

HPIPM: a high-performance quadratic programming framework for model predictive control[★]

Gianluca Frison^{*} Moritz Diehl^{*}

^{*} *Department of Microsystems Engineering, University of Freiburg,
email: {gianluca.frison, moritz.diehl} at imtek.uni-freiburg.de*

Abstract:

This paper introduces HPIPM, a high-performance framework for quadratic programming (QP), designed to provide building blocks to efficiently and reliably solve model predictive control problems. HPIPM currently supports three QP types, and provides interior point method (IPM) solvers as well (partial) condensing routines. In particular, the IPM for optimal control QPs is intended to supersede the HPMPC solver, and it largely improves robustness while keeping the focus on speed. Numerical experiments show that HPIPM reliably solves challenging QPs, and that it outperforms other state-of-the-art solvers in speed.

Keywords: quadratic programming, model predictive control, embedded optimization, software

1. INTRODUCTION

The aim of this paper is to introduce the quadratic programming (QP) framework in the high-performance software HPIPM in the context of model predictive control (MPC) (Maciejowski, 2002). Among other features, the HPIPM software provides an interior point method (IPM) for the solution of linear-quadratic optimal control problems (OCP), which has been used in numerous applications. In this regard it is intended to supersede the software HPMPC (Frison et al., 2014), which comprises both hardware-tailored linear algebra (LA) routines and solvers for QPs in the MPC form. The LA framework in HPMPC has been completely reimplemented and extended in the BLASFEO software (Frison et al., 2018), which targets routines operating on both the panel-major matrix format proposed in HPMPC, and the column-major matrix format used in the standard BLAS application programming interface (API). Analogously, the QP framework in HPMPC has been completely reimplemented and extended in the HPIPM software, which is the subject of the current paper.

More in detail, HPIPM is an open-source C-coded high-performance framework for QP, and it is designed in a modular fashion with two distinct goals: to be easily extendable to handle new QP types, and to serve as the basis for more generic algorithms such as solvers for nonlinear programming (NLP). At the time of writing, HPIPM supports three QP types: a dense QP, an OCP QP, and a tree-structured OCP QP. For each of them, it defines C structures for the QP itself, its solution and its residuals. These structures serve as arguments to a set of routines factorizing and solving the Karush-Kuhn-Tucker (KKT) system and implementing a family of IPM solvers.

^{*} This research was supported by the German Federal Ministry for Economic Affairs and Energy (BMWi) via eco4wind (0324125B) and DyConPV (0324166B), and by DFG via Research Unit FOR 2401.

The different IPM solvers are meant to provide different trade-offs between computational speed and robustness, with the fastest but least robust algorithm being analogous to the one implemented in HPMPC. At their core, the IPM solvers in HPIPM comprise a common set of routines that operate only on vectors and are agnostic of the QP type or even the matrix format employed. Conversely, the KKT modules, which account for the computationally most intensive operations, are tailored to the specific QP type, and are currently implemented using BLASFEO.

Alongside these QP types and respective IPM solvers, HPIPM implements routines to convert between them. Condensing and partial condensing (Axehill, 2015) routines convert an OCP QP into a dense QP and an OCP QP with shorter horizon, respectively. These (partially) condensed QPs can be solved with tailored IPM solvers, and their solution expanded back into the original OCP QP format by means of dedicated routines. The (partial) condensing and expansion routines can also be employed as stand-alone routines, and coupled with other QP solvers. As such, HPIPM provides a complete QP framework.

HPIPM uses encapsulation mechanisms for its C API: the C structures are not assumed to be accessed directly, but through setters and getters. This fact has several important consequences: for example, it allows the problem data to be stored in the C structures in a format efficient for the solver while setters and getters are designed to be convenient for the user. Additionally, this enhances maintainability, as the internal layout of the structures can freely change as long as the setters and getters expose the same API. Another key feature of HPIPM is its memory management, which eases its embedded use: no internal memory allocation takes place, as HPIPM operates on externally provided chunks of memory. On top of the C API, HPIPM provides interfaces to the high-level programming languages Matlab, Octave, Simulink and Python.

At the time of writing, HPIPM is employed as a QP solver in several NLP solvers such as acados (Verschuere et al., 2019), Control Toolbox (Gifftthaler et al., 2018), MPCC (Liniger et al., 2015) and MATMPC (Chen et al., 2018). In particular, in the case of acados it provides the entire QP framework and it is deeply embedded in the source code, in the sense that the members of the OCP QP type are directly accessed when building the QP approximation at each SQP iteration. Furthermore, the (partial) condensing routines provided by HPIPM are used as pre-processing step in the solution of these QPs by means of other QP solvers such as qpOASES (Ferreau et al., 2014), qpDUNES (Frasch et al., 2014), OSQP (Stellato et al., 2017), OQCP (Gertz and Wright, 2003). In the case of the Control Toolbox, HPIPM is used through its setter-based C API as a QP solver for both constrained and unconstrained linear-quadratic OCPs and it provides the feedback gains needed in the iterative linear quadratic regulator (iLQR) algorithm (Li and Todorov, 2004). In the case of MATMPC, HPIPM is used through its C API within mex sources in Matlab, for both the IPM solver and the (partial) condensing routines. In the case of MPCC, HPIPM is currently the solver of choice for real-time applications (replacing the originally employed FORCES solver (Domahidi et al., 2012)), and it is used through its Matlab interface.

2. QP FORMULATIONS

HPIPM currently defines three QP types: a dense QP, an OCP QP and tree-structured OCP QP. As a distinctive feature compared to most other QP solvers, all QP formulations define a special type of variable, the slacks, which do not enter the dynamics equality constraints and which give a diagonal contribution to Hessian and inequality constraint matrices. The exploitation of this structure allows their elimination from the QP formulation in computational complexity linear in the number of slack variables, making it computationally cheap to use them. The slack variables can be used to efficiently implement soft constraints with L1 and L2 penalties.

2.1 Dense QP

The dense QP type describes a generic QP where Hessian and constraint matrices are assumed to be dense. This formulation can handle the QP sub-problems arising in single-shooting discretization schemes or in state-condensing schemes in the OCP and MPC frameworks.

$$\begin{aligned} \min_{v,s} \quad & \frac{1}{2} \begin{bmatrix} v \\ 1 \end{bmatrix}^T \begin{bmatrix} H & g \\ g^T & 0 \end{bmatrix} \begin{bmatrix} v \\ 1 \end{bmatrix} + \\ & + \frac{1}{2} \begin{bmatrix} s^l \\ s^u \\ 1 \end{bmatrix}^T \begin{bmatrix} Z^l & 0 & z^l \\ 0 & Z^u & z^u \\ (z^l)^T & (z^u)^T & 0 \end{bmatrix} \begin{bmatrix} s^l \\ s^u \\ 1 \end{bmatrix} \\ \text{s.t.} \quad & Av = b \\ & \begin{bmatrix} \underline{v} \\ \underline{d} \end{bmatrix} \leq \begin{bmatrix} J^{b,v} \\ C \end{bmatrix} v + \begin{bmatrix} J^{s,v} \\ J^{s,g} \end{bmatrix} s^l \\ & \begin{bmatrix} J^{b,v} \\ C \end{bmatrix} v - \begin{bmatrix} J^{s,v} \\ J^{s,g} \end{bmatrix} s^u \leq \begin{bmatrix} \bar{v} \\ \bar{d} \end{bmatrix} \\ & s^l \geq \underline{s}^l \\ & s^u \geq \underline{s}^u \end{aligned}$$

The primal variables comprise generic variables v and slack variables s^l (s^u) associated to the lower (upper) constraints. The Hessian matrices of the slacks Z^l and Z^u are diagonal. The matrices J^{\cdot} are made of rows from identity matrices, and are employed to select only some components in box and soft constraints. Furthermore, the constraint matrices are the same for the upper and the lower constraints, meaning that all constraints in the formulation are two-sided. A mask (not represented in the above dense QP formulation) can be employed to dynamically activate or deactivate the single upper and/or lower constraints. The considerations in this paragraph apply also to the OCP and tree OCP QP.

2.2 OCP QP

The OCP QP type describes a QP formulation handling many common OCP and MPC problems such as linear-quadratic regulators (LQR), constrained linear MPC problems, and QP sub-problems in sequential quadratic programming (SQP) algorithms for non-linear OCP and MPC problems.

$$\begin{aligned} \min_{x,u,s} \quad & \sum_{n=0}^N \frac{1}{2} \begin{bmatrix} u_n \\ x_n \\ 1 \end{bmatrix}^T \begin{bmatrix} R_n & S_n & r_n \\ S_n^T & Q_n & q_n \\ r_n^T & q_n^T & 0 \end{bmatrix} \begin{bmatrix} u_n \\ x_n \\ 1 \end{bmatrix} + \\ & + \frac{1}{2} \begin{bmatrix} s_n^l \\ s_n^u \\ 1 \end{bmatrix}^T \begin{bmatrix} Z_n^l & 0 & z_n^l \\ 0 & Z_n^u & z_n^u \\ (z_n^l)^T & (z_n^u)^T & 0 \end{bmatrix} \begin{bmatrix} s_n^l \\ s_n^u \\ 1 \end{bmatrix} \\ \text{s.t.} \quad & x_{n+1} = A_n x_n + B_n u_n + b_n \quad , \quad n \in \mathcal{H} \setminus \{N\} \\ & \begin{bmatrix} \underline{u}_n \\ \underline{x}_n \\ \underline{d}_n \end{bmatrix} \leq \begin{bmatrix} J_n^{b,u} & 0 \\ 0 & J_n^{b,x} \\ D_n & C_n \end{bmatrix} \begin{bmatrix} u_n \\ x_n \end{bmatrix} + \begin{bmatrix} J_n^{s,u} \\ J_n^{s,x} \\ J_n^{s,g} \end{bmatrix} s_n^l \quad , \quad n \in \mathcal{H} \\ & \begin{bmatrix} J_n^{b,u} & 0 \\ 0 & J_n^{b,x} \\ D_n & C_n \end{bmatrix} \begin{bmatrix} u_n \\ x_n \end{bmatrix} - \begin{bmatrix} J_n^{s,u} \\ J_n^{s,x} \\ J_n^{s,g} \end{bmatrix} s_n^u \leq \begin{bmatrix} \bar{u}_n \\ \bar{x}_n \\ \bar{d}_n \end{bmatrix} \quad , \quad n \in \mathcal{H} \\ & s_n^l \geq \underline{s}_n^l \quad , \quad n \in \mathcal{H} \\ & s_n^u \geq \underline{s}_n^u \quad , \quad n \in \mathcal{H} \end{aligned}$$

where $\mathcal{H} = \{0, 1, \dots, N\}$. This problem has a multi-stage structure, with cost and inequality constraints defined stage-wise, and with dynamics equality constraints coupling pairs of consecutive stages. The primal variables are divided into state variables x_n , control (or input) variables u_n and slack variables associated to the lower (upper) constraints s_n^l (s_n^u). The size of all variables (number of states n_{x_n} , number of controls n_{u_n} and number of slacks n_{s_n}), as well as the number of box constraints n_{b_n} and general polytopic constraints n_{g_n} can freely vary stage-wise. All data matrices and vectors can vary stage-wise, as required e.g. to solve QP sub-problems arising in multiple-shooting discretization schemes (Bock and Plitt, 1984).

Note that the current formulation does not support equality constraints other than the dynamics equations, and therefore other types of equality constraints (as e.g. the constraint on the initial state value $x_0 = \hat{x}_0$) have to be reformulated as inequality constraints with equal upper and lower limits. The IPM variant employed in HPIPM (see Section 4.2) is able to handle such reformulation well. This also applies to the tree OCP QP.

2.3 Tree OCP QP

The tree OCP QP type can handle many common robust and scenario-based OCP and MPC problems, see (Kouzoupis, 2019) and references therein.

$$\begin{aligned} \min_{x,u,s} \quad & \sum_{n \in \mathcal{N}} \frac{1}{2} \begin{bmatrix} u_n \\ x_n \\ 1 \end{bmatrix}^T \begin{bmatrix} R_n & S_n & r_n \\ S_n^T & Q_n & q_n \\ r_n^T & q_n^T & 0 \end{bmatrix} \begin{bmatrix} u_n \\ x_n \\ 1 \end{bmatrix} + \\ & + \frac{1}{2} \begin{bmatrix} s_n^l \\ s_n^u \\ 1 \end{bmatrix}^T \begin{bmatrix} Z_n^l & 0 & z_n^l \\ 0 & Z_n^u & z_n^u \\ (z_n^l)^T & (z_n^u)^T & 0 \end{bmatrix} \begin{bmatrix} s_n^l \\ s_n^u \\ 1 \end{bmatrix} \\ \text{s.t.} \quad & x_m = A_m x_n + B_m u_n + b_m, \quad \begin{cases} n \in \mathcal{N} \setminus \mathcal{L} \\ m \in \mathcal{C}(n) \end{cases} \\ & \begin{bmatrix} u_n \\ x_n \\ d_n \end{bmatrix} \leq \begin{bmatrix} J_n^{b,u} & 0 \\ 0 & J_n^{b,x} \\ D_n & C_n \end{bmatrix} \begin{bmatrix} u_n \\ x_n \end{bmatrix} + \begin{bmatrix} J_n^{s,u} \\ J_n^{s,x} \\ J_n^{s,g} \end{bmatrix} s_n^l, \quad n \in \mathcal{N} \\ & \begin{bmatrix} J_n^{b,u} & 0 \\ 0 & J_n^{b,x} \\ D_n & C_n \end{bmatrix} \begin{bmatrix} u_n \\ x_n \end{bmatrix} - \begin{bmatrix} J_n^{s,u} \\ J_n^{s,x} \\ J_n^{s,g} \end{bmatrix} s_n^u \leq \begin{bmatrix} \bar{u}_n \\ \bar{x}_n \\ \bar{d}_n \end{bmatrix}, \quad n \in \mathcal{N} \\ & s_n^l \geq \underline{s}_n^l, \quad n \in \mathcal{N} \\ & s_n^u \geq \underline{s}_n^u, \quad n \in \mathcal{N} \end{aligned}$$

\mathcal{N} is the set of nodes in the tree and \hat{N} is its cardinality. The set \mathcal{L} contains the leaves of the tree, while $\mathcal{C}(n)$ denotes the set of the children of node n . All data matrices and vector can vary node-wise.

3. ALGORITHM

HPIPM implements in a modular fashion all algorithmic building blocks needed to implement many variants of a primal-dual IPM. This algorithmic richness can be used e.g. to trade off computational speed and robustness, to adapt it to the specific application.

3.1 Primal-dual IPM

This section contains some basic notion about primal-dual IPM algorithms. In particular, the IPM algorithm is derived from the KKT conditions, and not from the barrier methods theory. The interested reader can find a more detailed presentations in (Wright, 1997; Nocedal and Wright, 2006).

Let us consider the generic QP in the form

$$\begin{aligned} \min_y \quad & \frac{1}{2} y^T \mathcal{H} y + g^T y \\ \text{s.t.} \quad & \mathcal{A} y = b \\ & \mathcal{C} y \geq d \end{aligned} \quad (1)$$

The Lagrangian function for this QP reads

$$\mathcal{L}(v, \pi, \lambda) = \frac{1}{2} y^T \mathcal{H} y + g^T y - \pi^T (\mathcal{A} y - b) - \lambda^T (\mathcal{C} y - d)$$

where π and λ are the Lagrange multipliers of the equality and inequality constraints, respectively. The first order necessary KKT optimality conditions read

$$\mathcal{H} y + g - \mathcal{A}^T \pi - \mathcal{C}^T \lambda = 0 \quad (2a)$$

$$-\mathcal{A} y + b = 0 \quad (2b)$$

$$-\mathcal{C} y + d + t = 0 \quad (2c)$$

$$\lambda_i t_i = 0 \quad i = 1, \dots, n_c \quad (2d)$$

$$(\lambda, t) \geq 0 \quad (2e)$$

where the slack variables $t = \mathcal{C} y - d \geq 0$ have been introduced, and n_c denotes the total number of constraints. In case the Hessian matrix \mathcal{H} is positive definite, the QP is convex and a point satisfying the KKT conditions is the unique global minimizer.

Equations (2a)-(2d) are a system of nonlinear equations $f(y, \pi, \lambda, t) = 0$, where the only nonlinear equation is (2d). In a nut shell, a primal-dual IPM is Newton method applied to the system of equations $f_\tau(y, \pi, \lambda, t) = 0$ with equation (2d) relaxed as

$$\lambda_i t_i = \tau, \quad i = 1, \dots, n_c$$

The homotopy parameter τ is related to the barrier parameter in barrier methods, and it is shrunk toward zero as the iterations approach the solution of (2). A line search procedure is used to ensure the strict satisfaction of the inequalities (2e) on the sign of Lagrange multipliers and slacks of inequality constraints.

At every iteration k of the Newton method, the Newton step $(\Delta y_{\text{aff}}, \Delta \pi_{\text{aff}}, \Delta \lambda_{\text{aff}}, \Delta t_{\text{aff}})$ is found solving the linear system

$$\begin{aligned} \nabla f_{\tau_k}(y_k, \pi_k, \lambda_k, t_k) \begin{bmatrix} \Delta y_{\text{aff}}^T & \Delta \pi_{\text{aff}}^T & \Delta \lambda_{\text{aff}}^T & t_{\text{aff}}^T \end{bmatrix}^T = \\ = -f_{\tau_k}(y_k, \pi_k, \lambda_k, t_k) \end{aligned}$$

which takes the form

$$\begin{aligned} \begin{bmatrix} \mathcal{H} & -\mathcal{A}^T & -\mathcal{C}^T & 0 \\ -\mathcal{A} & 0 & 0 & 0 \\ -\mathcal{C} & 0 & 0 & I \\ 0 & 0 & T_k & \Lambda_k \end{bmatrix} \begin{bmatrix} \Delta y_{\text{aff}} \\ \Delta \pi_{\text{aff}} \\ \Delta \lambda_{\text{aff}} \\ \Delta t_{\text{aff}} \end{bmatrix} = \\ = - \begin{bmatrix} \mathcal{H} y_k - \mathcal{A}^T \pi_k - \mathcal{C}^T \lambda_k + g \\ -\mathcal{A} y_k + b \\ -\mathcal{C} y_k + t_k + d \\ \Lambda_k T_k e - \tau_k e \end{bmatrix} \doteq - \begin{bmatrix} r_g \\ r_b \\ r_d \\ r_m \end{bmatrix} \quad (3) \end{aligned}$$

The diagonal matrices Λ_k and T_k have on their diagonal the elements of the vector λ_k and t_k respectively. The function $f_{\tau_k}(y_k, \pi_k, \lambda_k, t_k)$ at the right hand side (RHS) is denoted as the residual function.

3.2 Delta and absolute IPM formulations

The formulation of the Newton step in (3) is denoted as ‘delta formulation’ in this paper, in the sense that the solution of the linear system (3) gives the Newton step. By exploiting the linearity of the first three equations in (2), and using for the RHS of the fourth equation in (3) the expression

$$\begin{aligned} T_k \Delta \lambda_{\text{aff}} + \Lambda_k \Delta t_{\text{aff}} &= T_k (\lambda_{\text{aff}} - \lambda_k) + \Lambda_k (t_{\text{aff}} - t_k) \\ &= T_k \lambda_{\text{aff}} + \Lambda_k t_{\text{aff}} - 2\Lambda_k T_k e \end{aligned}$$

the linear system (3) can be rewritten as

$$\begin{bmatrix} \mathcal{H} & -\mathcal{A}^T & -\mathcal{C}^T & 0 \\ -\mathcal{A} & 0 & 0 & 0 \\ -\mathcal{C} & 0 & 0 & I \\ 0 & 0 & T_k & \Lambda_k \end{bmatrix} \begin{bmatrix} y_{\text{aff}} \\ \pi_{\text{aff}} \\ \lambda_{\text{aff}} \\ t_{\text{aff}} \end{bmatrix} = - \begin{bmatrix} g \\ b \\ d \\ -\Lambda_k T_k e - \tau_k e \end{bmatrix} \quad (4)$$

In this paper, this formulation is denoted as ‘absolute formulation’, in the sense that the solution of the linear system (4) gives the next full-step Newton iterate, and the Newton step can be computed as the difference with respect to the current iterate

$$\begin{aligned} \begin{bmatrix} \Delta y_{\text{aff}}^T & \Delta \pi_{\text{aff}}^T & \Delta \lambda_{\text{aff}}^T & \Delta t_{\text{aff}}^T \end{bmatrix}^T = \\ = \begin{bmatrix} y_{\text{aff}}^T & \pi_{\text{aff}}^T & \lambda_{\text{aff}}^T & t_{\text{aff}}^T \end{bmatrix}^T - \begin{bmatrix} y_k^T & \pi_k^T & \lambda_k^T & t_k^T \end{bmatrix}^T \quad (5) \end{aligned}$$

The advantage of the absolute formulation is that (5) has a computational complexity linear in the number of variables, while the evaluation of the residual function at the RHS of (3) involves matrix-vector multiplications, which for dense matrices have a quadratic computational complexity. Therefore, especially in case of systems with few variables, the absolute formulation can be computationally faster. On the other hand, as the iterate approaches the solution, the computation of the Newton step as in (5) incurs in increasingly severe cancellation errors, as the two vectors at the RHS nearly coincide. This effect is exacerbated by the fact that typically in late IPM iterations the conditioning of the linear system gets increasingly worse.

3.3 KKT system solution

The ill-conditioning arises from the fact that in HPIPM, as in most IPM solvers, the slacks t_k and the Lagrange multipliers λ_k are eliminated as a pre-processing step before the actual factorization, as the resulting linear system is in the form of the KKT system of an equality-constrained QP, and therefore efficient factorization procedures for this class of systems can be employed.

In HPIPM, the variables Δt_{aff} and $\Delta \lambda_{\text{aff}}$ are eliminated, in this order, from the KKT system in (3), and the resulting system of linear equations

$$\begin{bmatrix} \mathcal{H} + \mathcal{C}^T T_k^{-1} \Lambda_k \mathcal{C} & -\mathcal{A}^T \\ -\mathcal{A} & 0 \end{bmatrix} \begin{bmatrix} \Delta y_{\text{aff}} \\ \Delta \pi_{\text{aff}} \end{bmatrix} = - \begin{bmatrix} r_g - \mathcal{C}^T (T_k^{-1} \Lambda_k r_d - T_k^{-1} r_m) \\ r_b \end{bmatrix} \quad (6)$$

which is in the form of the KKT system of an equality constrained QP, is solved with structure-exploiting routines operating on dense sub-matrices in case of all QP types currently supported. This approach corresponds to statically selecting the pivoting sequence based solely on sparsity, and hand-crafting efficient factorization algorithms tailored to the QP type at hand. The alternative approach, which is currently not employed in HPIPM, is to directly factorize the KKT system in (3) using sparse LA, which may additionally perform dynamic pivoting for stability in an attempt to keep the conditioning under control.

In HPIPM, in the case of the dense QP type, the linear system in (6) is solved using either the Schur-complement or the null-space methods (Nocedal and Wright, 2006). In case the dense QPs arise from the condensing of the OCP QP type, there are no equality constraints, and the factorization of the KKT matrix in (6) can be simply performed using the dense Cholesky factorization.

In the case of the OCP QP type, as proposed in the MPC context by (Steinbach, 1994) and similarly to HPMPC (Frison et al., 2014), the linear system in (6) is factorized using a structure-exploiting backward Riccati recursion. Two versions of the Riccati recursion are available (Frison, 2015): (a) the classical implementation (which requires the reduced Hessian with respect to the dynamics equality constraints to be positive definite, but allows the full-space Hessian to be indefinite), and (b) the square-root implementation (which in order to reduce the flop count employs the Cholesky factorization of the Riccati recursion matrix, and therefore requires the full-space Hessian to be positive definite). In case the number

of states x_n and inputs n_u is constant across stages, the computational complexity is of $\mathcal{O}(N(n_x + n_u)^3)$ flops, which is linear in the number of stages.

In the case of the tree OCP QP type, the linear system in (6) is factorized using a Riccati recursion modified to exploit the tree structure (Frison et al., 2017). This factorization proceeds from the leaves of the tree, and therefore it does not introduce any fill-in outside the problem data blocks. The Riccati recursion can be implemented using the same two variants described for the OCP QP type. In case the number of states x_n and inputs n_u is constant across stages and realizations, the computational complexity is of $\mathcal{O}(\hat{N}(n_x + n_u)^3)$, and therefore it is linear in the number of nodes in the tree \hat{N} .

3.4 Iterative refinement

In practice, due to the finite accuracy of numerical computations, any procedure for the solution of the KKT system (3) can compute the solution only up to a certain accuracy. Especially in case of unstable or ill-conditioned systems, and in late IPM iterations, the ill-conditioning of the KKT system (3) can make the accuracy in the Newton step too low to be useful for the IPM algorithm. In such cases, it may be beneficial to employ a few iterative refinement steps in order to attempt to improve the accuracy of the Newton step.

The iterative refinement residuals are defined as the residuals in the solution of the KKT system (3),

$$\begin{bmatrix} r_g^{\text{ir}} \\ r_b^{\text{ir}} \\ r_d^{\text{ir}} \\ r_m^{\text{ir}} \end{bmatrix} \doteq \begin{bmatrix} \mathcal{H} & -\mathcal{A}^T & -\mathcal{C}^T & 0 \\ -\mathcal{A} & 0 & 0 & 0 \\ -\mathcal{C} & 0 & 0 & I \\ 0 & 0 & T_k & \Lambda_k \end{bmatrix} \begin{bmatrix} \Delta y_{\text{aff}} \\ \Delta \pi_{\text{aff}} \\ \Delta \lambda_{\text{aff}} \\ \Delta t_{\text{aff}} \end{bmatrix} + \begin{bmatrix} r_g \\ r_b \\ r_d \\ r_m \end{bmatrix}$$

In case they are not small enough, it is possible to solve again a system in the form (3) with the same KKT matrix at the left hand side (LHS) and the iterative refinement residuals as RHS, and compute a correction term to iteratively improve the accuracy of the Newton step. Since no new matrix factorization is performed, the computation of this correction term is cheap compared to the factorization and solution of the KKT system in (3), especially in case of large systems.

3.5 Condensing and partial condensing

HPIPM provides a condensing module, which implements routines to convert an OCP QP into a dense QP by eliminating the state variables at all stages except stage 0. This is possible because such state variables are not real degrees of freedom, as they are determined uniquely from the inputs and the state at stage 0 by using the dynamics equality constraints. The partial (or block) condensing module, which converts an OCP QP into another OCP QP with smaller number of stages, is built on top of the condensing module in a modular fashion, by performing condensing on blocks of stages.

There exist at least three Hessian condensing algorithms with different computational complexities, ranging between quadratic and cubic in both the number of states and the number of stages (Frison, 2015). The algorithm

currently implemented in HPIPM is the algorithm proposed in (Frison et al., 2016), which has a computational complexity quadratic in the number of stages and cubic in the number of states. This algorithm is based on a backward recursion which has similarities to the Riccati recursion. As such, two versions of this condensing algorithm are available: one analogous to the classical Riccati recursion and one analogous to the square-root Riccati recursion, with similar requirements about Hessian positive definiteness. This condensing algorithm is the fastest when the state at stage 0 is kept as an optimization variable (such as in partial condensing) and in case of long horizons.

4. IMPLEMENTATION

4.1 Linear algebra

Algorithms in HPIPM are explicitly designed to exploit the overall structure of the optimization algorithms. Sparsity within data and working matrices is not considered. Therefore, computations are cast in term of operations on matrices with fixed structure, like e.g. dense, symmetric, triangular or diagonal.

In HPIPM, the KKT modules contain the most computationally expensive routines, such as KKT system factorization and solution, and residuals computation. Such operations have computational complexities cubic or quadratic in the number of (stage) variables. In case of all currently implemented QP types, in the KKT modules HPIPM employs the structure-based interface in the high-performance LA library BLASFEO (Frison et al., 2018) for such operations. These routines are designed to give close-to-peak performance for matrices of moderate size (assumed to fit in cache) and are optimized for several hardware architectures (e.g. exploiting different classes of SIMD instructions).

The IPM core operations in HPIPM are independent of the specific QP type and solely operate on unstructured arrays of doubles holding e.g. the set of all primal variables, or Lagrange multipliers or slack variables. These operations have computational complexity linear in the number of variables, and are currently offered in two implementations, one coded in generic C and one using AVX intrinsics.

4.2 IPM implementation choices

At a high-level, HPIPM implements an infeasible-start Mehrotra’s predictor-corrector IPM (Mehrotra, 1992), proposed in the MPC context e.g. in (Rao et al., 1998). This standard algorithmic scheme is tweaked by several arguments, whose different choice gives rise to a family of algorithms with different trade-offs between computational speed and reliability. A ‘mode’ argument is introduced in order to select pre-defined sets of arguments tailored to a particular purpose. The modes currently available are:

speed_abs This mode selects an algorithm analogous to the one implemented in HPMP. The focus lies on the highest solution speed, and the IPM is based on the absolute formulation. The residuals are optionally computed only before returning, and therefore the IPM loop exit conditions are only based on duality measure tolerance, maximum number of iterations and minimum

step length. This algorithm is indicated for small and well-conditioned systems, in case tight bounds on the solution accuracy are not required.

speed The focus lies on the high solution speed, but the more reliable IPM delta formulation is employed. The KKT residuals are computed at each IPM iteration, and exit conditions are additionally based on minimum residuals tolerance.

balance In this mode, the accuracy of the KKT factorization is checked by computing its iterative refinement residuals. In case accuracy is too low, the factorization is repeated by replacing all Cholesky factorizations of normal matrices in the form $A \cdot A^T$ with array algorithms based on QR factorizations, which have better numerical properties as they never explicitly form the worse-conditioned normal matrix $A \cdot A^T$. Furthermore, a few iterative refinement steps are performed on the prediction-centering-correction direction, in case this is not accurate enough.

robust In this mode, the KKT factorization is always performed using the more accurate array algorithms based on QR factorizations. A higher number of iterative refinement steps is performed on the prediction-centering-correction direction, in case this is not accurate enough.

Some additional implementation choices are common to all modes. The Mehrotra’s corrector step is implemented conditionally, and is only applied if it is not increasing the duality measure too much compared to the affine step; otherwise, a predictor-centering step is applied. The Hessian matrix of both primal and dual variables can be regularized by adding a scaled identity matrix during factorization; this regularization does not affect the original problem data, and therefore the effect of regularization on the solution can be offset by means of iterative refinement. The value of the inequality constraint Lagrange multipliers and slacks is bounded to a minimum value, in order to give the possibility to bound the ill-conditioning of the KKT system in late IPM iterations. Finally, it is possible to provide a guess to warm start either the primal, or both the primal and the dual variables (this second option is particularly effective in SQP algorithms, since close to the NLP solution a good guess for the primal variables is zero, and for the dual variables coincides with the Lagrange multipliers of the NLP).

5. NUMERICAL EXPERIMENTS

In this section, numerical experiments evaluate the relative speed (Section 5.1) and robustness (Section 5.2) of the different HPIPM modes, and compare HPIPM to other state-of-the-art MPC solvers (Section 5.3).

5.1 Linear mass spring system

The linear mass spring system is a scalable benchmark commonly employed to evaluate linear MPC solvers (Wang and Boyd, 2010). The variant proposed in (Domahidi et al., 2012) is considered, comprising a horizontal chain of M masses connected to each other and to walls using springs, and where the first $M - 1$ masses are controlled using a force. Therefore the number of states $n_x = 2M$, inputs $n_u = M - 1$ and box constraints $n_b = n_x + n_u = 3M - 1$ per stage scales with the number of masses.

Table 1. Comparison of solvers for box-constrained linear MPC: `speed_abs`, `speed` and `robust` modes for HPIPM, low-level interfaces for HPMPC. M is the number of masses; N is the horizon length. Runtimes in seconds; number of IPM iterations fixed to 10. Intel Core i7 4800MQ CPU with ‘Haswell’ target in BLASFEO and ‘AVX2’ target in HPMPC.

M	N	HPIPM <code>speed_abs</code>	HPIPM <code>speed</code>	HPIPM <code>robust</code>	HPMPC low-level
2	10	$6.71 \cdot 10^{-5}$	$1.17 \cdot 10^{-4}$	$1.80 \cdot 10^{-4}$	$5.7 \cdot 10^{-5}$
4	10	$1.10 \cdot 10^{-4}$	$1.82 \cdot 10^{-4}$	$3.67 \cdot 10^{-4}$	$8.9 \cdot 10^{-5}$
6	30	$6.36 \cdot 10^{-4}$	$9.00 \cdot 10^{-4}$	$1.96 \cdot 10^{-3}$	$4.88 \cdot 10^{-4}$
11	10	$4.62 \cdot 10^{-4}$	$6.22 \cdot 10^{-4}$	$1.61 \cdot 10^{-3}$	$4.05 \cdot 10^{-4}$
15	10	$8.29 \cdot 10^{-4}$	$1.05 \cdot 10^{-3}$	$2.94 \cdot 10^{-3}$	$7.51 \cdot 10^{-4}$
30	30	$1.21 \cdot 10^{-2}$	$1.45 \cdot 10^{-2}$	$4.53 \cdot 10^{-2}$	$1.17 \cdot 10^{-2}$

Table 1 compares the solution time for 3 HPIPM modes with HPMPC. The algorithm in HPIPM with `speed_abs` mode is analogous to the one in HPMPC, and the computational times are also within 10-15% of each other; HPIPM is slightly slower due to its more modular nature, and its use of an external LA library. The HPIPM `speed` mode is about 2 times slower for small system, but the difference decreases for large systems (since the same KKT factorization is employed). The HPIPM `robust` mode is roughly 3 times slower for all systems, since a slower but more accurate KKT factorization is employed.

5.2 CUTEr

This section evaluates the performance and robustness of HPIPM using the challenging Maros-Mészáros test set of convex QPs (Maros and Mészáros, 1999), which can be accessed through the CUTEr testing environment. In particular, the QPs available on the qpOASES COIN-OR repository are employed, which are the 43 problems with up to 250 variables and up to 1001 constraints, as discussed in (Ferreau et al., 2014). All matrices are in dense format.

Figure 1 contains a performance graph comparing the performance of the `speed` and `balance` modes in HPIPM with qpOASES; the required accuracy is set to 10^{-6} for the infinity norm of residuals. HPIPM in `speed` mode is the fastest solver in about 50% of the problems, but it fails to solve 4 problems. HPIPM in `balance` mode can solve all problems, it is the fastest solver in about 20% – and within a factor 2 of the fastest solver for more than 90% – of the problems. qpOASES is the fastest solver in the remaining 30% of the problems, but it gets quickly outperformed (often by a factor 10 or more) in the remaining problems.

HPIPM in `speed_abs` mode (not shown in Figure 1) would fail to converge in 32 problems, and return an inaccurate solution in the remaining problems. This shows the great improvement in reliability over HPMPC (which implements an algorithm analogous to HPIPM in `speed_abs` mode, but can not be directly compared as it does not provide a dense QP solver) in case of challenging QPs.

5.3 Chain of masses

This section compares the performance of HPIPM against other state-of-the-art QP solvers for MPC problems. In

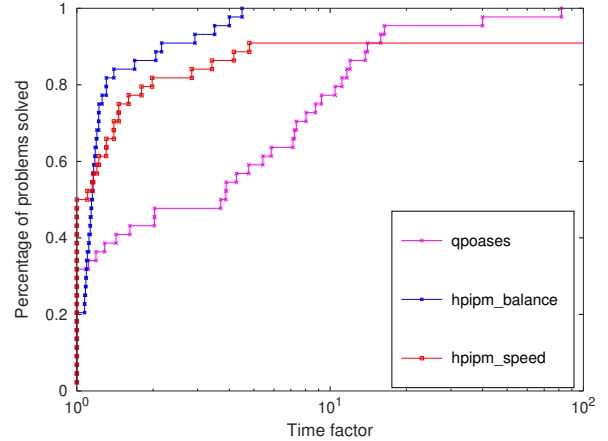


Fig. 1. Performance graph on the 43 problems from the Maros-Mészáros QP test set with up to 250 variables and up to 1001 constraints. Intel Core i7 4810MQ CPU with ‘Haswell’ target in BLASFEO.

particular, the NLP framework acados is employed to generate constrained time-varying linear MPC problems. The test problem is the nonlinear hanging chain problem used for the comparison of QP solvers in (Kouzoypis et al., 2018), where the chain of masses starts at rest position, and a strong perturbation is generated by overwriting the controller output in the first 5 simulation steps; afterwards the controller takes over again and takes the chain of masses back to the resting position. Four free masses are employed, accounting for a total of 24 states (the 3D position and velocity of each free mass), 3 controls (the 3D force acting on the last mass), 4 state and 3 control bounds; the control horizon is of 8 seconds, divided into 40 equidistant stages.

The numerical experiment consists of a closed loop simulation of 50 steps, where the real time iteration scheme in acados is used to generate one QP to be solved per simulation step. Optionally partial condensing is employed to reformulate the OCP QP into another OCP QP with 5 stages. The QP solution times are in Figure 2 for the solvers HPIPM (`speed` mode; OCP QP solver for the fully sparse or partially condensed QP; dense QP solver for the condensed QP), OSQP (sparse solver for the fully sparse or partially condensed QP) and qpOASES (for the condensed QP). (Partial) condensing is provided by the acados wrapper to the corresponding HPIPM routines. Note that only the call to the QP solver is timed: therefore, the (partial) condensing time is not included, nor the SQP algorithm time. Warm start is employed for OSQP and qpOASES, as these solvers are found to benefit from it; OSQP employs a polishing step to improve the low accuracy of the ADMM (Alternating Direction Method of Multipliers) solution.

HPIPM solving the partially condensed OCP QP stands out as the fastest solver, with solution times steadily below 1 millisecond. Both HPIPM and OSQP are significantly faster in solving the partially condensed QP compared to the corresponding fully sparse QP. In general, HPIPM (being an IPM) shows a rather constant solution time, while qpOASES (active set) and OSQP (ADMM) require more time (due to more iterations) when the QP problems are more challenging, at the beginning of the simulation.

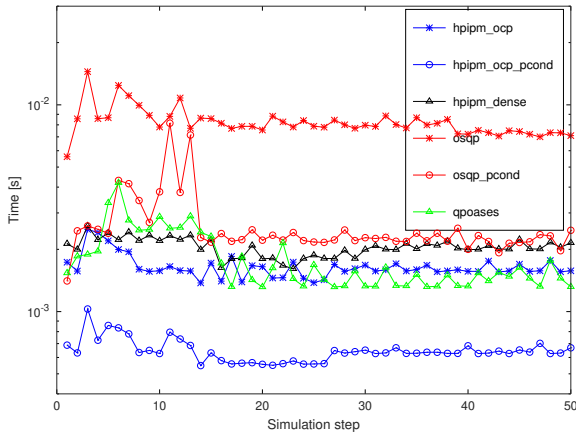


Fig. 2. Comparison of QP solvers in the solution of QPs generated using acados in a closed loop MPC simulation on the nonlinear chain of masses. Intel Core i7 4810MQ CPU with ‘Haswell’ target in BLASFEO.

6. CONCLUSION

This paper introduced HPIPM, a high-performance QP framework for the solution of MPC problems. HPIPM provides IPM solvers for dense, OCP and tree OCP QPs, as well as (partial) condensing routines. In particular, compared to the predecessor HPMPC solver, reliability is shown to be significantly improved in case of challenging QPs. Numerical experiments show that HPIPM speed excels against other state-of-the-art QP solvers for MPC.

REFERENCES

- Axehill, D. (2015). Controlling the level of sparsity in MPC. *Systems & Control Letters*, 76, 1–7.
- Bock, H.G. and Plitt, K.J. (1984). A multiple shooting algorithm for direct solution of optimal control problems. In *Proc. of the IFAC World Congress*, 242–247.
- Chen, Y., Bruschetta, M., Picotti, E., and Beghi, A. (2018). Matmpc - a matlab based toolbox for real-time nonlinear model predictive control. *arXiv preprint*.
- Domahidi, A., Zraggen, A., Zeilinger, M.N., Morari, M., and Jones, C.N. (2012). Efficient interior point methods for multistage problems arising in receding horizon control. In *Proc. of the IEEE Conference on Decision and Control (CDC)*, 668–674.
- Ferreau, H.J., Kirches, C., Potschka, A., Bock, H.G., and Diehl, M. (2014). qpOASES: a parametric active-set algorithm for quadratic programming. *Mathematical Programming Computation*, 6(4), 327–363.
- Frasch, J.V., Vukov, M., Ferreau, H.J., and Diehl, M. (2014). A new quadratic programming strategy for efficient sparsity exploitation in SQP-based nonlinear MPC and MHE. In *Proc. of the IFAC World Congress*.
- Frison, G., Kouzoupis, D., Jørgensen, J.B., and Diehl, M. (2016). An efficient implementation of partial condensing for nonlinear model predictive control. *Proc. of the IEEE Conference on Decision and Control*, 4457–4462.
- Frison, G., Kouzoupis, D., Sartor, T., Zanelli, A., and Diehl, M. (2018). BLASFEO: Basic linear algebra subroutines for embedded optimization. *ACM Transactions on Mathematical Software*, 44(4), 42:1–42:30.
- Frison, G., Sorensen, H.B., Dammann, B., and Jørgensen, J.B. (2014). High-performance small-scale solvers for linear model predictive control. In *Proc. of the European Control Conference*, 128–133.
- Frison, G. (2015). *Algorithms and Methods for High-Performance Model Predictive Control*. Ph.D. thesis, Technical University of Denmark (DTU).
- Frison, G., Kouzoupis, D., Diehl, M., and Jørgensen, J.B. (2017). A high-performance Riccati based solver for tree-structured quadratic programs. In *Proc. of the IFAC World Congress*, volume 50, 14399–14405.
- Gertz, E.M. and Wright, S.J. (2003). Object-oriented software for quadratic programming. *ACM Transactions on Mathematical Software*, 29(1), 58–81.
- Gifftahler, M., Neunert, M., Stäubli, M., and Buchli, J. (2018). The Control Toolbox - an open-source C++ library for robotics, optimal and model predictive control. In *IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots*.
- Kouzoupis, D. (2019). *Structure-exploiting numerical methods for tree-sparse optimal control problems*. Ph.D. thesis, University of Freiburg.
- Kouzoupis, D., Frison, G., Zanelli, A., and Diehl, M. (2018). Recent advances in quadratic programming algorithms for nonlinear model predictive control. *Vietnam Journal of Mathematics*, 46(4), 863–882.
- Li, W. and Todorov, E. (2004). Iterative linear quadratic regulator design for nonlinear biological movement systems. In *Proc. of the 1st International Conference on Informatics in Control, Automation and Robotics*.
- Liniger, A., Domahidi, A., and Morari, M. (2015). Optimization-based autonomous racing of 1:43 scale RC cars. *Optimal Control Applications and Methods*, 36(5), 628–647.
- Maciejowski, J.M. (2002). *Predictive Control with Constraints*. Prentice Hall.
- Maros, I. and Mészáros, C. (1999). A repository of convex quadratic programming problems. *Optimization Methods and Software*, 11, 431–449.
- Mehrotra, S. (1992). On the Implementation of a Primal-Dual Interior Point Method. *SIAM Journal on Optimization*, 2(4), 575–601.
- Nocedal, J. and Wright, S.J. (2006). *Numerical Optimization*. Springer, 2 edition.
- Rao, C.V., Wright, S.J., and Rawlings, J.B. (1998). Application of interior-point methods to model predictive control. *Journal of Optimization Theory and Applications*, 99, 723–757.
- Steinbach, M.C. (1994). A structured interior point SQP method for nonlinear optimal control problems. In R. Bulirsch and D. Kraft (eds.), *Computation Optimal Control*, 213–222. Birkhäuser, Basel Boston Berlin.
- Stellato, B., Banjac, G., Goulart, P., Bemporad, A., and Boyd, S. (2017). OSQP: An operator splitting solver for quadratic programs. *ArXiv e-prints*.
- Verschueren, R., Frison, G., Kouzoupis, D., van Duijkeren, N., Zanelli, A., Novoselnik, B., Frey, J., Albin, T., Quirynen, R., and Diehl, M. (2019). acados: a modular open-source framework for fast embedded optimal control. *arXiv preprint*.
- Wang, Y. and Boyd, S. (2010). Fast model predictive control using online optimization. *IEEE Transactions on Control Systems Technology*, 18(2), 267–278.
- Wright, S.J. (1997). *Primal-Dual Interior-Point Methods*. SIAM Publications, Philadelphia.