Let $(v, t) \in \mathbb{R}^m \times \mathbb{R}_+$ such that $Hv + tw \in \mathrm{int}(\mathcal{F}^*)$. If $t > 0$, set $v' = \frac{1}{t}v$. Since $Hv + tw = t(Hv' + w)$ and $\mathrm{int}(\mathcal{F}^*)$ is a cone, $Hv' + w \in \mathrm{int}(\mathcal{F}^*)$. Suppose now that $t = 0$. There exists $\epsilon > 0$

such that $Hv + B_\epsilon \subset \mathcal{F}^*$, where B_ϵ is the open ball centered at 0 with radius ϵ. Since $\mathcal{F}^*$ is a cone, for all $t > 0$, $tHv + B_{t\epsilon} \subset \mathcal{F}^*$. Let us choose $t > \frac{1}{\epsilon}\|w\|$. We then have $tHv + w \in \mathrm{int}(\mathcal{F}^*)$, which proves (i). □

We can prove now that the Slater's assumption is equivalent to the non-emptiness and compactness of the optimal set of the reduced problem (11).

Proof of Proposition 1. The assertions (i) and (ii) are direct consequences of the weak duality and of the strong convexity of the objective function of (9). The outcome (iii) can be proved by means of some useful tools from asymptotic analysis in convex optimization [11]. Let us define the function $g(r) = \frac{1}{2}r^\top Wr + q^\top r + \delta_{\mathcal{F}}(r)$, where $\delta_{\mathcal{F}}$ is the indicator function of the set $\mathcal{F}$. The asymptotic function of g is defined by

$$g_\infty(s) = \sup_{t>0} \frac{g(ts) - g(0)}{t} = \begin{cases} q^\top s & \text{if } s \in \mathcal{F} \text{ and } Ws = 0, \\ \infty & \text{otherwise,} \end{cases}$$

for $s \in \mathbb{R}^{nd}$, see [11, Proposition 2.5.2]. The optimal set $\widehat{R}$ coincides with the set of minima of g. It is non-empty and compact if and only if g is coercive, that is $g_\infty(s) > 0$ for all nonzero $s \in \mathbb{R}^{nd}$ [11, Proposition 3.1.3]. It follows that $\widehat{R}$ is nonempty and compact if and only if there is no nonzero $s \in \mathcal{F}$, such that $Ws = 0$ and $q^\top s \leq 0$. By Lemma 2, this is equivalent to the Slater hypothesis. □

B Numerical results by differentiable optimization solver

For each problem, the number of contact points (n), that is the number of 3-dimensional Lorentz cones, is indicated in bold. The number of degrees of freedom (m) is indicated in brackets.

	iter	time	$u^\top r$	$\mathrm{err}(r,u)$		iter	time	$u^\top r$	$\mathrm{err}(r,u)$
PER	316	4.7	1.5e-04	9.0e-04	PER	940	1.9	1.1e-04	7.5e−04
SQU	449	20.7	-5.8e-06	5.8e-05	SQU	392	4.4	-1.0e-05	4.6e-05
EXP	253	7.4	-8.3e-06	9.0e-05	EXP	523	2.9	-1.5e-05	6.3e-05
RAT	88	1.1	8.9e-08	6.7e-06	RAT	179	0.27	5.0e-08	2.1e-05
GFC3D	31	0.14	5.3e-11	1.2e-07	GFC3D	19	0.05	9.5e-11	2.5e-07

Table 7: Box_Stacks-i1000-**557**-13 (450) – Capsules-i097201-**313**-31328 (600)

	iter	time	$u^\top r$	$\mathrm{err}(r,u)$		iter	time	$u^\top r$	$\mathrm{err}(r,u)$
PER	15	0.02	5.2e-11	2.6e-07	PER	278	15.7	8.5e-06	4.8e-04
SQU	20	0.18	1.4e-12	1.7e-10	SQU	427	326.5	-2.2e-08	3.6e-05
EXP	22	0.49	8.1e-13	1.0e-10	EXP	475	337.3	-5.9e-07	4.6e-05
RAT	16	0.34	1.6e-10	2.0e-08	RAT	36	1.8	6.4e-07	5.0e-05
GFC3D	12	0.11	1.6e-11	2.2e-09	GFC3D	40	4.8	8.3e-11	3.1e-07

Table 8: KaplasTower-i000249-**898**-29491 (792) – Spheres-i1000-**4290**-14670 (12000)

	iter	time	$u^\top r$	$\mathrm{err}(r,u)$		iter	time	$u^\top r$	$\mathrm{err}(r,u)$
PER	10000	533.3	7.4e-01	3.8e-01	PER	10000	488.7	5.4e-01	1.7e-03
SQU	376	513.5	-5.7e-07	5.7e-04	SQU	597	504.4	1.9e-04	5.3e-04
EXP	405	703.9	-9.4e-08	2.2e-04	EXP	280	187.8	1.1e-04	3.3e-04
RAT	170	6.5	5.0e-07	1.4e-04	RAT	283	10.0	1.2e-05	4.1e-04
GFC3D	44	1.7	5.6e-11	1.6e-06	GFC3D	47	2.971	6.2e-11	3.8e-08

Table 9: PrimitiveSoup-ndof-6000-nc-**3123**-2922 (6000) – Chute-ndof-12672-nc-**3224**-3890 (12672)

C Euclidean Jordan algebra

Let us consider the set $\mathcal{K} = \prod_{i=1}^{n} \mathcal{K}_i$ where $\mathcal{K}_i$ is an n_i-dimensional Lorentz cone defined by

$$\mathcal{K}_i = \left\{ x = (x_{i0}, \bar{x}_i^\top)^\top \in \mathbb{R} \times \mathbb{R}^{n_i - 1} : \|\bar{x}_i\| \leq x_{i0} \right\}.$$

Let $N = \sum_{i=1}^{n}$. For $x \in \mathbb{R}^N$, we denote $x = (x_1, \ldots, x_n)$, where $x_i = (x_{i0}, \bar{x}_i)$. For two vectors x and y in $\mathbb{R}^N$, the Jordan product is defined by

$$x \circ y = \begin{pmatrix} x_1 \circ y_1 \\ \vdots, \\ x_n \circ y_n \end{pmatrix}, \text{ where } x_i \circ y_i = \begin{pmatrix} x_i^\top y_i \\ x_{i0}\bar{y}_i + y_{i0}\bar{x}_i \end{pmatrix}, \text{ for } i = 1, \ldots, n.$$

Let $x \in \mathbb{R}^N$ and $x^2 = x \circ x$. A fundamental property of the Jordan algebra for interior-point algorithms, is that the Lorentz cone is the cone of squares, that is $\mathcal{K} = \{x^2 : x \in \mathbb{R}^N\}$, see [9, pp. 18–19].

For matrices A and B, we define

$$A \oplus B := \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}.$$

For $x \in \mathbb{R}^N$, the arrow-shaped matrix is defined by

$$\mathrm{Arw}(x) = \mathrm{Arw}(x_1) \oplus \ldots \oplus \mathrm{Arw}(x_n), \quad \text{where } \mathrm{Arw}(x_i) = \begin{pmatrix} x_{i0} & \bar{x}_i^\top \\ \bar{x}_i & x_{i0} I \end{pmatrix} \text{ for } i = 1, \ldots, n.$$

For $i \in \{1, \ldots, n\}$, let $e_i = (1, 0) \in \mathbb{R} \times \mathbb{R}^{n_i - 1}$ and let $e = (e_1, \ldots, e_n) \in \mathbb{R}^N$ be the neutral element of the Jordan product. For vectors x and y in $\mathbb{R}^N$, one has

$$x \circ y = \mathrm{Arw}(x) y = \mathrm{Arw}(x)\mathrm{Arw}(y) e$$

For $i \in \{1, \ldots, n\}$, let $\det(x_i) = x_{i0}^2 - \|\bar{x}_i\|^2$ be the determinant of $x_i \in \mathbb{R}^{n_i}$. If x_i is nonsingular, i.e., $\det(x_i) \neq 0$, the inverse of x_i is the unique vector of $\mathbb{R}^{n_i}$ such that $x_i \circ x_i^{-1} = e_i$ and is given by $x_i^{-1} = \det(x_i)^{-1} R_{n_i} x_i$, where R_{n_i} is the reflexion matrix defined by

$$R_{n_i} = \begin{pmatrix} 1 & 0^\top \\ 0 & -I \end{pmatrix} \in \mathbb{R}^{n_i \times n_i}.$$

If for all $i \in \{1, \dots, n\}$ x_i is nonsingular, we have

$$x^{-1} = \begin{pmatrix} x_i^{-1} \\ \vdots \\ x_n^{-1} \end{pmatrix} = \mathrm{Arw}(x)^{-1} e.$$

For $i \in \{1, \dots, n\}$, the spectral decomposition of a vector $x_i \in \mathbb{R}^{n_i}$ is defined by $x_i = \lambda_i c_i + \lambda_i' c_i'$, where

$$\lambda_i = x_{i0} + \|\bar{x}_i\|, \quad \lambda_i' = x_{i0} - \|\bar{x}_i\|, \quad c_i = \frac{1}{2}\begin{pmatrix} 1 \\ \frac{\bar{x}_i}{\|x_i\|} \end{pmatrix} \quad \text{and} \quad c_i' = \frac{1}{2}\begin{pmatrix} 1 \\ -\frac{\bar{x}_i}{\|x_i\|} \end{pmatrix}.$$

The scalars λ_i and λ_i' are eigenvalues of $\mathrm{Arw}(x_i)$, with the corresponding eigenvectors c_i and c_i'. This pair of eigenvectors is called the Jordan frame of x_i. It is said that two vectors commute if they share the same Jordan frame. In that case the corresponding arrow matrices commute. From these definitions, it follows that $x_i \in \mathcal{K}_i$ (resp. $x_i \in \mathrm{int}(\mathcal{K}_i)$) if and only if $\lambda_i' \geq 0$ (resp. $\lambda_i' > 0$). If x_i is nonsingular, then

$$x_i^{-1} = \lambda_i^{-1} c_i + {\lambda'}_i^{-1} c_i'.$$

If $x_i \in \mathrm{int}(\mathcal{K}_i)$, then

$$x_i^{1/2} = \lambda_i^{1/2} c_i + {\lambda'}_i^{1/2} c_i'.$$

More generally, for a continuous function f we can define

$$f(x_i) = f(\lambda_i) c_i + f({\lambda'}_i) c_i'.$$

At last, the quadratic representation of a vector $x_i \in \mathcal{K}_i$ is defined as

$$Q_{x_i} = \mathrm{Arw}^2(x_i) - \mathrm{Arw}(x_i^2) = 2 x_i x_i^\top - \det(x_i) R_{n_i}.$$

The scalars λ_i^2 and ${\lambda'}_i^2$ are eigenvalues of Q_{x_i} with the eigenvectors c_i and c_i'. In particular, two vectors commute if and only if their quadratic representation matrices commute. For $x \in \mathcal{K}$, we define $Q_x = Q_{x_1} \oplus \ldots \oplus Q_{x_n}$. For $x \in \mathrm{int}(\mathcal{K}$, we have $\nabla_x(\log\det(x)) = 2x^{-1}$ and $\nabla_x^2(\log\det(x)) = -2Q_{x^{-1}}$.

These operators Arw and $Q_\cdot$ are fundamental for the design of interior algorithms. See [9, §4] for a review of their interesting and useful properties.

D Proofs of Propositions 2 and 3

Proof of Proposition 2. Let $\mu > 0$. Let us show that the barrier function defined in (19) is coercive. As in the proof of Proposition 1, we show that for all nonzero $s \in \mathbb{R}^{nd}$, $(\varphi_\mu)_\infty(s) > 0$. Let us write $\varphi_\mu = p + \psi_\mu$, where p is the quadratic part. We have $(\varphi_\mu)_\infty = p_\infty + (\psi_\mu)_\infty$ [11, Proposition 2.6.1.]. We have

$$p_\infty(s) = \sup_{t>0} \frac{p(ts) - p(0)}{t} = \begin{cases} q^\top s & \text{if } Ws = 0, \\ \infty & \text{otherwise.} \end{cases}$$

To compute the asymptotic derivative of the function ψ_μ, let us look at one element of the sum. Let $s \in \mathbb{R}^d$ and let us define $\rho(s) := -\log\det s$. We have

$$\begin{aligned}\rho_\infty(s) &= \sup_{t>0} \tfrac{1}{t}(\rho(e+ts) - \rho(e)) \\ &= \sup_{t>0} \tfrac{-1}{t}\log(1 + 2te^\top s + t^2 \det s) \\ &= \delta_{\mathcal{L}}(s).\end{aligned}$$

We then have $(\psi_\mu)_\infty = \delta_{\mathcal{L}^n}$. It follows that $(\varphi_\mu)_\infty = g_\infty$, where g is defined in the proof of Proposition 1. Therefore, the same conclusion holds, that is: under Assumption 1, $(\varphi_\mu)_\infty(s) > 0$ for all nonzero $s \in \mathcal{L}^n$. This implies that the set of minima of φ_μ is nonempty and compact. Since φ_μ is strictly convex, this set is reduced to a singleton.

To show that the central path is bounded, let us proceed to a reasoning by contradiction. Suppose that there exists a positive sequence $\{\mu_k\}$, converging to zero and such that $\{(r_k, u_k)\}$ is unbounded, where we denote $r_k := r(\mu_k)$ and $u_k := u(\mu_k)$. Let $t_k := \|(r_k, u_k)\|$. By taking a subsequence, suppose that $\{\frac{1}{t_k}(r_k, u_k)\} \to (r^*, u^*) \neq 0$. By definition of the central path, for all $k \in \mathbb{N}$ we have

$$Wr_k + q = u_k \quad \text{and} \quad r_k \circ u_k = 2\mu_k e. \tag{44}$$

Dividing the first equation by t_k and the second one by t_k^2, then taking the limit $k \to \infty$, we obtain $Wr^* = u^*$ and $r^* \circ u^* = 0$. Multiplying the first equation by r^* on the left, we obtain $r^{*\top}Wr^* = r^{*\top}u^* = 0$. Since $W = HM^{-1}H^\top$ and M is positive definite, it follows that $H^\top r^* = 0$. Multiplying the first equation of (44) by r_k we have $r_k^\top W r_k + q^\top r_k = r_k^\top u_k = 2n\mu_k$. Since W is positive semi-definite, we then get $q^\top r_k \leq 2n\mu_k$. Dividing by t_k and passing through the limit, we obtain $q^\top r^* \leq 0$. Since for all $k \in \mathbb{N}$, $r_k \in \operatorname{int}(\mathcal{L}^{2n})$, we also have $r^* \in \mathcal{L}^{2n}$. Finally, we have shown that there exists a non-zero vector $r^* \in \mathcal{L}^{2n}$, such that $H^\top r^* = 0$ and $w^\top r^* = q^\top r^* \leq 0$. In view of Lemma 2, this is equivalent to the fact that the Slater hypothesis does not hold, which is in contradiction with the assumption. □

The proof of Proposition 3 is done in two steps. The convergence is proved first. It follows the one of [27, Theorem A.3]. It is based on the use of the curve selection lemma from algebraic geometry. In the second part of the proof, it is shown that the limit is maximally complementary. It is inspired by [21].

Proof of Proposition 3. Let $(\hat{r}, \hat{u})$ be a limit point of the central path. Since $\hat{u}$ is unique, it remains to show that $r(\mu) \to \hat{r}$ as $\mu \downarrow 0$. Let us define the subsets of $\mathbb{R}^{nd+1}$:

$$\mathcal{V} = \{(r, \mu) : (W(\hat{r} + r) + q) \circ (\hat{r} + r) = 2\mu e\} \quad \text{and} \quad \mathcal{U} = \{(r, \mu) : \det(\hat{r} + r) > 0, \mu > 0\}.$$

The set $\mathcal{V}$ is a real algebraic set and $\mathcal{U}$ is an open set defined by a finite number of polynomial inequalities. Moreover, $0 \in \mathbb{R}^{nd+1}$ is in the closure of $\mathcal{U} \cap \mathcal{V}$. By the curve selection lemma [37, Lemma 3.1], there exists $\varepsilon > 0$ and a real analytic curve $p : [0, \varepsilon) \to \mathbb{R}^{nd+1}$, with $p(0) = 0$ and for all $t \in (0, \varepsilon)$, $p(t) \in \mathcal{U} \cap \mathcal{V}$. By setting $p(t) = (\rho(t), \phi(t))$, we have $(W(\hat{r}+\rho(t))+q)\circ(\hat{r}+\rho(t)) = 2\phi(t)e$ and $\phi(t) = \mu$, for all $t \in (0, \varepsilon)$. Since the central path is uniquely defined by (18), we deduce that $r(\mu) = \hat{r} + \rho(t)$, for all $t \in (0, \varepsilon)$. Since the function ϕ is continuous, $\phi(0) = 0$ and $\phi(t) > 0$ for $t > 0$, it is locally nondecreasing and so is invertible on an interval $[0, \bar{\varepsilon}]$, for some $0 < \bar{\varepsilon} < \varepsilon$. We then have $r(\mu) = \hat{r} + \rho(\phi^{-1}(\mu))$ for $\mu > 0$ small enough. It follows that $\lim_{\mu\downarrow 0} r(\mu) = \hat{r}$.

Let us show now that $\hat{r}$ is maximally complementary. Let $r \in \widehat{R}$ be a maximally complementary optimal solution. For all $\mu > 0$, by using (17) and (18) and the positive semi-definiteness of W, we have

$$(r(\mu) - r)^\top (u(\mu) - \hat{u}) = (r(\mu) - r)^\top W (r(\mu) - r) \geq 0.$$

It follows that

$$\sum_{i=1}^{n} r_i(\mu)^\top \hat{u}_i + \sum_{i=1}^{n} r_i^\top u_i(\mu) \leq 2n\mu. \tag{45}$$

Since each term in the sums is nonnegative and $u_i(\mu) = 2\mu r_i^{-1}(\mu)$ for all i, we deduce that for all $i \in \mathsf{B} \cup \mathsf{R} \cup \mathsf{T}_1$, we have $r_i^\top r_i^{-1}(\mu) \leq n$. It follows that for all $i \in \mathsf{B} \cup \mathsf{R} \cup \mathsf{T}_1$, $\lim_{\mu \downarrow 0} r_i(\mu) = \hat{r}_i$ is nonzero, which allows to conclude that $\hat{r}$ is maximally complementary. □