Embedded Code Generation Using the OSQP Solver

Goran Banjac*, Bartolomeo Stellato*, Nicholas Moehle, Paul Goulart, Alberto Bemporad, and Stephen Boyd

Abstract—We introduce a code generation software package that accepts a parametric description of a quadratic program (QP) as input and generates tailored C code that compiles into a fast and reliable optimization solver for the QP that can run on embedded platforms. The generated code is based on OSQP, a novel open-source operator splitting solver for quadratic programming. Our software supports matrix factorization caching and warm starting, and allows updates of the problem parameters during runtime. The generated C code is library-free and has a very small compiled footprint. Examples arising in real-world applications show that the generated code outperforms state-of-the-art embedded and desktop QP solvers.

I. INTRODUCTION

Convex optimization has become a standard tool across many engineering fields. In recent years, these methods have increasingly been applied on embedded systems where data are processed in real time and on low-cost computational platforms [1, Ch. 1], [2]. Current applications include, e.g., model predictive control (MPC) [3], real-time signal processing [4], [5] and onboard trajectory planning in space missions [6], [7].

Real-time applications of embedded optimization impose special requirements on the solvers used [8]. First, embedded solvers must be reliable even in the presence of poor quality data, and should avoid exceptions caused by division by zero or memory faults caused by dynamic memory allocation. Second, the solver should be implementable on low-cost embedded platforms with very limited memory resources. In particular, solvers should have very small compiled footprint, should consist only of basic algebraic operations, and should not be linked to any external libraries, which also makes the solver easily verifiable. Finally, real-time applications typically require that the solver is fast and able to correctly identify infeasible problems.

On the other hand, optimization problems arising in embedded applications have certain features that can be exploited when designing an embedded solver [8]. First, embedded optimization is typically applied to the repeated solution of parametrized problems in which the problem data, but not its dimensions or sparsity pattern, change between problem instances. For such problems, the solver initialization and some part of its computations can be performed offline during the solver design phase. Second, requirements on the solution accuracy in embedded applications are often moderate because of noise in the data and arbitrariness of the objective function. Also, in embedded applications one can typically assume that problems are reasonably scaled. As an example, the authors in [9] show that acceptable control performance of an MPC controller is achievable even when using a very low accuracy solver. This argument supports the use of first-order optimization methods, which are known to return solution of medium accuracy with low computational cost. Although the performance of first-order methods is known to depend strongly on problem scaling, this dependency can be reduced significantly by preconditioning the problem data [10].

A. Related work

In some cases solution of a parametrized convex optimization problem can be precomputed offline using multiparametric programming techniques [11], [12]. However, the memory required for storing such solutions grows exponentially with the problem dimensions, making this approach applicable only to small problems.

In the last decade tools for generating custom online solvers for parametric problems have attracted increasing attention. CVXGEN [8] is a code generation software tool for small-scale parametric QPs. The generated solver is fast and reliable, but its main disadvantage is that the code size grows rapidly with the problem dimensions. This issue is overcome in FORCES [13], [14] where the code size of the compiled code is broadly constant with respect to the problem dimensions. In HPMPC [15] tailored solvers for MPC are combined with high-performance optimized libraries for linear algebra. ECOS [16], [17] and Bsocp [18] are embedded solvers for a wider class of second-order cone programs (SOCPs). All of the aforementioned solvers are based on primal-dual interior point methods that are tailored for their specific problem classes. A known limitation of these methods is that they cannot use the warm starting technique, which is one of the dominant acceleration factors in applications such as MPC [19].

In contrast, qpOASES [20] is based on a parametric active-set method which can effectively use a priori information to speed-up computation of a QP solution. On the other hand, since qpOASES is based on dense linear algebra it cannot exploit sparsity in the problem data. Moreover,
the computational complexity of active-set methods grows exponentially with the number of constraints.

FiOrdOs [21] uses first-order gradient methods as the basis for the embedded solvers it generates. In the case of a general QP, the methods require a Lipschitz constant of the gradient of the objective function in order to compute the stepsize. Alternatively, FiOrdOs implements an adaptive rule for the stepsize selection, but it requires a new matrix factorization each time the stepsize is updated. QPgen [22] uses optimal preconditioning of the problem data that can improve performance of first-order methods considerably. The main disadvantage of FiOrdOs and QPgen is their inability to detect infeasible problems.

In this paper we introduce a software package that generates tailored C code that compiles into a solver for a user-specified parametric quadratic program (QP). The generated code is based on the open-source solver OSQP [23] which exploits sparsity in the problem data, supports matrix factorization caching and warm starting, and is able to detect infeasible problems.

B. Structure of the paper

In Section II we introduce an operator splitting method which is the basis for the solvers generated. Two different classes of parametric QPs are described in Section III. The code generation software package is presented in Section IV. In Section V we present numerical results on problems arising in real-world applications. Finally, Section VI concludes the paper.

II. OSQP SOLVER

We consider the following QP

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} x^T P x + q^T x \\
\text{subject to} & \quad l \leq Ax \leq u,
\end{align*}
\]

where \( x \in \mathbb{R}^n \) is the optimization variable. The objective function is defined by a positive semidefinite matrix \( P \in \mathbb{S}^n_+ \) and a vector \( q \in \mathbb{R}^n \), and the constraints by a matrix \( A \in \mathbb{R}^{m \times n} \) and vectors \( l \in \mathbb{R}^m \cup \{-\infty\}^m \) and \( u \in \mathbb{R}^m \cup \{+\infty\}^m \) such that \( l \leq u \). Linear equality constraints can be enforced by setting \( l_i = u_i \in \mathbb{R} \).

A. Algorithm

The algorithm used in OSQP [23] is based on the alternating direction method of multipliers (ADMM) [24] and is described in Algorithm 1. Scalars \( \rho > 0 \) and \( \sigma > 0 \) are penalty parameters, \( \alpha \in (0, 2) \) is a relaxation parameter, and \( \Pi \) is the Euclidean projection onto the set \( \{ z \in \mathbb{R}^m \mid l \leq z \leq u \} \) which has a simple closed-form solution

\[
\Pi(z) = \max \left( \min(\min(z, u), l) \right),
\]

where the max and min operators should be taken element-wise.

Note that steps 4 to 7 involve only scalar-vector multiplications, vector additions and the projection operator \( \Pi \). We describe in the next subsection how to solve efficiently the linear system in step 3.

Algorithm 1 OSQP solver

1: given initial values \( x^0, z^0, y^0 \) and parameters \( \rho, \sigma, \alpha \)
2: repeat
3: \hspace{1em} solve \( \begin{bmatrix} P + \sigma I & A^T \\ A & -\frac{1}{\rho} I \end{bmatrix} \begin{bmatrix} z^{k+1} \\ y^{k+1} \end{bmatrix} = \begin{bmatrix} \sigma x^k - q \\ -\alpha z^k \end{bmatrix} \)
4: \hspace{1em} \tilde{z}^{k+1} \leftarrow z^k + \frac{1}{\rho} (y^{k+1} - y^k)
5: \hspace{1em} x^{k+1} \leftarrow \alpha \tilde{z}^{k+1} + (1 - \alpha) x^k
6: \hspace{1em} z^{k+1} \leftarrow \Pi (\alpha \tilde{z}^{k+1} + (1 - \alpha) z^k + \frac{1}{\rho} y^k)
7: \hspace{1em} y^{k+1} \leftarrow y^k + \rho (\alpha \tilde{z}^{k+1} + (1 - \alpha) z^k - z^{k+1})
8: until termination conditions are satisfied

B. Solving the linear system

Since the coefficient matrix in Algorithm 1 line 3 does not change between iterations of the algorithm, it must be factorized only once. This factorization is then cached and used in forward and backward solves in all subsequent iterations.

The coefficient matrix is quasi-definite, i.e., it can be written as a 2-by-2 block-symmetric matrix where the (1,1)-block is positive definite, and the (2,2)-block is negative definite. It therefore always has a well defined \( LDL^T \) factorization, with \( L \) being a lower triangular matrix with unit diagonal elements and \( D \) a diagonal matrix with nonzero diagonal elements [25]. For any sparse quasidefinite matrix \( K \), efficient algorithms can be used to compute a permutation matrix \( P \) for which the factorization \( PKP^T = LDL^T \) results in a sparse matrix factor \( L \) [26]. We use the approximate minimum degree (AMD) code [27] to compute such a permutation matrix.

In order to perform the \( LDL^T \) factorization of a sparse matrix \( K \) (or permuted matrix \( PKP^T \)) efficiently, the sparsity pattern for the factor \( L \) should be found before performing any numerical operations. Determining this sparsity pattern is known as symbolic factorization and requires only the nonzero structure of the matrix \( K \), but not its numerical values. After the symbolic factorization finds the pattern of nonzero elements of \( L \), the numerical values of these elements can be computed. This procedure is known as numerical factorization. Note that if the non-zero entries of the matrix \( K \) change, but the sparsity pattern remains the same, then only the numerical factorization step needs to be performed again and the memory required to store the new factorization does not change.

We use the SuiteSparse package [28] to perform symbolic and numerical factorizations. The computational complexity of factorization and forward and backward solves with SuiteSparse, as well as the memory required for storing the factor \( L \), depend only on the number of nonzero elements in \( K \) and not on its dimensions. The code is thus efficient in terms of both the memory and computational effort required to solve the linear system in Algorithm 1 with sparse matrices \( P \) and \( A \).

Observe that the matrix \( P \) is not required to be positive semidefinite for the linear system to be solvable. In particular, as long as \( P + \sigma I \) is positive definite, the \( LDL^T \) factoriza-
tion of the matrix in Algorithm 1 will return a factor $D$ with nonzero diagonal elements, and exceptions caused by division by zero cannot occur when solving the linear system. This is critical in embedded applications in the presence of poor quality data.

C. Termination criterion

The algorithm used in OSQP generates at each iteration $k$ a tuple $(x^k, z^k, y^k)$ for which we define the primal and dual residuals as

\[ r_{\text{prim}}^k = Ax^k - z^k, \]
\[ r_{\text{dual}}^k = Px^k + q + A^T y^k. \]

If the problem $P$ is solvable then the residuals converge to zero vectors as $k \to \infty$ [24]. If the residuals are small, we say that $(x^k, y^k)$ is an approximate primal-dual solution of problem $P$. A reasonable termination criterion is that the norms of the residuals $r_{\text{prim}}^k$ and $r_{\text{dual}}^k$ are smaller than some tolerance levels $\varepsilon_{\text{prim}} > 0$ and $\varepsilon_{\text{dual}} > 0$, respectively. Note that the tolerance levels are often chosen relative to the scaling of the algorithm iterates; see [24, Sec. 3.3] for details.

Detection of infeasible problems has been historically a weak point of first-order methods. An exception is the algorithm in [10] that can detect infeasibility in conic problems, and the works in [29], [30] that detect infeasible QPs under additional assumptions on the problem data. OSQP can detect primal and dual infeasibility when the problem is unsolvable. In particular, if the problem is primal infeasible, the algorithm generates a certificate of infeasibility, i.e., a vector $v \in \mathbb{R}^m$ that satisfies

\[ A^T v = 0, \quad u^T v_+ + l^T v_- < 0, \]

where $v_+ = \max(v, 0)$ and $v_- = \min(v, 0)$. Likewise, if the problem is dual infeasible the algorithm generates a vector $s \in \mathbb{R}^n$ that satisfies

\[ P s = 0, \quad q^T s < 0, \quad (As)_i \begin{cases} \leq 0 & l_i \in \mathbb{R}, \ u_i \in \mathbb{R} \\ \geq 0 & l_i \in \mathbb{R}, \ u_i = +\infty \\ \leq 0 & u_i \in \mathbb{R}, \ l_i = -\infty \end{cases} \]

which is a certificate of dual infeasibility. We refer the reader to [31] for details.

III. PARAMETRIC PROGRAMS

In many applications problem $P$ is solved for varying data, and is thus referred to as a parametric program. We make a distinction between two cases depending on which of the problem data are to be treated as parameters. We assume throughout that the problem dimensions $m$ and $n$, and the sparsity patterns of $P$ and $A$ are fixed.

A. Vectors as parameters

If the vectors $q$, $l$ and $u$ in problem $P$ are the only parameters, then the coefficient matrix in Algorithm 1 does not change across different instances of the parametric program. The matrix can then be pre-factored offline and only backward and forward solves are performed during code execution. This enables a significant reduction of the code footprint. If the diagonal matrix $D^{-1}$ is stored instead of $D$, then the resulting algorithm is division-free.

This important class of parametric programs arises, for instance, in linear MPC [32], linear regression with (weighted) $\ell_1$ regularization [33], [34] and portfolio optimization [1].

B. Matrices and vectors as parameters

If values in matrices $P$ and $A$ are updated, then the coefficient matrix in Algorithm 1 must be refactored. We assume, however, that the sparsity patterns of $P$ and $A$ are fixed. In this case symbolic factorization of the matrix does not change, and only numerical factorization needs to be redone. This implies that no dynamic memory allocation is performed during the code execution.

This class of problems arises, e.g., in nonlinear MPC, nonlinear moving horizon estimation (MHE) [32] and sequential quadratic programming (SQP).

C. Warm starting

When a series of similar optimization problems is solved, the solutions across problem instances are often similar. Since the algorithm running time depends largely on the distance between the algorithm’s initial iterate and the problem’s set of optimizers, one can set a solution of the previous problem instance as the initial iterate in the next instance. This strategy is known as warm starting and often improves running times of iterative optimization algorithms [19].

IV. CODE GENERATION WITH OSQP

OSQP is an open-source operator splitting QP solver written in the C language, with interfaces to high-level languages including Matlab, Python and Julia.

Listing 1 shows a simple Python script that generates code for a given problem family. To generate a solver, the end-user must provide the problem data and (optionally) configure the solver settings. The end-user has some flexibility to customize the solver prior to code generation. For instance, if the setting early_terminate is set to 1, then the solver will terminate when one of the termination criteria is satisfied, or when the maximum number of iterations is reached, whichever happens first. Checking the termination

```python
import osqp

# Create an OSQP object
m = osqp.OSQP()

# Solver initialization
m.setup(P, q, A, l, u, settings)

# Generate code
m codegen(’code’, project_type=’Makefile’,
parameters=’vectors’)
```

Listing 1. A simple Python script for generating the code for a given parametric QP.
#include "osqp.h"
#include "workspace.h"

int main(int argc, char **argv) {
    // Solve problem
    osqp_solve(&workspace);
    return 0;
}

Listing 2. A simple C program that loads the problem data from header file workspace.h and solves the problem.

Fig. 1. The tree structure of the generated code. The main program is stored in example.c.

criteria in each iteration is computationally expensive since it involves several matrix-vector multiplications, and may slow down the code execution considerably. If the user instead sets early_terminate to 0, then the algorithm will run for the maximum number of iterations without checking the termination criteria. Alternatively, one can check the termination criteria every \( N \) iterations; this is specified with the setting early\_terminate\_interval. For the complete list of solver settings we refer the reader to [23].

To generate the code, the codegen method is called with the specified name of a directory where the generated code is stored. Using the keyword argument project\_type the user can define the build environment, e.g., Makefiles or several supported IDEs such as Eclipse, Apple Xcode or Microsoft Visual Studio. The keyword argument parameters allows the user to specify which of the data are parameters. The option vectors assumes that only vectors \( q, l \) and \( u \) in problem \( P \) are parameters, while the option matrices allows matrices \( P \) and \( A \) to be parameters as well.

A. Generated files

Figure 1 shows the tree structure of the generated code. Directories \(<\text{dir\_name}>\)/src/osqp and \(<\text{dir\_name}>\)/include contain the solver source code. The generated code is self-contained, has small footprint, does not perform dynamic memory allocation, and is thus suitable for embedded applications. CMakeLists.txt is a CMake configuration file that manages the compilation process in a compiler- and platform-independent manner [35].

The main program is stored in example.c whose content is shown in Listing 2. The program loads the problem data from the header file workspace.h, and solves the problem. For solving a different instance of parametric problems, the problem data must be updated. Listing 3 provides illustrative function calls for updating vectors in problem \( P \); for the complete documentation we refer the reader to [23].

V. Numerical results

We benchmarked the generated solvers against the open-source code generation tools CVXGEN [8], FiOrdOs [21], the open-source solver qpOASES [20], and the commercial solver GUROBI [36]. All the solvers were selected with their default options. We performed benchmarks on a MacBook Pro 2.8GHz Intel Core i7 with 16GB RAM running Python 3.5. The code for the examples is available at [37].

A. Portfolio optimization

We consider a portfolio optimization problem [1, p. 185–186], [17] where we want to maximize risk-adjusted return. The problem is

\[
\begin{align*}
\text{maximize} & \quad \mu^T x - \gamma (x^T \Sigma x) \\
\text{subject to} & \quad 1^T x = 1, \quad x \geq 0,
\end{align*}
\]

where \( x \in \mathbb{R}^n \) represents the portfolio, \( \mu \in \mathbb{R}^n \) is the vector of expected returns, scalar \( \gamma > 0 \) is the risk-aversion parameter and \( \Sigma \in \mathbb{S}_+^n \) is the asset return covariance. A common assumption is the \( k \)-factor risk model [38], where the return covariance matrix is the sum of a diagonal matrix and a matrix of rank \( k \), i.e.,

\[
\Sigma = D + FF^T,
\]

where \( F \in \mathbb{R}^{n \times k} \) and \( D \in \mathbb{R}^{n \times n} \) is diagonal with nonnegative diagonal elements. Problem (1) can be written in the standard form \( \mathcal{P} \) with the linear cost \( q \) depending on the parameter \( \gamma \); see Appendix I for details.

In order to obtain Pareto optimal portfolios, one needs to solve problem (1) for varying risk-aversion parameter \( \gamma \). Since the parameter appears only in the linear cost, one does not need to perform any matrix factorization once the code is generated. Moreover, seeing that the optimal solution does not significantly differ with small changes in \( \gamma \), we can make use of warm starting and get a range of Pareto optimal portfolios with minimal computational effort.

The data are generated as follows: \( k = \lceil n/10 \rceil \), \( F \) has half of the elements set to zero, and the other half are drawn from \( \mathcal{N}(0,1) \), diagonal elements of \( D \) are drawn from \( \mathcal{U}(0, \sqrt{k}) \), and the elements of \( \mu \) are drawn from \( \mathcal{N}(0,1) \). We generate 11 values of \( \gamma \) equally spaced on a logarithmic scale between \( 10^{-2} \) and \( 10^2 \). For each solver and each dimension \( n \) we
is the data matrix whose columns are potential regressors, where \( x \) is the vector of parameters, \( n \) is the number of assets. The size of the compiled code for all the tested instances of problem (1).

### Table I

<table>
<thead>
<tr>
<th>Assets ( n )</th>
<th>Nonzeros in ( P ) and ( A )</th>
<th>Size of file (kB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>235</td>
<td>46</td>
</tr>
<tr>
<td>80</td>
<td>496</td>
<td>58</td>
</tr>
<tr>
<td>100</td>
<td>720</td>
<td>70</td>
</tr>
<tr>
<td>120</td>
<td>984</td>
<td>82</td>
</tr>
<tr>
<td>150</td>
<td>1455</td>
<td>102</td>
</tr>
<tr>
<td>200</td>
<td>2440</td>
<td>142</td>
</tr>
<tr>
<td>250</td>
<td>3675</td>
<td>190</td>
</tr>
<tr>
<td>300</td>
<td>5160</td>
<td>246</td>
</tr>
<tr>
<td>400</td>
<td>8880</td>
<td>382</td>
</tr>
<tr>
<td>500</td>
<td>13600</td>
<td>554</td>
</tr>
</tbody>
</table>

The data are generated as follows: \( n \) varies from 10 to 400, \( m = 10n \), matrix \( C \) has 40\% nonzero elements drawn from \( \mathcal{N}(0,1) \), \( \hat{x} \in \mathbb{R}^n \) has half of the elements set to zero, and the other half are drawn from \( \mathcal{N}(0,1/n) \). We then set \( d = C\hat{x} + \epsilon \), where \( \epsilon \) represents the noise whose elements are drawn from \( \mathcal{N}(0,1/4) \). We generate \( 21 \) values of \( \gamma \) equally spaced on the logarithmic scale from \( 10^{-2} \) to \( 10^{2} \). For each solver and each dimension \( n \) we solve the generated problem for the \( 21 \) values of \( \gamma \) and average the execution time.

The data are generated as follows: \( n \) varies from 10 to 400, \( m = 10n \), matrix \( C \) has 40\% nonzero elements drawn from \( \mathcal{N}(0,1) \), \( \hat{x} \in \mathbb{R}^n \) has half of the elements set to zero, and the other half are drawn from \( \mathcal{N}(0,1/n) \). We then set \( d = C\hat{x} + \epsilon \), where \( \epsilon \) represents the noise whose elements are drawn from \( \mathcal{N}(0,1/4) \). We generate \( 21 \) values of \( \gamma \) equally spaced on the logarithmic scale from \( 10^{-2} \) to \( 10^{2} \). For each solver and each dimension \( n \) we solve the generated problem for the \( 21 \) values of \( \gamma \) and average the execution time.

### B. Sparse regressor selection

We seek a sparse solution of the regressor selection problem, which is in general a hard combinatorial problem [1, Ch. 6.3]. Lasso [33] is a popular heuristic for enforcing sparsity of the solution by adding the \( \ell_1 \) regularization term in the objective. The problem is

\[
\text{minimize} \quad \frac{1}{2} \|Cx - d\|_2^2 + \gamma \|x\|_1, \tag{2}
\]

where \( x \in \mathbb{R}^n \) is the vector of parameters, \( C \in \mathbb{R}^{m \times n} \) is the data matrix whose columns are potential regressors, \( d \in \mathbb{R}^m \) is the vector of measurements that is to be fit by a subset of regressors, and \( \gamma > 0 \) is the weighting parameter. Problem (2) can be written in the standard form \( P \) with the linear cost \( q \) depending on the parameter \( \gamma \); see Appendix II for details.

In order to find a sparse solution with a given degree of sparsity, i.e., not more than \( k \) nonzero elements in \( x \), the above problem should be solved for varying weighting parameter \( \gamma \). As in Section V-A, we can exploit the fact that \( \gamma \) enters only in the linear part of the cost function to cache the matrix factorization and warm start the solver.

The results are shown in Figure 3. FiOrdOs does not converge within 50,000 iterations for this problem type and CVXGEN is not able to generate code for \( n > 10 \). OSQP performs clearly better than GUROBI and qpOASES for all the considered number of parameters. Note that qpOASES is not able to find solution in less than 10 seconds for \( n > 100 \).

### VI. Conclusion

This paper introduces a new code generation software tool based on the OSQP solver. The generated code is very efficient and robust. Moreover, it is general purpose and does not require additional assumptions on problem data beyond convexity. Numerical results show that the generated code not only outperforms state-of-the-art embedded code-generation tools but also desktop solvers.
APPENDIX I
DATA FOR THE PORTFOLIO OPTIMIZATION EXAMPLE

Problem (1) can be written in form $P$ with the following problem data

$$
P = \begin{bmatrix} D & 0 \\ 0 & I \end{bmatrix}, \quad q = \begin{bmatrix} \frac{1}{25} \mu \\ 0 \end{bmatrix},$$

$$
A = \begin{bmatrix} 1^T \\ F^T & -I \end{bmatrix}, \quad l = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad u = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.
$$

APPENDIX II
DATA FOR THE LASSO EXAMPLE

Problem (2) can be written in form $P$ with the following problem data

$$
P = \begin{bmatrix} 0 \\ I \\ 0 \end{bmatrix}, \quad q = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

$$
A = \begin{bmatrix} C & -I \\ I & 0 \\ -I & I \end{bmatrix}, \quad l = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad u = \begin{bmatrix} d \\ +\infty \\ +\infty \end{bmatrix}.
$$

REFERENCES


