UNIVERSITY OF SOUTHAMPTON

Implementing Improved Model Predictive Control of Selective Catalytic Reduction

by

Jonathan A. Sowman

A thesis submitted for the degree of
Doctor of Engineering

in the
Faculty of Engineering and the Environment
Energy Technology Group

October 2017

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND THE ENVIRONMENT
ENERGY TECHNOLOGY GROUP

Doctor of Engineering

IMPLEMENTING IMPROVED MODEL PREDICTIVE CONTROL OF SELECTIVE CATALYTIC REDUCTION

by Jonathan A. Sowman

Oxides of nitrogen (NO_x), as emitted by lean-burn direct injection internal combustion engines, have been shown to be harmful to human health, which has resulted in a tightening of road vehicle emissions limits for these pollutants. NO_x abatement systems, the most popular of which is selective catalytic reduction (SCR), have been developed and proven effective in production, but increasingly stringent legislation mandates further improvement. In particular, SCR suffers from suboptimal NO_x conversion at the low catalyst temperatures occurring in slow moving urban traffic, yet it is precisely these environments where pollutants cause maximal harm. In this thesis, the problem of control of aftertreatment catalysts is considered, in the light of the significant body of recent research into real-time constrained optimal control, otherwise known as model predictive control (MPC). This paradigm, which has so far seen relatively slow adoption by the automotive industry, promises numerous advantages over classical techniques, including systematic handling of constraints, improved closed loop performance, and ease of tuning and calibration. However, the computational demand of MPC has traditionally been considered insurmountable for real-time application to fast systems. The efficacy of MPC as applied to SCR is examined, focusing particularly on challenging operating conditions, including multivariable control by the addition of a catalyst heater in order to cope with urban traffic conditions. It is demonstrated that SCR can exploit the many benefits offered by MPC. Furthermore, this thesis demonstrates that solving the associated optimal control problem is possible in real-time on a low power automotive grade embedded hardware platform, thereby indicating its feasibility for production SCR control and providing a pathway to wider adoption by industry.

Contents

List of Figures

List of Tables

Acronyms

ACP	Accelerator coherency port
AFR	Air-to-fuel ratio
ANR	Ammonia to NO_x ratio
ARM	Advanced RISC machine
ASAM	Association for standardisation of automation and measuring systems
AXI	Advanced extensible interface
AUS	Aqueous urea solution
AWGN	Additive white gaussian noise
BRAM	Block random access memory
BSFC	Brake specific fuel consumption
BVP	Boundary value problem
CAN	Controller area network
CAS	Computer algebra system
CPU	Central processing unit
CSTR	Continuously stirred tank reactor
CUC	Clean-up catalyst
DOC	Diesel oxidation catalyst
DOF	Degrees of freedom
DPF	Diesel particulate filter
DSP	Digital signal processing
ECU	Engine/Electronic control unit
EHC	Electrically heated catalyst
EKF	Extended Kalman filter
EGR	Exhaust gas recirculation
FGM	Fast gradient method
FPGA	Field programmable gate array
GPU	Graphics processing unit
IPM	Interior point method
IQ	Injection quantity
IRK	Implicit Runge-Kutta
IFT	Implicit function theorem
IVP	Initial value problem

KKT	Karush-Kuhn-Tucker
LLSQ	Linear least squares
LNT	Lean NO_x trap
LPV	Linear parameter varying
LQR	Linear quadratic regulator
MIMO	Multiple-input, multiple-output
MPC	Model predictive control
NEDC	New European drive cycle
NLP	Nonlinear programme
NMPC	Nonlinear model predictive control
NO_x	Nitrogen oxides
NRTC	Non-Road test cycle
OBD	On-board diagnostics
OCM	On-chip memory
OCP	Optimal control problem
ODE	Ordinary differential equation
QP	Quadratic programme
PEMS	Portable emissions measurement system
PID	Proportional-Integral-Derivative
PWA	Piecewise affine
PWL	Piecewise linear
RAM	Random access memory
RHC	Receding horizon control
RMS	Root mean square
RTI	Real time iterations
SCR	Selective catalytic reduction
SCRF	SCR on filter
SoC	System-on-Chip
TP-BVP	Two-point boundary value problem
UKF	Unscented Kalman filter
XCP	Universal measurement and calibration protocol
ZOH	Zero order hold

Declaration of Authorship

I, Jonathan A. Sowman, declare that this thesis entitled Implementing Improved Model Predictive Control of Selective Catalytic Reduction and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Either none of this work has been published before submission, or parts of this work have been published as: Sowman et al. 2017.

Signed:

Date:

Acknowledgements

I am indebted to my academic supervisors, Dr Dina Laila and Prof. Andrew Cruden, as well as my industrial supervisor, Dr Peter Fussey, for their collective help, guidance and inspiration. The time spent on discussions, proof-reading papers and this thesis, and countless other tasks is enormously appreciated. Additionally, I must thank my examiners, Prof. Jan Maciejowski and Dr Edward Richardson, for their time and effort in examining this work.

I would also like to thank Dr Tony Truscott at Ricardo for the many discussions regarding aftertreatment control and modelling, your experience and expertise has been essential to this project. Additionally, my thanks to the innumerable people at Ricardo who have participated in helpful discussions and generated ideas for research directions.

My gratitude to the Engineering and Physical Sciences Research Council (EPSRC) for providing funding under grant EP/G036896/1, and Ricardo UK Ltd for supporting this project, both financially and by provision of an industrial supervisor.

My fellow doctoral candidates, Katie, David and Alice, have been a great source of camaraderie throughout this period of study.

To my parents and sister Aneka: thank you for the unending support, not only during this venture but during all the years preceding it, without which I would not have found myself in a position to undertake this project, let alone successfully complete it.

Finally, to my fiancée Georgia: for your enduring love, friendship and encouragement – thank you.

"Inquiry is fatal to certainty."

—Will Durant

Chapter 1

Introduction

Internal combustion engines have been a staple prime mover in innumerable applications for more than a century, including transport, power generation and plant equipment. In a continuing drive towards increased efficiency and power, the combination of direct injection (DI), in which fuel is delivered directly into the combustion chamber, and lean-burn combustion, in which the power generated is controlled via the quantity of fuel injected, rather than via an intake throttle, has proven successful due to inherently improved fuel economy owing to reduced throttling losses. The majority of this type of engine are diesel fuelled, which have long been preferred in heavy duty applications, such as generators and large shipping vehicles, due to their torque delivery compared to their gasoline counterparts. In recent years, this trend has begun to emerge in passenger vehicles as well. The number of newly registered cars powered by diesel has risen steadily, with the total exceeding one million for the first time in 2012, where just over 50% of registrations were diesel whilst only 48% were gasoline (the remaining cars were alternative fuel). Almost one third of the cars on UK roads (around 9.4 million) were diesel, a marked increase from only 7.4% in 1994 [1].

Whilst pollutant emissions from diesel and gasoline engine are significantly different in general, generation of oxides of nitrogen (NO_x, largely comprising NO and NO_2) is largely distinguished by the combustion mode of the engine: lean or stoichiometric, where the former produces significantly more NO_x than the latter. Emissions from stoichiometric gasoline engines have been largely addressed in recent years by use of the three way catalyst (TWC). These engines emit relatively high levels of carbon monoxide (CO), unburnt hydrocarbons (HC) and oxides of nitrogen. The introduction of the TWC has improved these to tolerable levels, and are a legal requirement. The older two-way catalyst performs oxidation of carbon monoxide to carbon dioxide, and oxidation of hydrocarbons to carbon dioxide and water, whilst the later three-way catalytic converter also adds the reduction of nitrogen oxides to nitrogen and oxygen. In contrast, lean combustion creates significantly increased quantities of NO_x and particulate matter (PM).

Lean combustion implies an excess of oxygen, which is also present in the exhaust. As a result of this, the TWC is inapplicable to this variety of internal combustion engines.

The effects of NO_x on human health have been extensively studied and recommendations for safe exposure levels have been published and used as guidelines for emissions limits. It has been shown that both short and long term exposures to oxides of nitrogen can introduce negative health effects. Short term exposure to high levels of NO_x can result in severe pulmonary damage and in people suffering with asthma can cause short term responses such as changes in lung function. Long term effects have been associated with increased respiratory symptoms [2]. A significant proportion of ground level NO_x comes from vehicles – concentrations have been shown to be up to 100% higher within 50m of roads. NO_x reacts with ammonia and moisture in the air to produce minute particles that penetrate deeply into sensitive parts of the lungs and can cause or worsen respiratory disease, such as emphysema and bronchitis, and can aggravate existing heart disease, leading to increased hospital admissions and premature death [3]. In the presence of sunlight, NO_x reacts to form ozone (O_3), which has similar respiratory effects.

As a result of the increased awareness of the effects of NO_x on human health and the environment, combined with the ever increasing uptake of diesel engines in passenger vehicles, the emissions of vehicles sold in EU member states have been regulated since 1993 by the European Emissions Standards (or *Euro* standards). These standards are updated every few years and vehicles produced must conform to the latest standards. Updates are designed to continually reduce pollutants, meaning manufacturers are having to continuously find new technology to improve the emissions of their vehicles. Figure 1.1 shows how the emission limits have gradually decreased over time, as a percentage of their permitted values when the standards were introduced in 1993. It is likely that this trend will continue, with future standards requiring a reduction in pollutant emission, fuel consumption, or both.

Particulate matter emissions have been improved drastically by the use of the gasoline and diesel particulate filters (GPF/DPF). These units, which are effectively made compulsory for diesel by current Euro 5 emissions legislation, are an extremely effective method for lowering the particulate output by trapping soot particles which are subsequently removed in a *regeneration* process. One of the most widely accepted and deployed techniques for reducing the formation of NO_x in internal combustion engines has been exhaust gas recirculation (EGR). As the name suggests, this technique involves recycling some of the inert gas from the exhaust manifold back to the intake manifold. In doing so, several effects which help to reduce NO_x production come into play; these are discussed in more detail in Subsection 2.1.2. One of the downsides of EGR is a reduction in brake specific fuel consumption (BSFC) [4], which may limit its use in future emissions regulations. Even with EGR, NO_x emissions are coming into focus again as Euro 6 emissions standards are on the horizon and fuel consumption targets as well as emissions targets need to be met. On its own, EGR cannot achieve the required NO_x

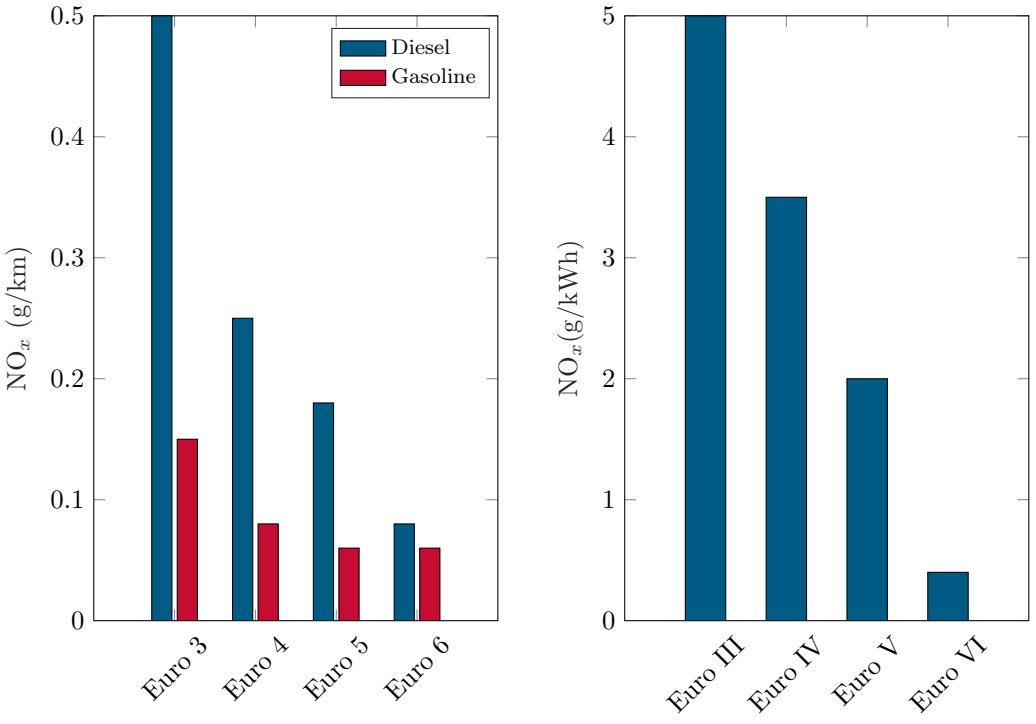


FIGURE 1.1: Emissions requirements for passenger and heavy duty vehicles.

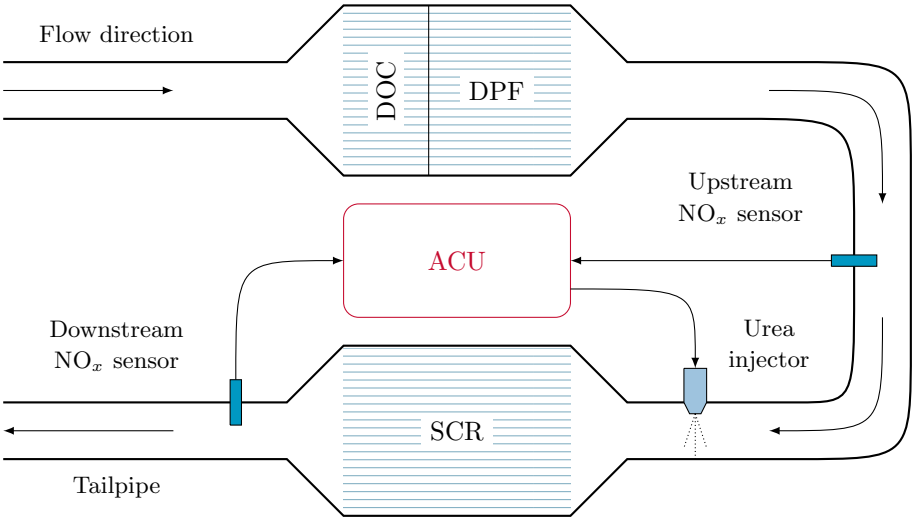


FIGURE 1.2: Schematic of a diesel exhaust system, including a diesel oxidation catalyst (DOC), diesel particulate filter (DPF) and selective catalytic reduction (SCR) reactor, as well as the aftertreatment control unit (ACU).

emissions that current standards require, due to restrictions upon its use to lower engine loads for reasons of both engine performance and efficiency (see Subsection 2.1.2 for further information). Furthermore, it is possible that future emissions standards will begin to tighten fuel consumption limits as well as emissions levels. In this case, the combined target of meeting emissions regulations whilst reducing fuel consumption will place more reliance on the aftertreatment system.

Two major aftertreatment techniques have emerged with a view to improving the NO_x performance even further: the NO_x adsorber or Lean NO_x Trap (LNT) and selective catalytic reduction (SCR). The LNT is a molecular sponge, absorbing NO_x from the exhaust gases and later being regenerated by injection of fuel before the LNT, an operation principle very similar to that of the DPF. However, the requirement for regeneration means that overall engine efficiency suffers. Conversely, SCR comprises the reduction of NO_x by ammonia over a catalyst. Storage and transport of ammonia is problematic due to its toxic nature, instead it is stored in the vehicle as an aqueous urea solution (AUS) which is injected into the catalyst and hydrolyses to produce ammonia owing to the heat of the exhaust gas. The requirement to transport and refill AUS is not too onerous, due to its relatively safe nature and low cost (around £2 per litre at the time of writing). Figure 1.2 shows a typical schematic for the exhaust system of a diesel engine fitted with Diesel Oxidation Catalyst (DOC), DPF and SCR. The details of the chemical and physical mechanisms of SCR are described further in Subsection 2.1.4.

1.1 Motivation

Selective catalytic reduction is an active dynamical system that requires input and thus its operation is possible to control. The principal input is the quantity of urea (which is converted to ammonia) to inject upstream of the SCR catalyst periodically, with injection carried out by an injector unit which is fed by a urea pump from the additive tank on the vehicle, as shown in Figure 1.2. Historical emissions legislation was less stringent, meaning that emissions limits could be met by applying industry standard techniques, such as feedforward map-based controllers. These controllers comprise a lookup table (or *map*) which defines the required urea injection quantity for a given engine speed and load. However, a combination of increased transient operation of SCR (such as in vehicles in urban environments) and more tightly constrained emissions regulation means that these approaches no longer suffice. Indeed, in start-stop city traffic, this transient behaviour is most pronounced, yet it is in these locations that minimisation of pollutants is of greatest importance for human health. Improving the performance of SCR can be achieved via three main avenues: improved chemical catalysis, improved packaging of the catalyst, or improved control of the aftertreatment system. Whilst new chemical technologies for SCR are appearing, the gains are relatively modest, and the physical packaging of catalysts in a vehicle is typically highly constrained by other factors. As such, this thesis focuses on the application of advanced control methods to improve the performance of SCR whilst reducing calibration expense.

The difficulty in control of SCR is principally due to the combination of transport delays and nonlinear dynamics. Whilst the control objectives are very clear (minimise NO_x output as well as NH_3, typically with some weighting terms), the dynamics of the system make its control challenging. Whilst a NO_x sensor is placed upstream of the SCR

catalyst, the injected ammonia has to be adsorbed onto the catalyst surface before it reacts with NO_x. The delay between the urea injection, its adsorption onto the catalyst surface, and its subsequent reaction with exhaust NO_x is significantly longer than the transport time of the gas through the SCR reactor.

In Chapter 3 of this thesis, we collect data from SCR-equipped vehicle during real-world operation which highlights this difficulty. Even relatively modern vehicles are found to be in violation of emissions limits under many circumstances. These conditions are typically met in geographic locations where the minimisation of pollutants is of the most importance. This is further evidenced by plans to abandon diesel vehicles in urban environments in the near future [5], and some cities having banned older diesel vehicles entirely [6]. As these moves will be expensive, this thesis examines a new advanced control approach to SCR in order to improve its efficiency. Since the timescale of new control systems for aftertreatment is relatively short, this will have a positive short-term impact on internal combustion engine road vehicles. In addition, it will permit permanently improved emissions for applications where internal combustion engines cannot so easily be abandoned.

Model predictive control appears to offer several advantages when considered for an SCR process. Firstly, its model-based nature means that providing a model can be provided which includes the transport delay dynamics, the controller should natively cope with such things. Secondly, constraints on the process are present, including the maximum unreacted ammonia that may leave the catalyst, and minimum/maximum constraints of the actuators, with which MPC can natively cope. Thirdly, the natural extension of MPC to multiple-input, multiple-output systems proves to be useful in adapting SCR systems to meet stricter emissions targets via thermal management. Lastly, tuning of an MPC controller amounts to choosing a cost function for the closed loop, which can be relatively simply designed by consideration of the objectives of the control system. In the case of SCR, the cost function can be formulated as some function of the tailpipe emissions, in order that the controller find the optimal operating point with respect to this cost. The combination of these potential advantages comprises the motivation for this thesis: to examine whether in practice, MPC for SCR aftertreatment achieves the performance that might be expected, and how barriers to its adoption by industry might be overcome.

1.2 Novel Contributions

This thesis makes the following contributions:

[C1] It has been shown that model predictive control enables constraint-adhering optimal control of SCR, enabling constraint certification of the controller and providing reduced calibration effort. In particular, the novel contribution is in the demonstration that model predictive control can obtain this performance across the full operational temperature range of the catalyst, thereby removing calibration effort in gain scheduling across temperatures. These claims are supported by publications [P1] and [P2].

[C2] For the first time, an economic nonlinear MPC formulation for SCR control has been developed. This improves existing results on storage-based MPC for SCR found in the literature, both by reduced computational effort by avoiding calculation of the optimal steady state, and by achieving lower closed-loop costs compared to tracking MPC. Furthermore, closed-loop stability of this formulation is certified. This is documented fully in publication [P7].

[C3] Detailed evidence of urea SCR systems performing poorly in real world operation has been collected, which is collated in publication [P9]. This data relates to urban environments and public transport vehicles – an area of specific concern regarding pollution at present. The results have value not only in their exhibition of the discrepancy between design and real world operation, but also suggest thermal control as a mitigation measure, since catalyst temperature is a key factor in conversion efficiency. This work supports and informs the subsequent control design research.

[C4] It has been demonstrated that combined chemical and thermal control of an SCR catalyst leads to significant improvements in emissions reduction. Additionally, it permits the designer the freedom to balance the fundamental trade-off in SCR control between NO_x and CO_2 emissions. This is supported by publication [P6].

[C5] Two state estimators have been developed, in order to estimate the ammonia stored quantity in an SCR catalyst, and to reject the cross-sensitivity between NO_x and ammonia that is typical of commercial NO_x sensors. These methods are based on the Unscented Kalman Filter (UKF) and the particle filter. The UKF has improved performance over the Extended Kalman Filter found in the literature whilst being computationally acceptable for online implementation. The particle filter provides even better estimation at the cost of computational expense, but may be suitable for real-time operation in the future or on parallel computation hardware. This claim is supported by publication [P4].

[C6] A novel method for extending the idea of explicit model predictive control into the nonlinear MPC domain has been established [P3]. If the state spaces and input spaces are finite, this method works by laying a multidimensional grid over the union of these spaces and solving the optimal control problem at the vertices. The online operation of the controller is then reduced to multidimensional interpolation

over the solution grid, which can be upper bounded in terms of computation expense.

[C7] Adoption of new technologies by the automotive industry is always a gradual process. As such, reference governor (RG) theory has been leveraged to add model-based constraint certification to a industry standard map-based primal controller [P8]. This is the first application of reference governor theory to the SCR plant. Acting in a supervisory layer above the normal controller, this method does not provide the flexibility or performance of MPC, but instead provides a pathway to adoption of MPC.

TABLE 1.1: Evidence of novel contributions made by this thesis

Claim	Publication
[C1]	[P1], [P2]
[C2]	[P7]
[C3]	[P9]
[C4]	[P6]
[C5]	[P4]
[C6]	[P3]
[C7]	[P8]

1.3 List of Publications Associated with this Thesis

Many of the contributions made by this thesis have been published separately, in the works listed below.

1.3.1 Published Works

[P1] J. Sowman, D. S. Laila, A. J. Cruden, P. Fussey, A. Truscott. *A Predictive Control Approach to Diesel Selective Catalytic Reduction.* In: Proc. European Control Conference, 2015. Linz.
Reduction of oxides of nitrogen (NO_x) emitted from diesel exhaust systems is a current problem due to increased stringency in worldwide emissions legislation. One of the most successful approaches to reduce tailpipe NO_x is to reduce NO_x by ammonia over a catalyst, known as Selective Catalytic Reduction (SCR). Control of the ammonia injection in such systems is typically a map-based approach, often augmented by feedback from NO_x sensors to account for mechanical variation and ageing. We show that a predictive control approach to this system yields several compelling improvements over such industry standard controllers during a representative test cycle. These include better NO_x conversion performance whilst

simultaneously minimising the quantity of ammonia released to the environment, along with reduced design effort.

[P2] J. Sowman, D. S. Laila, A. J. Cruden, P. Fussey. *Nonlinear Model Predictive Control for Cold Start Selective Catalytic Reduction.* In: Proc. IFAC Conference on Nonlinear Model Predictive Control, 2015. Seville.
Selective catalytic reduction (SCR) is emerging as a key technology for reducing emissions of nitrogen oxides (NO_x) from diesel vehicles, but the temperature dependence of the governing chemical kinetics are highly nonlinear and industry standard techniques of limiting ammonia injection until the catalyst reaches operating temperature leave room for improvement of NO_x reduction. Cold start emissions constitute a significant fraction of urban NO_x emissions, due to low road speeds and short journeys precluding the catalyst from reaching operating temperature quickly. We demonstrate that nonlinear model predictive control (NMPC) provides the desired control performance in adhering to the required constraints and meeting the complex control objectives regardless of catalyst temperature. The results include improved overall NO_x reduction during a typical test cycle including cold start, without design effort specifically for low temperature operation. We also show that the controller is amenable to real-time implementation for use in a vehicle.

[P3] J. Sowman, D. S. Laila, S. Longo. *Real-Time Approximate Explicit Nonlinear Model Predictive Control for the Swing-Up of a Reaction Wheel Pendulum.* In: Proc. Conference on Decision and Control, 2015. Osaka.
In this paper, we show that nonlinear model predictive control (NMPC) demonstrates excellent performance in driving a reaction wheel pendulum to its unstable equilibrium at which the pendulum is inverted. We show that NMPC is capable of driving the system to this point from the stable equilibrium (i.e. the non-inverted position), often known as swing-up. This is as opposed to many common implementations in which two controllers are used one for swing-up and another for stabilisation or balancing around the equilibrium. We derive an explicit version of the controller which provides a close approximation to the required control input, but can do so within the required sampling period. We demonstrate that this approach generates a real-time controller of a size and speed appropriate for embedded implementation on a microprocessor.

[P4] J. Sowman, D. S. Laila, A. Truscott, P. Fussey, A. J. Cruden. *Real-Time Rejection of Ammonia Cross Sensitivity in Sensors for Diesel Aftertreatment Systems by Parallel Particle Filtering.* In: Proc. European Control Conference, 2016. Aalborg.
Selective catalytic reduction (SCR) continues to be the preferred technology for abatement of oxides of nitrogen (NO_x) from diesel engines, yet challenges remain to bring SCR efficiency up to the level required by legislation. Such efficiency

gains are predicated on improved control, which in turn relies in part upon improved state estimation, due to suboptimal plant observability. SCR relies on the injection of urea solution to convert harmful NO_x to nitrogen and water, whilst avoiding excess ammonia leaving the tailpipe. Ammonia sensors to detect such events are costly and not yet production ready. Knowledge of the quantity of ammonia stored in the catalyst is necessary for control, but this cannot be measured directly. Finally, NO_x sensors are cross sensitive to ammonia by an unknown nonlinear function of temperature. We design a high accuracy state estimator in the form of a particle filter to estimate these quantities. Furthermore, we demonstrate a fast implementation of this estimator using a graphics processing unit (GPU) to demonstrate the real time accuracy possible with modern computing platforms.

[P5] M. F. Brejza, J. Hooker, J. Sowman, D. Oakley, R. G. Maunder. *Design of Digital Testbeds for Undergraduate Microelectronics Teaching.* In: Proc. European Workshop on Microelectronics Education, Southampton, 2016.
This work is not relevant to the EngD research, but is a result of collaboration with the authors during the EngD candidature.
In this paper we detail the series of digital microelectronics testbed boards which have been developed to meet the requirements of the microelectronics undergraduate programmes at the University of Southampton. We discuss how our boards solve many of the issues with the digital testbeds used previously for the microelectronics programmes and those which are available on the market. The digital testbed solutiondescribed in this paper is modular, and comprises multiple boards which each perform a specific function. This modular design is compact and easily expandable, while being cost effective such that they can be given to students to use at home. The boards are well suited for use with modern digital components, and fit in well with teaching of microcontrollers, programmable logic, and discrete logic.

[P6] J. Sowman, D. S. Laila, P. Fussey, A. Truscott, A. J. Cruden. *Nonlinear MPC applied to multivariable thermal and chemical control of SCR aftertreatment.* Symposium on Combustion Control, Aachen, 2017.
The ammonia storage capacity of the SCR catalyst has a negative correlation with its temperature, thus the achievable rate of NO_x removal from the exhaust stream is temperature dependent. At low temperature, reaction rates are temperature limited, resulting in suboptimal NO_x removal, whereas at high temperatures, they are mass transfer limited. SCR operates optimally in the latter regime, wherein precise control is mandated by the requirement to simultaneously maximise conversion of NO_x and to avoid unreacted ammonia exiting the tailpipe, known as ammonia slip. Maintaining a temperature sufficient to allow optimal NO_x reduction whilst retaining safety is a challenge, particularly at cold start, as well as at any other operating conditions in which exhaust heat is insufficient, such as

prolonged periods of low load or idling. Several techniques have been investigated in the literature to address this issue, including exhaust throttling with increased fuelling, post-injection, physical positioning of the catalyst, and direct electrical heating. However, the use of effective control strategy has not been explored much. In this work, we consider nonlinear model predictive control applied to thermal management in addition to the typical urea control, resulting in the multivariable control of an electric catalyst heater and urea injection. The controller targets maximal NO_x conversion, focussing on the temperature limited conditions, whilst respecting constraints and minimising consumption of fuel and reactants. We show that this achieves highly accurate temperature control of the SCR catalyst which enables optimal NO_x conversion, permitting tuning of the balance between fuel consumption and NO_x emissions across the full range of operating conditions.

1.3.2 Works Under Review and In Progress

[P7] J. Sowman, D. S. Laila, P. Fussey, A. J. Cruden. *A Stability Analysis of Economic MPC for a Selective Catalytic Reduction System.* IEEE Transactions on Control Systems Technology. Submitted 2016.

Model predictive control has shown promise for operation of a selective catalytic reduction plant, a chemical process for reduction of oxides of nitrogen (NO_x) in diesel engine exhaust systems. Aqueous ammonia is injected into the gas stream and is adsorbed onto a catalytic surface before reacting with NO_x, resulting in harmless products. Under-injection of ammonia results in suboptimal NO_x conversion, whilst over- injection causes ejection of harmful ammonia from the tailpipe. Previous applications of predictive control to SCR have been of the tracking MPC variety, in which a target storage of ammonia in the catalyst is determined a-priori; MPC is used to drive the storage to this target. In this work we replace this hierarchical scheme with an economic MPC (eMPC) controller which models tailpipe gas concentrations, bringing the closed loop closer to optimality with respect to cumulative emissions, particularly during transients. We show that this formulation achieves better closed loop performance than the tracking-MPC scheme. Furthermore, we certify stability of the eMPC closed loop, and comment on feasibility of the optimisation programme in the presence of constraints.

[P8] J. Sowman, D. S. Laila, A. Truscott, P. Fussey, A. J. Cruden. *A Real-Time Reference Governor for Ammonia Slip Control in Urea SCR Systems.* SAE International Journal of Engines. Submitted 2017.

Urea selective catalytic reduction (SCR), comprising the injection of urea over a catalyst in the exhaust system, is a key nascent technology for the reduction of oxides of nitrogen (NO_x) from lean-boost direct injection internal combustion engines. The quantity of ammonia injected is critical to maximise NO_x reduction whilst minimising unreacted ammonia exiting the tailpipe, known as ammonia slip.

Feedback control is necessitated by the uncertainty in plant parameters, such as ammonia injector drift, and by response to disturbances, such as upstream NO_x concentration, temperature and flow rate. Classic control methods tuned to respect the ammonia slip constraint lack performance in other parts of the operating range, and provide no guarantee of constraint satisfaction in transient conditions. In this work, we employ a reference governor framework to decouple tuning of the primal controller for performance and enforcing constraint adherence of the closed loop. This has the advantage that the reference governor can be applied above the existing feedback control scheme, only affecting the primal controller when constraint violation is predicted. As well as improved constraint adherence, the results show improved NO_x reduction and natural removal of integrator wind-up. We also demonstrate the controller developed in this work using a rapid prototyping automotive electronic control unit in a hardware-in-the-loop setup, where it displays execution significantly faster than real time as well as upper-bounded runtime and fixed memory requirements

[P9] J. Sowman, D. S. Laila, S. Box, P. Fussey, A. J. Cruden. *In-Use Emissions Testing of Diesel-Driven Buses in Southampton: is Selective Catalytic Reduction as Effective as Fleet Operators Think?*. Intelligent Transport Systems, IET. Submitted 2017.

Despite the continuously tightening emissions legislation, urban concentrations of nitrogen oxides (NO_x) remain at harmful levels. Road transport is responsible for a large fraction, wherein diesel engines are the principal culprits. Turbocharged diesel engines have long been preferred in heavy duty applications, due to their torque delivery and low fuel consumption. Fleet operators are under pressure to understand and control the emissions of their vehicles, yet the performance of emissions abatement technology in real-world driving is largely unquantified. The most popular NO_x abatement technology for heavy duty diesel vehicles is selective catalytic reduction. In this work, we empirically determine the efficiency of a factory-fitted SCR system in real-world driving by instrumenting passenger buses with both a portable emissions measurement system (PEMS) and a custom built telematics unit to record key parameters from the vehicle diagnostics systems. We find that even in relatively favourable conditions, while there is some improvement due to the use of SCR, the vehicles operate far from the design emissions targets. The archival value of this paper is in quantification of real world emissions versus design levels and the factors responsible for the discrepancy, as well as in examination of technologies to reduce this difference.

[P10] J. Sowman, D. S. Laila, P. Fussey, A. J. Cruden. *Hardware Accelerated Implicit Integrators for Real-Time Embedded Model Predictive Control.*

Real-time solution of constrained optimal control problems requires that the plant model be evaluated online. In many practical control applications, this model is

phrased in the form of differential equations. As such, numerical integration is required to predict the future behaviour of the plant over the prediction horizon. This can incur significant computational expense, particularly when the system of differential equations is stiff, and when a long prediction horizon is required to achieve the desired closed loop performance. Implicit integration methods typically outperform other methods when faced with stiff systems, and their amenity to sensitivity generation makes them popular for optimal control applications. Whilst fast solution of the optimisation problem is possible thanks to recent advances in structure-exploiting solvers, the numerical integration can impede real-time implementation. One of the most popular numerical methods for optimal control is direct multiple shooting, in which state trajectories are discretised over the prediction horizon and each segment can be integrated independently. In this work, we exploit this fact to design a customised pipelined hardware accelerator to solve the Newton steps which arise in implicit integration methods. We implement the accelerator on FPGA fabric with a close-coupled microprocessor, wherein acceleration is achieved by using the hardware to remove the computational demand from the microprocessor and thus rendering it free to begin work on the next multiple shooting segment. We demonstrate that the speedup achieved by this topology permits real-time embedded model predictive control of a stiff plant with long horizons.

1.4 Outline

This thesis aims to reconcile considerable academic progress in predictive control techniques with the increasing stringency in automotive emissions legislation. It aims to bring the advantages of predictive control to bear on the problems of NO_x emissions from lean-burn internal combustion engines in order to meet emissions legislation in a manner that is practical from the point of view both of design engineers and vehicle manufacturers. As such, Chapter 2 introduces both the automotive and control aspects of the project, including the current industrial and academic standards and the problems to be overcome in each area. Chapter 3 describes a methodology for examining the real-world emissions from SCR-equipped vehicles and discusses the reasons for which design emission levels are not achieved in practice. The key results are summarised in publication [P9]. Chapter 4 examines dynamic models of SCR systems and constructs several varieties of models, both to use in lieu of a real SCR for hardware-in-the-loop testing, and control oriented models to form the basis of model-based control schemes.

Chapter 5 examines the reference governor approach to model based control as an interim step towards model predictive control, and the results are contained in publication [P8]. Subsequently, in Chapter 6, model predictive control is applied to SCR. A linear formulation is attempted first, which is published in publication [P1]. A nonlinear MPC

approach was tried next, which proved it could maintain optimal performance across a wide temperature range, and these results are published in publication [P2]. Finally, as a result of the low-temperature issues identified in Chapter 3, Chapter 6 introduces thermal management of SCR using a catalyst heater; the results of this application are published in publication [P6]. Chapter 7 examines the observability issues with SCR plants, and the complications introduced by off-the-shelf sensors. It applies various state estimation techniques to extract the information required by the predictive controllers designed in the previous chapter, with results summarised in publication [P4]. Chapter 8 examines a novel approach for nonlinear MPC on cheap multipurpose embedded hardware, that could provide a suitable route to hardware implementation of the controller for automotive use. Chapter 9 examines embedding of MPC, particularly the computational issues introduced by particular features of the dynamic models of SCR. A modern automotive grade system-on-chip device, comprising a microprocessor along with FPGA fabric, is configured to accelerate execution of the MPC algorithm, thus permitting real-time execution. The results of this approach are published in publication [P10].

Finally, Chapter 10 summarises the key findings and contributions. It also examines remaining open questions and discusses possible future directions for research.

Chapter 2

State of the Art

In this chapter, a review of the state of the art in optimal control and automotive aftertreatment control is presented. The review covers the emissions challenges facing the automotive industry and the current technologies that exist to manage them, focusing particularly on aftertreatment. Specifically, management of NO_x and the opportunities for improving abatement technology are the main focus. It then examines the theory of optimal control and the most recent developments in numerical methods to achieve real-time predictive control on embedded or resource-limited hardware, and the limitations of these methods. The chapter concludes by discussing previous applications of predictive control in the automotive industry, and examines the strengths and limitations of these, indicating the challenges that will be faced in this project.

2.1 Emissions Control Technologies

Variation and tightening of emissions legislation has led automotive manufacturers to adopt various technologies over the past two decades in order to control emissions from vehicles. This section examines those relevant to NO_x emissions from lean-burn internal combustion engines and how they interact with each other. Various control strategies that are relevant to the particular abatement technology are also discussed.

2.1.1 Engine Out Emissions

Engine operating conditions have a significant impact on its emissions. The temperature at which lean-burn engines operate dictates a fundamental tradeoff between soot production and NO_x emissions [7]. When older emissions standards were in force, it was possible to meet the soot and NO_x limits by calibration of the engine maps on the testbed during development. This was standard practice until Euro 6/Euro VI, at which

relying solely on engine calibration to meet legislative limits was no longer sufficient and aftertreatment abatement technology became necessary. As a result, this review concentrates on aftertreatment, with the exception of exhaust gas recirculation (EGR) due to its direct impact on NO_x production.

2.1.2 Exhaust Gas Recirculation

The formation of NO_x occurs largely when nitrogen and oxygen are present in a high temperature region, such as during combustion. Since direct injection engines lack a physical throttle, the air to fuel ratio (AFR) often rises significantly above stoichiometric. The air-fuel equivalence ratio (λ) is defined as

$$\lambda = \frac{AFR}{AFR_{stoic}}, \tag{2.1}$$

and this can rise to $\lambda > 10$ in some circumstances in diesel engines [8]. This is the reason that NO_x emissions are much greater from direct injection engines as compared to stoichiometric, where there is little or no excess oxygen.

Exhaust gas recirculation (EGR) has several consequences, some of which help to reduce NO_x production. The recirculated gas has a higher specific heat capacity due to the higher levels of CO_2 and H_2O than the intake air, which results in lower gas temperatures during combustion. Furthermore, the inert exhaust gas replaces some of the intake O_2. The combination of these two effects gives the NO_x reduction that is associated with diesel EGR systems. The lower combustion temperature and the lack of excess oxygen combine to make NO_x formation significantly smaller [9]. Whilst these are the main processes involved in EGR, there are additional ones which are detailed in [9].

One significant downside to EGR is that its applicability is restricted to certain engine conditions. The excess oxygen in the combustion chamber is load dependent: as engine load increases, there is less access to oxygen in the chamber and as a result the formation of soot is more likely to occur. The application of EGR at high loads further reduces oxygen in the inlet air, and so exacerbates the problem. Additionally, the displacement of fresh charge air by exhaust gases reduces the quantity of charge air available for combustion, which limits the engine peak performance. As a result of all of these factors, EGR is typically employed only at low to mid engine loading [8].

2.1.3 Particulate Filtering

With filtration efficiencies up to around 90% [10], the DPF has become the de-facto standard in modern diesel engine vehicles for reducing particulate matter. Filtering of particulate matter by the filter leads to a buildup of particles, which, to retain suitably

low back pressure in the exhaust system, have to be removed once they begin to restrict exhaust gas flow. There are two principal methods of this *regeneration* process; active:

$$C + \alpha O_2 \longrightarrow 2(\alpha - 0.5)CO_2 + 2(1 - \alpha)CO \tag{2.2}$$

and passive:

$$C + \alpha NO_2 \longrightarrow (\alpha - 1)CO_2 + (2 - \alpha)CO + \alpha NO \tag{2.3}$$

where α is a measure of the completeness of the reaction, taking values $\alpha \in [0.55, 0.9]$ in (2.2) and $\alpha \in [1.2, 1.8]$ in (2.3) [10].

The oxygen (O_2) required by the active process is abundant in diesel exhausts due to their lean burn nature. However, the reaction requires high temperatures in order to function: i.e., in the range $400 - 600°C$, which are not typical of diesel exhaust systems. As such, the vehicle electronic systems are required to take additional measures to raise the exhaust gas temperature. These are often varied, but include late injection (post-injection) of diesel fuel and turning on additional electronic loads such as the rear windscreen heater. Further, some vehicles include an additive that is mixed with the fuel in order to reduce the combustion temperature of the trapped soot in the DPF unit.

While passive regeneration is preferable, it requires significant NO_2 in the exhaust stream, which is something to be avoided. However, since the SCR is typically placed downstream of the DPF (see Figure 1.2), if the SCR control strategy is good enough, higher engine out NO_x can be allowed, for example via reduced EGR activity. This higher NO_x will passively remove soot from the DPF and SCR can be relied upon to clean up any remaining NO_x after the DPF to ensure the vehicle remains within emission limits.

2.1.4 Selective Catalytic Reduction

Selective catalytic reduction (SCR) is a technology for removal of oxides of nitrogen (NO_x) from a gas stream. The fundamental principle is to react gaseous NO_x with ammonia (NH_3). The kinetics of this reaction mean that catalysis is required for the reaction rate to be acceptable. This is typically achieved in practice by coating a substrate monolith with the catalyst, known as a *washcoat*. Since NO_x comprises both NO and NO_2, there are several possible reactions that can take place. The most important of these are the *standard SCR* reaction with 1:1 stoichiometry of NH_3 and NO:

$$4NH_3^* + 4NO + O_2 \longrightarrow 4N_2 + 6H_2O \tag{2.4}$$

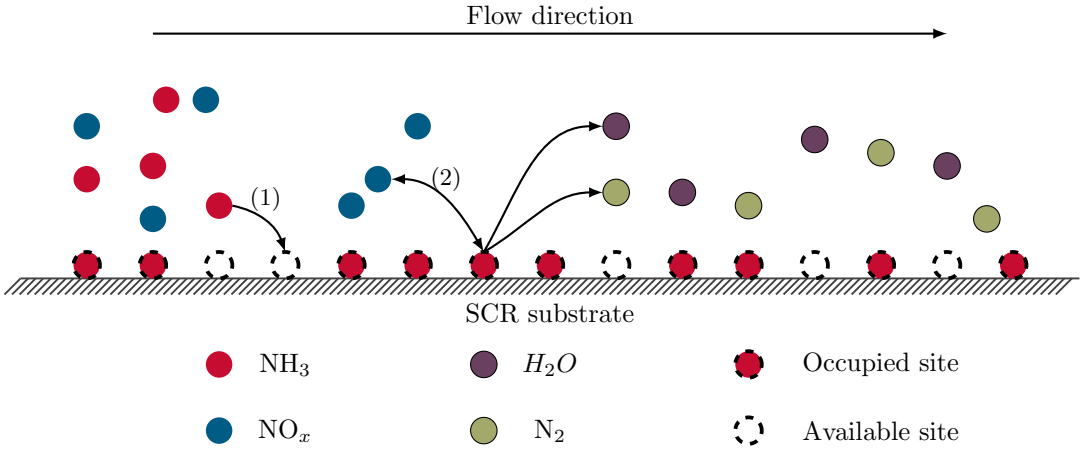


FIGURE 2.1: Depiction of the SCR processes, showing adsorption of ammonia (1) onto the catalyst surface and its subsequent reaction with NO_x (2) to produce nitrogen and water.

and *fast SCR* which consumes equal molar quantities of NO and NO_2 [11]:

$$4NH_3^* + 2NO + 2NO_2 \longrightarrow 4N_2 + 6H_2O \tag{2.5}$$

where the superscript * is used to denote a species on the surface of the catalyst. This distinction is made since the SCR reactions are governed by the Eley-Rideal mechanism, in which the reaction takes place between NO_x in the free stream and adsorbed ammonia [12]. As a result, reactions take place on the surface of the catalyst, and the quantity of ammonia stored on the catalyst is the important quantity rather than the concentration of ammonia in the free stream.

The fast SCR reaction (2.5) is known to have a reaction rate that is at least one order of magnitude higher than the standard SCR reaction, but requires the presence of equal molar quantities of NO and NO_2, and dominates at temperatures above 200°C [13]. These reactions and the Eley-Rideal mechanism are depicted in Figure 2.1, where the gas flow direction is from left to right. NO_x and ammonia enter on the left hand side, after which ammonia is adsorbed onto available sites on the catalyst surface. This adsorbed ammonia then reacts with NO_x to produce nitrogen and water.

Remark 2.1. Both standard and fast SCR reactions consume equal quantities of stored ammonia (NH_3^*) and NO_x in the steady state.

Due to its higher reaction rate, it is desirable to leverage the fast SCR reaction (2.5) where possible. Since it requires equal molar quantities of NO and NO_2, it is necessary to coerce the NO:NO_x ratio to 0.5. However, the typical NO:NO_x ratio from a diesel exhaust is higher than this, which suggests the use of a diesel oxidation catalyst (DOC), to oxidise NO to NO_2.

It is worth mentioning that there exists the equivalent (2.4) which consumes only NO_2, but it is neglected because engines have a NO:NO_x ratio of at least 0.5 under typical operation conditions, meaning the fast SCR reaction (2.5) is always kinetically favoured. Furthermore, the SCR reaction that does not consume oxygen:

$$4NH_3^* + 6NO \longrightarrow 5N_2 + 6H_2O \tag{2.6}$$

is very slow, and is also therefore often excluded from catalyst models [14].

Due to considerations of safety and ease of transport, gaseous ammonia is not directly stored within the vehicle. Instead, an aqueous urea solution (AUS) (such as AdBlue®) is stored. The heat in the exhaust stream causes hydrolysis:

$$CO(NH_2)_{2(aq)} \longrightarrow CO(NH_2)_{2(l)} + H_2O \tag{2.7}$$

which then decomposes, again due to thermal effects, to ammonia and isocyanic acid:

$$CO(NH_2)_{2(l)} \longrightarrow NH_{3(g)} + HNCO_{(g)} \tag{2.8}$$

Isocyanic acid hydrolyses easily with water vapour (from (2.7) and from the combustion process):

$$HNCO_{(g)} + H_2O_{(g)} \longrightarrow NH_{3(g)} + CO_{2(g)} \tag{2.9}$$

The reactions (2.7)-(2.9) combine to form the overall urea decomposition:

$$CO(NH_2)_2 + H_2O \longrightarrow 2NH_3 + CO_2 \tag{2.10}$$

meaning that one mole of urea generates two moles of ammonia [11]. The ammonia is then adsorbed onto the surface of the catalyst, after which it can take part in the SCR reactions required for NO_x conversion:

$$NH_3 \Longleftrightarrow NH_3^* \tag{2.11}$$

where the two way arrow denotes that this reaction can take place in both directions – i.e. ammonia can desorb from the catalyst surface back into gaseous ammonia if the diffusion gradient is in this direction. Additionally, oxidation of stored ammonia can take place. This is particularly the case at high temperatures, due to the abundance of oxygen in the exhaust stream:

$$4NH_3^* + 3O_2 \longrightarrow 2N_2 + 6H_2O. \tag{2.12}$$

An important phenomenon in the control of SCR is so-called *ammonia slip*. Not all of the ammonia injected into the reactor will adsorb onto the catalyst surface; part of it escapes into the environment, and some of the adsorbed ammonia will spontaneously desorb and leave the catalyst. This is particularly undesirable due to the environmental toxicity of ammonia. As such, there is a limit placed on the maximum ammonia slip that is permitted from the system (10 ppm for Euro 6 diesel) and minimising this is a key objective of the control strategy.

Low Temperature Operation of SCR

Relatively little attention has been paid to operation of the SCR below operating temperature (also known as the *light-off* temperature), since the NO_x conversion is small. The extension of map-based controllers to lower temperatures is typically not economically viable. As a result, disabling ammonia injection during this period has become a typical approach. This eschewing of NO_x conversion during cold start is becoming increasingly inadvisable. Firstly, NO_x emissions in urban environments are an increasing health concern. Low road speeds and start-stop conditions make urban driving conducive to low catalyst temperatures. Secondly, journeys are often short within cities, and will terminate before the SCR reaches operating temperature. Thirdly, continually increasing engine thermal efficiency reduces exhaust gas temperatures even further, lengthening the heating time for the catalyst. As a result, optimising NO_x reduction in the cold start region could have significant impact on diesel emissions, particularly in locations where it is most beneficial. Therefore, cold start SCR optimisation is a problem that has only recently become worthy of research.

There have been several suggested solutions in the literature for improving this situation. The first and most obvious is the catalytic material used for the SCR reactor. It is well known that some materials exhibit better low temperature NO_x reduction performance than others. Several possible catalysts were investigated by [15]. The chemical background to suboptimal temperature SCR operation is described in [16], which exposes the different reaction pathways that appear when the catalyst is significantly below its lightoff temperature (from 100°C upwards). They note that water adsorption onto the catalyst surface is responsible for a lot of the temperature effects that are observed in low temperature operation. Also noted by [11] is the fact that if AUS is injected rather than pure ammonia, the lack of heat in the exhaust gas hinders the thermal decomposition into ammonia, which can prevent the SCR control system working as intended.

The physical layout of the exhaust system is important, since moving the SCR catalyst closer to the engine means that less heat is lost in the transmission of the exhaust mass

flow from the engine to the SCR. This is studied under a mathematical framework in [17], which aims to find the optimal packaging for the exhaust system layout to prioritise operation of the SCR. Another option to bring the catalyst to operating temperature in a shorter period is explored by [18], which uses an electrical heater to heat the SCR catalyst. The results are impressive, but the additional electrical load of the heater implies a fuel penalty.

An interesting area for further research is into the design of a controller that attempts to optimise performance of the SCR all the way from cold start to operating temperature, whilst respecting the control objectives and constraints – this topic has, as yet, seen no significant research.

SCR on Filter

Recently, there has been interest in the possibility of combining the SCR and DPF units [19, 20, 21], by coating the walls of the DPF with the SCR catalyst; the resulting device is known as SCR on Filter (SCRF). This arrangement allows the SCRF to be mounted further upstream (i.e. towards the engine) which results in increased regeneration efficiency and quicker startup of the SCR device from a cold start. There are, of course, challenges in combining the unit – reactions in the DPF and SCR are not independent (see passive regeneration in Subsection 2.1.3), and the thermal stresses on the monolith are greater due to proximity to the engine and active regeneration events. Nevertheless, more than 90% deNO_x average steady state efficiency has been demonstrated over a large temperature range [11].

2.1.5 Classical Control of SCR

Industry standard approaches for automotive control have typically been the basis of SCR control, which are typically feedforward map-based controllers whose maps are calibrated on the testbed over test cycles during engine development. There has been speculation that closed-loop control will be required in the future to cope with the combination of stricter emissions requirements, highly transient plant operation, and variation in sensors and actuators both during production and over the life cycle of the aftertreatment system [22]. As a result, elements of feedback control have begun to make their way into industry standard SCR controllers, including PI [23, 24, 25] and sliding mode [26, 27]. The calibration effort required to tune a map-based controller is extensive, and constraints cannot be enforced in a systematic manner. The response of a linear controller (such as PI) means that steady states must typically be placed further from constraint boundaries to achieve sufficiently low probability of constraint violation. All of these reasons comprise the motivation for exploring model-based control for SCR.

More advanced control methods, such as model based approaches, have recently begun to appear in the literature. This includes the reference governor, which provides a way to integrate a model-based supervisory layer into existing feedback control laws, as well as model predictive control, which is introduced and discussed in Section 2.4.

2.2 Modelling of SCR

An accurate dynamic model of the SCR system is important for several reasons. Firstly, repeated testing on real hardware is expensive and time consuming. An accurate model can be used in model-based design in lieu of real hardware. This is not only cheaper, but can be expected to run faster than real time. Secondly, model-based control schemes require a model of the plant under control. This is typically physically inspired, or achieved by system identification.

The standard physical architecture of the SCR reactor is a chamber containing several axial cells in a honeycomb structure, the walls of which are coated with the catalytic substance [28]. The SCR system is a chemical process [29], the relevant chemical reactions during typical diesel operation and during low temperature operation (e.g. during cold start sequences) respectively are provided in [30] which constructs the SCR models. These models are a good starting point for modelling of a generic SCR since they are founded in physical principles but validated against measured test data from real engines. These works build on earlier research, mostly by the same authors, in modelling the fundamentals of the SCR process (these are cited in [29] and [30]). It is suggested in [30] that the developed models could be feasible for automotive application, but points out that the fidelity of the models during transient operation could cause issues in terms of computational load; this is something to be examined for the purposes of implementation in a vehicle environment. It is also mentioned that the dosing strategies may require knowledge of an optimal ammonia surface coverage, something which is likely therefore to be a requirement in the calibration of the model-based controller. Whilst no data is provided in [30], the calibration values of the various parameters in the rate equations for the SCR are provided in [31]. However, the rate equations are different and this is a full chemical model which is not designed for real-time simulation. Nevertheless, it provides a base calibration; having such a starting point available can make calibration for a different catalyst much faster.

The models developed in [29] and [30] are physically inspired. As such, in an attempt to apply model-based predictive control to an SCR system, [28] goes into significant detail about the chemical reactions involved and modelling of the physical system, including details of each of the relevant reactions and an expression for their simulation. The model resulting from this approach is actually two linked submodels, one describing the thermal behaviour of the gases and SCR reactor, the other describing the chemical processes in

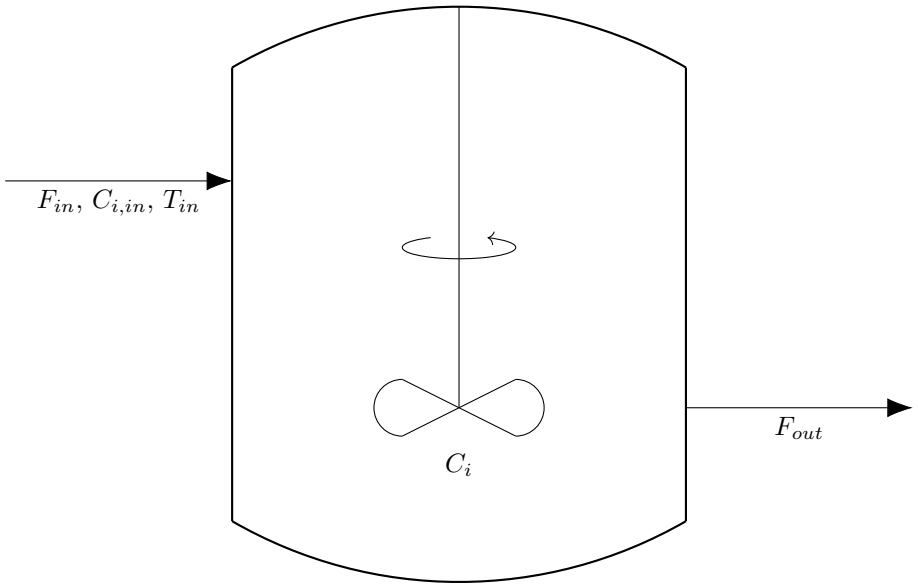


FIGURE 2.2: Schematic of a CSTR reactor

the conversion of NO_x. This is likely to be useful in terms of vehicle implementation since the models can be executed at different time scales to save computational effort. Finally, the models developed in [28] are validated against a heavy goods vehicle (HGV) automotive application.

It is important to define the quantities relevant to the control process, such that time is not wasted calculating (for example) the concentrations of exhaust gases which are not relevant to the SCR system, or about which the controller does not care. The models developed in [28] use the composition of gas at the SCR inlet (in terms of the mole fraction of several species of interest) as an input to the model, which is determined from the inlet air to fuel ratio (AFR), a reading from an upstream NO_x sensor, plus the estimated exhaust flow rate and NO_x fraction exiting the DPF. Herein lies the importance of state estimation to optimal performance of the controller.

As discussed at the beginning of Chapter 4, improvements are likely to be conceivable in the modelling, estimation and control of the entire exhaust and aftertreatment system, and one fairly complete attempt at doing this is found in [32] which models the SCR system as a series of continuously stirred tank reactors (CSTR). The CSTR is a variety of idealised chemical reactor, depicted schematically in Figure 2.2, where the mixture is assumed to be perfectly mixed such that outlet composition is the same as the composition in the reactor. Specifically, the author uses six tanks, which afford the right compromise of accuracy and computational demand for that particular application. This may or may not be the case in the model implementation developed for the predictive controller, depending on the hardware to be used and any computational efficiency gains that can be found from tailoring of the algorithms. Another CSTR modelling attempt is made in [33], which provides a full four state nonlinear state space model of the SCR

system. Along with the subsequent paper [34], it also validates these models against a testbed engine.

During modelling of the SCR system and other exhaust system components, an offline parameter estimation is likely to be required if the models are derived from or calibrated with real data. Many approaches to parameter estimation are available, including the MATLAB System Identification toolbox. For nonlinear systems such as the ones that are present in the exhaust aftertreatment system, a promising set-based approach is given by [35].

2.3 Exhaust Gas Sensors, Measurement and Estimation

2.3.1 Sensors and Measurement

In order for a feedback based control system to function, some measurement must be made of the plant outputs to be controlled. As can be seen in Figure 1.2, a typical architecture for an SCR system is to have an upstream NO_x sensor to measure the gas composition entering the catalyst, coupled with a NO_x and ammonia sensor pair at the tailpipe. The ammonia sensor is often not present in real world systems. Design and manufacture of these sensors presents a significant challenge for several reasons: they are required to last many thousands of hours in a harsh and varying temperature exhaust environment, during which they must sense very small concentrations of gases (typically on the order of 100s of ppm for NO_x and 10s of ppm for ammonia). If the control system is operating well, at the tailpipe, these concentrations fall very close to zero, especially in steady state.

There are several types of sensors available for detection of the chemical species relevant to SCR control. Organic sensors were the first available but have not found success in automotive systems due to their intolerance of extended high temperature operating conditions. Yttria Stabilised Zirconia (YSZ) sensors have been more successful in this respect, and such sensors are typically used in modern SCR systems [36, 37]. These sensors are equally sensitive to NO and NO_2, thus resulting in a so-called *total* NO_x sensor. Should the control system need to know the NO:NO_x ratio in the catalyst inlet or outlet gas, this needs to be estimated by knowledge of the DOC characteristics (which are typically very temperature dependent). The principle disadvantage of such sensors is that they typically exhibit cross sensitivity between NO_x and ammonia, necessitating filtering techniques for estimation of the true NO_x and ammonia concentrations [38]. Recently, there has been interest in doping the electrodes with tungsten trioxide (WO_3) in order to improve the sensitivity [39, 40].

A combined NO_x and ammonia sensor suite is developed by [41], which uses WO_3 films of different thicknesses is determined that provides the best selectivity of the two sensors

to their respective gases. It is likely that a combination of choosing sensors with minimal cross sensitivity and state estimation techniques will be required to achieve the sensor accuracy that is required for improved control of the SCR process.

In terms of production sensors that are currently used on vehicles, several major manufacturers are now producing NO_x sensors, including Continental and NGK. Typically, these are YSZ sensors which communicate via CAN or a similar interface. Off-the-shelf units can typically sense up to around 1500ppm NO_x and have an accuracy of $\pm$10ppm. It is therefore clear that if these sensors are to be used in high efficiency SCR systems, some filtering and estimation will be required in order to improve accuracy at low NO_x concentrations. Other sensors used in current commercial SCR applications are O_2 and temperature sensors, which have been standard equipment on production vehicles for many years and can therefore be considered a mature technology. Ammonia sensors are typically not fitted in current industry standard SCR systems. As a result, they are typically expensive and have not seen as much development. High efficiency SCR may require the use of such sensors [42], in which case it can be expected that the price of these units will decrease and availability will increase over the coming years. One recently released example by Delphi senses NH_3 in the range 0-100ppm and avoids cross-sensitivity to other species, such as NO_x [43].

2.3.2 State Estimation

It has been discussed that the SCR reactor, and indeed the exhaust aftertreatment system as an entirety, is a complex nonlinear system (see equation (10) in [34]). Knowledge of certain quantities in the SCR is vital to achieving good control performance. One of the most important of these is the estimation of ammonia coverage on the SCR catalyst surface [30], for which there is currently no direct sensor available. The only alternative is to use the rest of the sensor suite to estimate this parameter. A controller for an SCR system that incorporates a feedback element will also require measurements from sensors. As has been previously noted, currently commercially available NO_x sensors are typically equally sensitive to NO and NO_2, but additionally cross sensitive to ammonia, such that the measured NO_x concentration is [44]:

$$C_{NO_x,measured} = C_{NO} + C_{NO_2} + k_T C_{NH_3} \tag{2.13}$$

where k_T is a temperature dependent cross sensitivity factor, whose characteristic will typically differ across different sensors. Ammonia sensors are typically not fitted to production vehicles with SCR, so at present, the goal of the estimator is to estimate the true NO_x and NH_3 concentrations from the concentration measured by the NO_x sensor. The lumping of NO and NO_2 into a single NO_x concentration is typical of current SCR control practices. However, as has been shown, the NO:NO_x ratio will estimate the

rates of standard- and fast-SCR reactions that take place in the catalyst. As a result, extension of the observer in order to determine the true NO and NO_2 concentrations has been the subject of recent research, including [45] and [46], the former of which shows that separation of these species in the control oriented model can lead to performance improvements. It is also noted in [45] that tracking NO_2 can be useful to improve the estimation of DPF soot loading, since this species is involved in passive regeneration (2.3).

The most commonly used state estimation framework is the Kalman filter (KF) which is a recursive state estimator [47]. For a linear system such as:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + w_k \\ y_k &= Cx_k + v_k \end{aligned} \tag{2.14}$$

where the state and output sequences are subject to additive white Gaussian noise (AWGN) (w_k and v_k, known as the *process* noise and *measurement noise* respectively), the Kalman filter provides the *optimal* estimate of the system state x_k. The outputs of the system as measured by the sensors are given as y_k.

The Kalman filter provides an estimate of the system state x_k, denoted by $\hat{x}_k$, at each time step k via use of the system model (2.14) as well as measurements y of the system output. The filter accomplishes this recursion in a two step process known as *predict* and *correct*, where prediction is carried out using the model information to find the *a priori* estimate of the state and covariance matrix, which is then corrected using measurements to provide the *posterior* state estimate $\hat{x}$ and covariance.

$$\begin{aligned}
\text{Predict:} \quad \hat{x}_k^- &= Ax_{k-1} + Bu_k && \textit{a priori} \text{ state estimate} && (2.15a)\\
P_k^- &= AP_{k-1}A^T + Q_k && \textit{a priori} \text{ covariance} && (2.15b)\\
\text{Correct:} \quad r_k &= y_k - C\hat{x}_k^- && \text{innovation} && (2.15c)\\
S_k &= CP_k^-C^T + R_k && \text{covariance residual} && (2.15d)\\
K_k &= P_k^-C^TS_k^{-1} && \text{Kalman gain} && (2.15e)\\
\hat{x}_k^+ &= \hat{x}_k^- + K_kr_k && \textit{a posteriori} \text{ state estimate} && (2.15f)\\
P_k^+ &= (I - K_kC)P_k^- && \textit{a posteriori} \text{ covariance} && (2.15g)
\end{aligned}$$

This sequence is shown schematically in Figure 2.3, where x^- denotes an *a priori* estimate and x^+ denotes a *posterior* estimate. With Q_k and R_k the (possibly time varying) process and measurement noise covariances are denoted. The confidence in the state estimate is tracked in the *error covariance matrix* P.

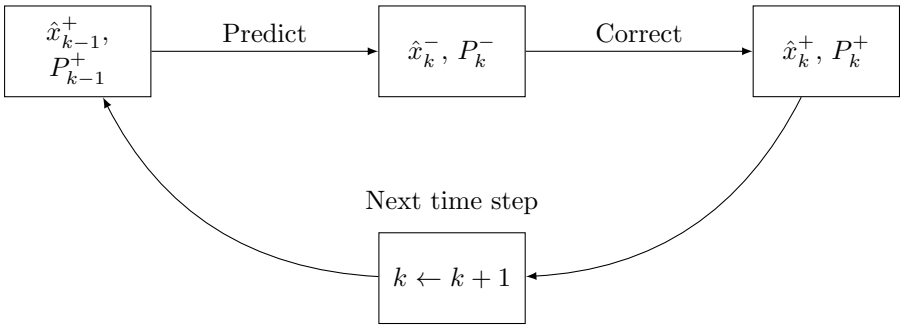


FIGURE 2.3: The Kalman Filter *update* and *correct* recursive process to generate the optimal linear estimate of the system state $\hat{x}$

The need for NO:NO_x and ammonia coverage estimation has been the subject of recent research. However, the nonlinearities in the SCR dynamic equations mean that the linearity assumption of the Kalman Filter is invalid. A naive yet effective answer to this is to successively linearise the nonlinear system as time progresses. This is known as the Extended Kalman filter (EKF), which for nonlinear differentiable discrete time system:

$$\begin{aligned} x_{k+1} &= f(x_k, u_k) \\ y_k &= h(x_k) \end{aligned} \tag{2.16}$$

the linearisation is obtained using

$$A_k = \left.\frac{\partial f}{\partial x}\right|_{\hat{x}_{k-1}, u_k} \tag{2.17a}$$

$$C_k = \left.\frac{\partial h}{\partial x}\right|_{\hat{x}_k} \tag{2.17b}$$

to let A and C become time varying due to the relinearisation. Equations (2.15a), (2.15b) and (2.15c) instead become:

$$\hat{x}_k^- = f(\hat{x}_{k-1}^+, u_k) \qquad \textit{a priori} \text{ state estimate} \tag{2.18a}$$

$$P_k = A_k P_{k-1} A_k^T + Q_k \qquad \textit{a priori} \text{ covariance} \tag{2.18b}$$

$$r_k = y_k - h(\hat{x}_k^-) \qquad \text{innovation} \tag{2.18c}$$

Applications of the EKF to ammonia storage estimation in SCR can be found both in the literature and in industrial practice, for instance in [46, 48, 49], the latter two of which also employ it to distinguish NO and NO_2. These works show that the EKF performs reasonably well, but in the case of highly nonlinear systems, a first order linearisation

as used in the EKF is likely to have only a small region of validity, meaning some loss of performance will be encountered. The Jacobians required by the EKF are typically stored in analytical form and evaluated online. For more complex systems, these can become large and thus nontrivial to find closed form expressions, though this is often aided by the use of a computer algebra system (CAS). Nonlinear estimators based on classical methods (e.g., sliding mode) can also be found in the literature [50, 51], though model-based approaches tend to be more common. As such, if the EKF is not a viable solution, other observers such as the Unscented Kalman Filter (UKF) may well be more so. The total storage capacity of the catalyst has also been the subject of estimation, as it decreases over time due to ageing and poisoning [48].

It is a generally accepted requirement for vehicle applications that performance of the aftertreatment system be maintained over time, often over the full useful service life of the vehicle. As such, estimation of ageing of the catalytic substrate and its effect on performance is vital. The EKF approach to age-related catalyst degradation appears to work well in the conditions tested in [48], however the same caveats apply as with coverage estimation. Further, this work does not consider how the observer performs in the face of model uncertainty – this requires further examination and testing.

Estimation and correction of NO_x sensor cross-sensitivity has seen much recent work due to the development of the sensors and requirements for accurate NO_x control. These can be model based [52], fuzzy logic based [53], or $\mathcal{H}_2/\mathcal{H}_\infty$ based [54]: these works aim to estimate the cross sensitivity factor directly. A tangential approach is to estimate the quantity of interest, usually the catalyst ammonia storage, but include the cross-sensitivity as an unknown [50]. Some works on SCR control have included cross-sensitivity rejection to attempt to bring the control problem closer to reality [55].

2.4 Predictive Control

Classical feedback control dates back to the late 19th Century in the design of speed regulators, called *governors*, for steam engines. The fundamental concept is comparison of the output of a plant to a predefined reference, to produce an error signal $e(t)$. The control action $u(t)$ is formulated as a function of the error signal along with the integral and derivative of the error signal, such that:

$$u(t) = k_p\, e(t) + k_i \int_0^t e(t)\,\mathrm{d}t + k_d \frac{\mathrm{d}}{\mathrm{d}t} e(t) \tag{2.19}$$

hence the common term for these controllers is proportional-integral-derivative (PID). The gains k_p, k_i and k_d, known as the proportional, integral and derivative gains respectively, are parameters to be tuned to achieve the control objectives in a reasonable

manner (typical criteria are stability, a settling time within n seconds and an overshoot of less than $x\%$). A sizeable portion of classical control theory concerns the determination of the optimal PID gains for a given application.

Predictive control represents a paradigm shift in the approach to the problem of control, and it has proved highly successful. It was developed almost simultaneously by two research institutions. In France, Jacques Richalet developed *Model Predictive Heuristic Control* [56], and subsequently founded ADERSA, which became the first entity to commission an industrial application of the control process. In Texas, Cutler & Ramaker of Shell Oil Corporation developed *Dynamic Matrix Control* [57]. It is no surprise that these initial applications were in the process industry, which operate on relatively slow timescales, meaning computational requirements are therefore less onerous. Furthermore, they operate on such a scale that marginal improvements in efficiency have considerable financial impact. These applications were typically a tiered hierarchy in which predictive control played only a small part. However in recent years, it has begun to find its way into faster and more fault intolerant applications, such as the aerospace [58] and automotive sectors [59]. A 2003 survey reported more than 4,500 industrial applications [60]. Advances in computational power, particularly in the embedded electronics world with faster microprocessors and a fall in the cost of programmable logic devices, has led to interest in predictive control in areas where the computation requirement has previously been considered prohibitive. Perhaps unsurprisingly, many formulations of the technique have evolved over the years, often resulting in confusion in the literature. The dynamic matrix control formulation of Cutler & Ramaker is still in common use, as are generalised predictive control (GPC) and predictive functional control (PFC), amongst others. These formulations typically have something specific to their algorithmic or mathematical structure in order to distinguish them from other flavours. The formulation that will be described here is a generic formulation known as model predictive control (MPC), also known as *receding horizon control* (RHC), full details of which are given in [61]. Despite this diversity, all predictive controllers have certain elements which distinguish them from other techniques. Fundamentally, these are an internal model for prediction of plant response faster than real time, a receding horizon approach to generation of optimal inputs, and in particular, an online optimisation algorithm which determines the next and future control inputs.

Predictive control boasts several powerful and compelling advantages over classical control approaches, including a natural and intuitive extension to multivariable control problems, a solution algorithm that is aware of actuator constraints and a nonlinear controller response near constraints that permits operation very close to such boundaries [62]. This last trait is particularly profound – it is characteristic of many physical systems, including SCR, that optimal operation of the plant lies on or infinitesimally close to a constraint boundary. In the specific case of SCR, optimal NO_x reduction occurs at the ammonia slip limit (see Subsection 2.1.4).

2.4.1 Review of MPC

Predictive control has its roots in classical control theory, particularly in optimal control. The optimal control paradigm deals with the concept of finding a control law which satisfies a particular optimality condition, known as the *cost function*. This is a function of the states and inputs of the plant, or the plant outputs which are a function of the states, which is chosen in order to drive the plant to a predetermined state. For a continuous time system:

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{2.20}$$

an *infinite horizon* cost function is defined as:

$$J(x,u) = \int_{t_0}^{\infty} x^T Q x + u^T R u + 2x^T S u \, \mathrm{d}t. \tag{2.21}$$

Here the matrix properties of *positive definiteness* or *positive semidefiniteness* are introduced:

$$\begin{aligned} \text{Positive definite: } Q \succ 0 &:= x^T Q x > 0 \ \ \forall\, x \neq 0 \\ \text{Positive semidefinite: } Q \succeq 0 &:= x^T Q x \geq 0 \ \ \forall\, x \neq 0 \end{aligned} \tag{2.22}$$

It is typically ensured that $S \succ 0$, $Q \succ 0$ and $R \succeq 0$ in (2.21), resulting in the desirable and important property that the cost function is *quadratic* and *convex*. This means that the solution to the problem, which will give the optimal control input sequence, is both *unique* and *global*. A *state feedback* control law is sought, of the form:

$$u(t) = -Kx(t) \tag{2.23}$$

where $x(t)$ is the current state of the system and K is the controller gain. In the case described here with a linear system and a quadratic cost function, the solution is given by:

$$K = R^{-1}(B^T P + S^T) \tag{2.24}$$

in which P is the solution to the algebraic Riccati equation (the derivation of this is outside the scope of this report):

$$A^T P + PA - (PB + S)R^{-1}(B^T P + S^T) + Q = 0. \tag{2.25}$$

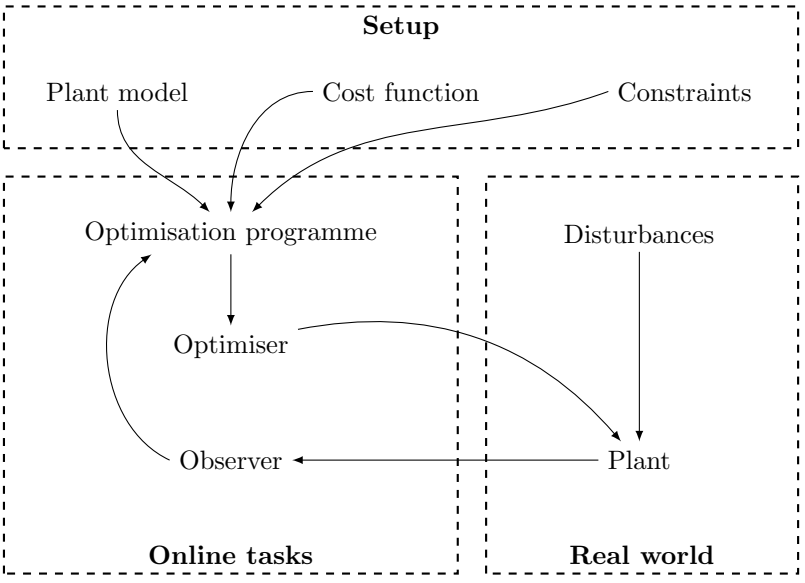


FIGURE 2.4: Ingredients of model predictive control

The solution laid out here is known as the *linear quadratic regulator* (LQR) and is by far the most commonly used technique in optimal control, providing a huge amount of flexibility. It is important to note that there are no constraints present on the inputs $u(t)$ and the states $x(t)$, the only tuning is via the weighting matrices S, Q and R in (2.21). It is this lack of constraints that means the value of the state feedback matrix K is time invariant and can be found offline. This ceases to be true in the case of MPC.

Predictive control aims to address some of the practical issues in optimal control. Numerical evaluation of the infinite horizon cost functional (2.21) is not feasible; this is instead approximated by repeated solution of a finite horizon cost function:

$$J(x,u) = \int_{t_0}^{t_f} x^T Q x + u^T R u + 2x^T S u \, \mathrm{d}t \tag{2.26}$$

where t_f is known as the prediction horizon. Of equal importance is the inclusion of constraints on the inputs u and the states x. By constraining the solution space, the closed form analytical expression for the state feedback law K (2.24) is lost. The canonical formulation of predictive control begins with a linear dynamic discrete time system:

$$x_{k+1} = Ax_k + Bu_k \tag{2.27a}$$

$$y_k = Cx_k \tag{2.27b}$$

where $x \in \mathcal{X} \subseteq \mathbb{R}^{n_x}$ is a state vector and $\mathcal{X}$ is the set of admissible states, $u \in \mathcal{U} \subseteq \mathbb{R}^{n_u}$ is the input vector and $\mathcal{U}$ is the set of admissible inputs, and y is the output vector. The

prediction horizon t_f is divided into N steps over which the system is simulated – it is over this horizon that the cost function is minimised:

$$V(x,u) := \sum_{k=0}^{N} l(x_k, u_k) + E(x_N) \tag{2.28a}$$

$$l(x_k, u_k) = x_k^T Q x_k + u_k^T R u_k \tag{2.28b}$$

$$E(x_N) = x_N^T P x_N \tag{2.28c}$$

under the constraint that when $k = 0$, x is equal to the measured value of the state and x_k are the states at time k, with x_N being the final state (at the end of the prediction horizon), whilst u_k are the inputs at time k. Notice that the stage cost $l(\cdot,\cdot)$ and the terminal cost $E(\cdot)$ are quadratic functions of their arguments. Writing the constraints in a linear fashion, the optimisation problem is:

$$\underset{u}{\text{minimise}} \quad V(x,u) \tag{2.29a}$$

$$\text{subject to} \quad G \begin{bmatrix} x \\ u \end{bmatrix} = g \tag{2.29b}$$

$$H \begin{bmatrix} x \\ u \end{bmatrix} \leq h \tag{2.29c}$$

Note that this is a quadratic programme (QP), since the cost function is convex and quadratic and the constraints linear. Furthermore, this QP is in *uncondensed* form, meaning that it is over-parameterised, the states x_i appear as decision variables, yet each state is uniquely determined by the previous state and input thanks to (2.27a). This results in very sparse and highly structured problem data, features which can be exploited for fast solution of the problem. Alternatively, the system (2.29) can be condensed to produce a much smaller but dense problem. To do so, rewrite (2.28) in matrix notation (see [61] for the details of this derivation):

$$V(x,u) = x_0^T Q x_0 + x^T \bar{Q} x + u^T \bar{R} u \tag{2.30}$$

where x_0 is the initial state, and x and u are the vectors of future states and inputs respectively. Now the states x_k can be substituted in (2.28) by recursively substituting (2.27a) for each time index k. In doing this, the states are eliminated, making $V(\cdot)$ a function of only the inputs u and the initial condition x_0 to create:

$$V(x,u) = \frac{1}{2}u^T\Gamma u + u^T\Psi x + x^T(Q + \Phi^T\Omega\Phi)x \tag{2.31a}$$

$$\Gamma := 2(\bar{R} + B^T\bar{Q}B) \tag{2.31b}$$

$$\Psi := 2B^T\bar{Q}A \tag{2.31c}$$

with $R \succ 0$, $P \succeq 0$ and $Q \succeq 0$ guaranteeing convexity of (2.31a). The QP now becomes the small and dense problem:

$$\underset{u}{\text{minimise}} \quad f(u) := \frac{1}{2}u^T\Gamma u + u^T\Psi x \tag{2.32a}$$

$$\text{subject to} \quad Hu \leq h \tag{2.32b}$$

$$Gu = g \tag{2.32c}$$

where the function to be optimised, $f(\cdot)$, is equivalent to $V(x,u)$ in (2.31) but lacking the term that does not involve u since this does not affect the optimisation. The constraints are written the forms of equality and inequality. Note that they are written in the form of constraints in the inputs u – state constraints can also be included in this format (see [61] for details). This format is relatively flexible, with algebraic tricks both saturation-type as well as rate-type constraints can be imposed (i.e., constraints on how quickly the input or state can change over time).

At each time step, the optimal input set is found using the quadratic programme solution, and the first lement of the inputs is applied to the plant. The rest of the sequence is discarded (since at the next time step, new information is available from sensors, and thus the calculated input is no longer necessarily optimal) and the optimisation runs again at the next time step (this is shown in Figure 2.5). For this reason, MPC is also known as *receding horizon control* (RHC).

2.4.2 Example

As an example to demonstrate this theory, take a continuous time linear system describing an undamped mass $m = 1\ kg$ on a spring of stiffness $k = 1\ Nm^{-1}$. The objective will be to drive it to a position of our choice using an MPC controller which controls an actuator that applies a force to that mass, and using a sensor that measures the position of the mass (but not its velocity), as shown in Figure 2.6.

Assume no disturbances and no model uncertainty in this case, and begin with the governing ordinary differential equation of motion:

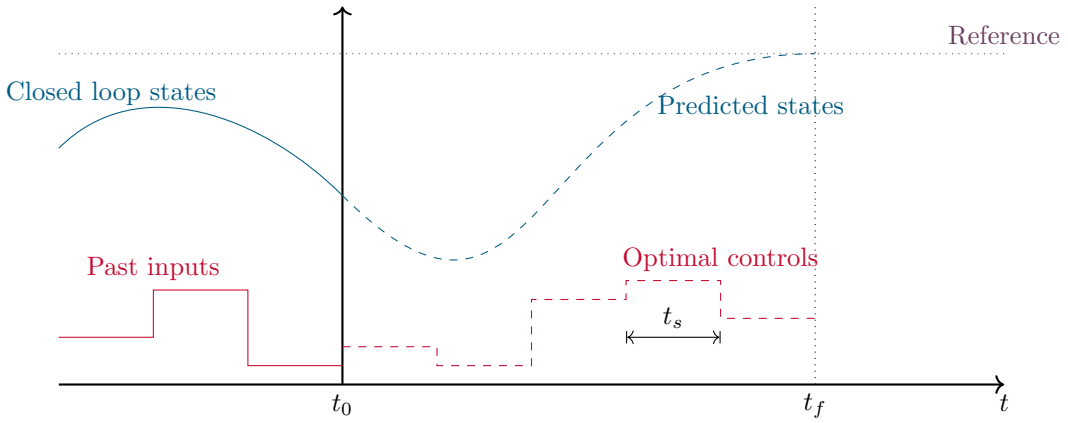


FIGURE 2.5: MPC actions at each time step. The optimal input sequence is calculated to bring the closed loop output to its reference value.

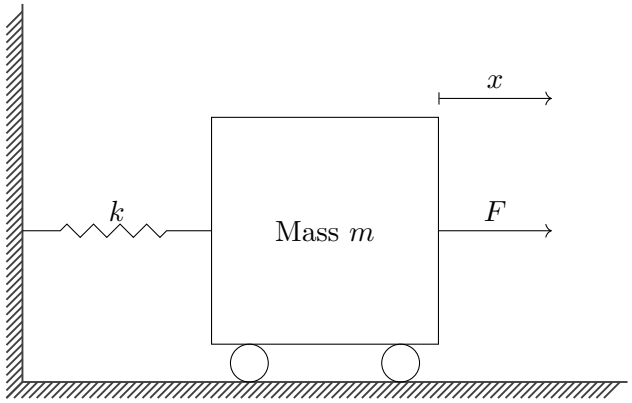


FIGURE 2.6: Schematic of the controlled mass-spring system employed in the MPC example.

$$\ddot{x}(t) + \frac{k}{m}x(t) = u(t) \tag{2.33}$$

which can be converted into the standard state space form for continuous time systems:

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{2.34}$$

where $\dot{x}$ denotes the time derivative of x, by defining the state vector:

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{2.35}$$

where x_1 and x_2 are the position and velocity of the mass respectively. This provides the state space form of the mass spring damper:

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ -\frac{1}{m} \end{bmatrix} u(t) \tag{2.36}$$

where $u(t)$ is scalar in this case since there exists only one input. Now choose $k = m = 1$ to produce the continuous time numerical state space model:

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ -1 \end{bmatrix} u(t) \tag{2.37}$$

and use MATLAB's `c2d` tool with the zero order hold (`zoh`) option and a time step of $T_s = 0.1s$ to convert this system to a discrete time representation as used for the MPC controller. This yields the system:

$$\begin{aligned} x_{k+1} = Ax_k + Bu_k \Longleftrightarrow x_{k+1} &= \begin{bmatrix} 0.1 & 1 \\ 1 & -0.1 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 0.1 \end{bmatrix} u_k \\ y_k = Cx_k \Longleftrightarrow y_k &= \begin{bmatrix} 1 & 0 \end{bmatrix} x_k \end{aligned} \tag{2.38}$$

To keep the algebra straightforward and clear, take equal prediction and control horizons of $N = 3$ in this simple example. The cost function can now be formed from (2.30), keeping the (scalar) weights P, Q and R symbolic here for clarity:

$$V(x,u) = x_0^T Q x_0 + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}^T \begin{bmatrix} Q & 0 & 0 \\ 0 & Q & 0 \\ 0 & 0 & P \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} u_0 \\ u_1 \\ u_2 \end{bmatrix}^T \begin{bmatrix} R & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & R \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \end{bmatrix} \tag{2.39}$$

in which the states $x_i, i = \{1, 2, 3\}$ can be replaced using the system model to produce the MPC cost function (2.31) which is a function only of the future inputs u and the current state x, as required:

$$
\begin{aligned}
V(x_0, u) = \begin{bmatrix} u_0 & u_1 & u_2 \end{bmatrix} \Bigg(& \begin{bmatrix} R & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & R \end{bmatrix} \\
& + \begin{bmatrix} B & 0 & 0 \\ AB & B & 0 \\ A^2B & B & B \end{bmatrix} \begin{bmatrix} Q & 0 & 0 \\ 0 & Q & 0 \\ 0 & 0 & P \end{bmatrix} \begin{bmatrix} B & 0 & 0 \\ AB & B & 0 \\ A^2B & B & B \end{bmatrix} \Bigg) \begin{bmatrix} u_0 \\ u_1 \\ u_2 \end{bmatrix} \\
& + 2 \begin{bmatrix} u_0 & u_1 & u_2 \end{bmatrix} \begin{bmatrix} B & 0 & 0 \\ AB & B & 0 \\ A^2B & B & B \end{bmatrix} \begin{bmatrix} Q & 0 & 0 \\ 0 & Q & 0 \\ 0 & 0 & P \end{bmatrix} \begin{bmatrix} A \\ A^2 \\ A^3 \end{bmatrix} x_0 \\
& + x_0^T \left(Q + \begin{bmatrix} A & A^2 & A^3 \end{bmatrix} \begin{bmatrix} Q & 0 & 0 \\ 0 & Q & 0 \\ 0 & 0 & P \end{bmatrix} \begin{bmatrix} A \\ A^2 \\ A^3 \end{bmatrix} \right) x_0 \quad (2.40)
\end{aligned}
$$

By choosing $P = Q = 1$ and $R = 0.01$, this results in the cost function for the undamped mass-spring system as:

$$
\begin{aligned}
V(x_0, u) = & \begin{bmatrix} u_0 & u_1 & u_2 \end{bmatrix} \begin{bmatrix} 0.0806 & -0.0040 & 0.0202 \\ -0.0040 & 0.0602 & -0.0020 \\ 0.0202 & -0.0020 & 0.0400 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \end{bmatrix} \\
& + \begin{bmatrix} u_0 & u_1 & u_2 \end{bmatrix} \begin{bmatrix} 0.6060 & -0.0606 \\ 0 & 0.4060 \\ 0.2020 & -0.0202 \end{bmatrix} x_0 \ (+ \text{ terms not involving } u) \quad (2.41)
\end{aligned}
$$

which is exactly the QP cost function in (2.32) to be solved at each timestep. The optimiser chooses the inputs $u = \begin{bmatrix} u_0 & u_1 & u_2 \end{bmatrix}$ in order to minimise $V(u)$, and then applies u_0 to the plant. Choosing an initial state vector $x_0 = \begin{bmatrix} -0.5 & 0 \end{bmatrix}^T$ (corresponding to the mass having initial position -0.5 and zero velocity), a target mass position at 0, and with all other variables as described above, the closed loop simulation of the system is shown in Figure 2.7. Here, it can be seen that the controller chooses an initial positive force to move the mass upwards followed by a negative force to slow it to rest, bringing the mass to its target position at 0 in around 10 seconds. The performance of this controller is relatively poor – this is due to the fact that short prediction and control horizons and simple gains were chosen to make the algebra simple. In a real application, the horizons would be longer (typically $H_p > H_u$) in order to capture a significant proportion of the system dynamics, and there may be constraints on the force the actuator can supply.

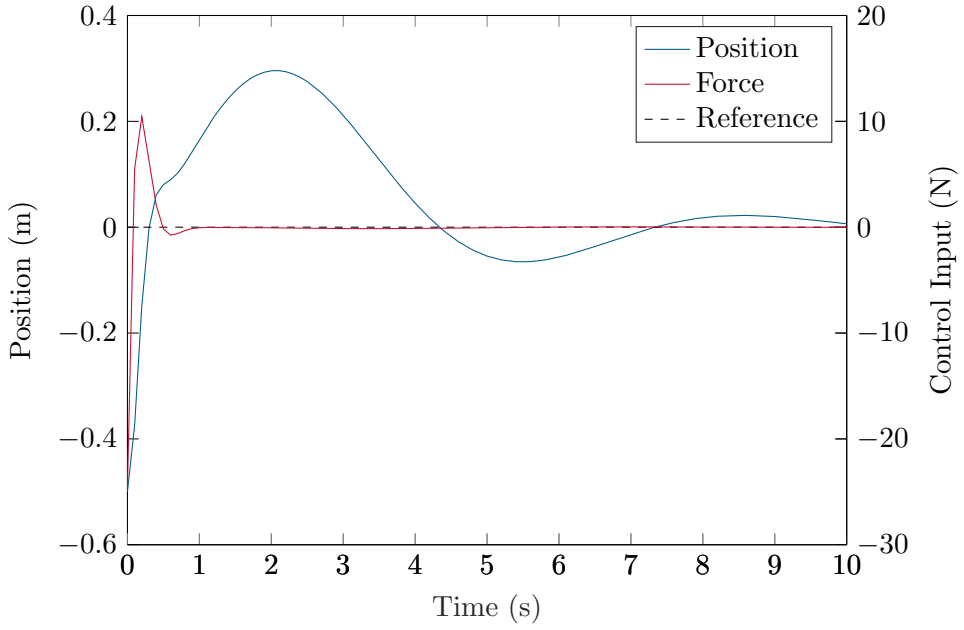


FIGURE 2.7: Closed-loop state and input trajectories for the system in (2.38) from an initial state of $x_0 = [-0.5, 0]^T$.

An interesting and promising concept that has found some success in recent years is that of *explicit* MPC. This was introduced in [63] and describes a technique to *offline* the optimisation inherent in MPC. This work proves that the feedback law for a convex finite horizon problem is piecewise linear and continuous, implying that the online controller action is only to evaluate a piecewise affine function of the state, which is computationally cheap. The disadvantage of this approach is the significant storage required for the control laws. An upper bound on the number of regions of state space required is given by

$$N_r \leq \sum_{k=0}^{2^q-1} k! q^k \tag{2.42}$$

where each region is defined by $q2^q$ inequalities. N_r is the number of regions in the controller and q is the number of constraints defining the region. q increases with N_u, N_c, m and p, which are the input horizon, control horizon, number of inputs (dimension of u in (2.27a)) and number of outputs (dimension of y in (2.27b)) respectively.

2.4.3 Stability and Feasibility in MPC

In general, the difficulties in stability analysis and guarantees in predictive control are analogous to those of adaptive control, in that the control law is being continuously redesigned online. However, there has been, and continues to be, much research in this area. From [61], an equilibrium of a discrete time system $x_{k+1} = f(x_k, u_k)$ is defined as:

Stability The origin is defined as stable if, for any $\epsilon > 0$, there exists a $\delta > 0$ such that if $\left\| \begin{bmatrix} x_0^T, & u_0^T \end{bmatrix} \right\| < \epsilon$ then $\left\| \begin{bmatrix} x_k^T, & u_k^T \end{bmatrix} \right\| < \delta \ \forall \ k$.

where x_k and u_k are the states and inputs at timestep k respectively. The restriction to the origin is without loss of generality of the definition, since it is possible to engineer a change of coordinates such that any equilibrium in the state space is shifted to the origin. Further to this, by requiring that:

$$\left\| \begin{bmatrix} x_k^T, & u_k^T \end{bmatrix}^T \right\| \to 0 \text{ as } k \to \infty, \tag{2.43}$$

the system is *asymptotically stable* [61].

The function $V : S \to \mathbb{R}$, defined on a region $S \subset \mathbb{R}^n$, is said to be a Lyapunov function, if:

C1. $V(0) = 0$

C2. $V(x) > 0 \ \forall x \neq 0 \in S$

C3. $V(f(x,u)) - V(x) \leq 0 \ \forall \ x \in S$ (discrete time) or $\dot{V}(x) \leq 0 \ \forall \ x \in S$ (continuous time).

If there exists such a function V meeting these conditions then the origin is an asymptotically stable equilibrium whose region of attraction is S. Correspondingly if $S = \mathbb{R}^n$ (i.e. encompasses the entire state space), and $V(x) \to \infty$ as $\|x\| \to \infty$ then the system is *globally* asymptotically stable.

It is possible to rewrite (C3.), recalling that a deterministic feedback control law can be written as $u = \mathcal{K}(x)$. If a linear feedback law $u = \mathcal{K}x$ is assumed, then:

$$V\Big(f(x, \mathcal{K}x)\Big) - V(x) < 0 \ \forall \ x \neq 0 \in S \tag{2.44}$$

and all other conditions are satisfied, then V is a *Control Lyapunov function*. If there exists such a function, then the origin is an asymptotically stable equilibrium for the *closed loop* system $x(k+1) = f(x(k), \mathcal{K}x(k))$.

Recalling that $Q \geq 0$ and $R > 0$ in (2.28), the origin is guaranteed to be an asymptotically stable equilibrium under predictive control providing that $P \geq 0$ is chosen to satisfy:

$$(A - B\mathcal{K})^T P (A - B\mathcal{K}) - P \leq -Q - \mathcal{K}^T R \mathcal{K} \tag{2.45}$$

in which $\mathcal{K}$ is any matrix chosen such that $\rho(A - B\mathcal{K}) < 1$, where $\rho(M)$ is known as the *spectral radius* of the matrix M and is defined as:

$$\rho(M) \triangleq \max\{|\lambda_1|, |\lambda_2|, ..., |\lambda_n|\} \tag{2.46}$$

where λ_i are the eigenvalues of M.

In this case, the value function V in (2.28) can be shown to be a Lyapunov function. It is clear that:

1. $V(0) = 0$
2. $V(x) \geq x^T Q x > 0 \; \forall \; x \neq 0$
3. $V(x) \to \infty$ as $\|x\| \to \infty$

and thus the only thing remaining to prove is that:

$$V^\star(x(k+1)) - V^\star(x(k)) < 0 \; \forall \; x \neq 0. \tag{2.47}$$

If a *candidate* input sequence $\widetilde{u}$ can be found which satisfies:

$$V(x_{k+1}, \widetilde{u}) - V^\star(x_k) < 0 \; \forall \; x \neq 0, \tag{2.48}$$

then from the definition of optimality,

$$V^\star(x_{k+1}) \leq V(x_{k+1}, \widetilde{u}), \tag{2.49}$$

and therefore (2.47) is satisfied. The candidate sequence $\widetilde{u}$ is not used for control, but its existence provides proof of stability. To show that such a sequence can always be found, consider a terminal control law

$$\widetilde{u}_N = -\mathcal{K} x_N \tag{2.50}$$

where $\mathcal{K}$ is as chosen in (2.45) and with the same condition that $\rho(A - BK) < 1$. In (2.28), the stage cost is defined as:

$$l(x_k, u_k) := x_k^T Q x_k + u_k^T R u_k \tag{2.51}$$

and the terminal cost:

$$E(x_N) := x_N^T P x_N \tag{2.52}$$

The cost at the next time step under the terminal control law is given by:

$$V^\star(x_k) - l(x_k, u^\star_{k,0} - V_N(x_N^T(x_k)) + l(x^\star_N(x_k), \mathcal{K}x^*_N(x_k)) + V_N((A - B\mathcal{K})x^\star_N(x_k)) \tag{2.53}$$

so that if it can be ensured

$$- V_N(x^\star_N(x_k)) + l(x^\star_N(x_k), \mathcal{K}x^\star_N(x_k)) + V_N((A - B\mathcal{K})x^\star_N(x_k)) < 0 \tag{2.54}$$

then (2.49) will be satisfied. The conditions ensuring this will hold is to choose V_N to be a Control Lyapunov function:

$$V_N(Ax - B\mathcal{K}x) - V_N(x) \leq -l(x, \mathcal{K}x) < 0 \; \forall \, x \neq 0 \tag{2.55}$$

which will always hold as long as $\mathcal{K}$ is chosen according to (2.45). Therefore $V^\star$ is a Control Lyapunov function and stability is guaranteed. For discrete-time or sampled data system, in practice it is difficult to achieve global asymptotic stability due to the effect of sampling. Instead, it is more common to achieve stability in a practical sense. If there exists $\delta > 0$ such that

$$V_N(Ax - B\mathcal{K}x) - V_N(x) \leq -l(x, \mathcal{K}x) + \delta < 0 \; \forall \, x \neq 0 \tag{2.56}$$

then V_N is said to be a *practical* Lyapunov function and the system is called *practically* asymptotically stable.

Feasibility is an issue that is specific to MPC. This is because it arises from the desire to ensure that the online optimisation has a solution at each time step. Since the optimisation is subject to constraints, it may be the case that no feasible solution exists. In the most general case of predictive control, there exist constraints on inputs and states:

$$x_k \in \mathcal{X} \subseteq \mathbb{R}^n \; \forall \, k \tag{2.57}$$

$$u_k \in \mathcal{U} \subseteq \mathbb{R}^m \; \forall \, k \tag{2.58}$$

where n and m are the dimensions of the state and input respectively. $\mathcal{X}$ and $\mathcal{U}$ are the sets of admissible states and inputs respectively. Using the above framework in which a stabilising terminal control law $\widetilde{u}_N = -\mathcal{K}x_N$ is applied at the end of the prediction horizon, consider so-called *dual-mode* MPC, in which MPC is used to drive the system into an invariant terminal set Ω, after which the terminal control law takes over. The idea of *recursive feasibility* introduces a new terminal constraint

$$x(k+N|k) \in \Omega \tag{2.59}$$

which ensures that the MPC control law will always drive the state x into the terminal set Ω by the end of the prediction horizon [64]. This implies that

$$u_{min} \leq -\mathcal{K}x(k+N|k) \leq u_{max} \tag{2.60}$$

$$x_{min} \leq x(k+N|k) \leq x_{max} \tag{2.61}$$

which is to say that the terminal control law is input- and state-admissible within Ω [65]. One further terminal constraint is required to ensure that the next input also leads the state into the terminal set Ω at the end of the prediction horizon, i.e. that

$$x(k+N+1|k) \in \Omega. \tag{2.62}$$

In this case, it remains only to guarantee that Ω is invariant under the terminal control law. Determination of the terminal set Ω and the guarantee of its invariability is out of the scope of this discussion, but it is relatively simple. If this is the case, then the MPC optimisation is guaranteed to be feasible at all future times $k > 0$ as long as it is feasible at time $k = 0$ [65]. A method for detecting initial states for which recursive feasibility cannot be guaranteed is given in [66].

2.4.4 Nonlinear MPC

The fundamental concepts behind predictive control do not stipulate a linear model. However, due to historic preference for the retention of a QP, the term MPC has come,

by popular usage, to mean linear MPC. As a result, modern applications which use nonlinear models have typically been explicitly termed nonlinear MPC (NMPC). In this technique, the linearised internal model is replaced with a nonlinear one, which is especially useful if the plant under control exhibits significant nonlinearities, since in this case the linearised model could, in some regions of the state space, be wildly inaccurate. The cost of this change is that the convexity of the problem is then lost. The naive answer to this is to rely on the optimiser to cope with the solution (although this cannot utilise a QP solver; a more general optimiser must be used). However, due to the time required for solution of the optimisation programme, it is typically necessary to exploit some of the structure in the problem introduced by the MPC formulation. Given a continuous time dynamic system:

$$\dot{x}(t) = f\big(x(t), u(t)\big) \tag{2.63}$$

where $x(t)$ and $u(t)$ are the state and input vectors respectively, t is time, and x_0 is the initial state. Restrict the state and input to some *admissible sets*:

$$\begin{aligned} x(t) &\in \mathcal{X} \subseteq \mathbb{R}^{n_x}, \\ u(t) &\in \mathcal{U} \subseteq \mathbb{R}^{n_u}. \end{aligned} \tag{2.64}$$

Define a cost function to be minimised as the sum of the integral of a stage cost $l(\cdot,\cdot)$ over the finite time horizon t_f and a terminal cost $E(\cdot)$:

$$J(x,u) = \int_{t_0}^{t_f} l(x,u)\,\mathrm{d}t + E\big(x(t_f)\big). \tag{2.65}$$

As in linear MPC, the solution of the infinite time optimal control problem (OCP) is approximated by the successive solution of finite time OCPs:

$$\begin{aligned} \underset{u}{\text{minimise}} \quad & J(x,u) && \text{(2.66a)} \\ \text{subject to} \quad & \dot{x}(t) = f\big(x(t),\, u(t)\big) && \text{(2.66b)} \\ & u(t) \in \mathcal{U} \quad \forall\, t \in [t_0, t_f] && \text{(2.66c)} \\ & x(t) \in \mathcal{X} \quad \forall\, t \in [t_0, t_f] && \text{(2.66d)} \\ & x(0) = x_0 && \text{(2.66e)} \end{aligned}$$

2.4.5 Economic MPC

Traditionally, stage costs have been defined to drive some or all of the states x and inputs u to some reference values x_{ref} and u_{ref}

$$l(x,u) = \|x - x_{ref}\|_Q^2 + \|u - u_{ref}\|_R^2 \tag{2.67}$$

where Q and R are weighting matrices used to prioritise tracking of certain states or inputs. This model fits tracking objectives, where the aim is for the controller to move the plant to a known fixed steady state. However, in many applications, such as process and chemical plants, the best operation condition x_s is not known *a priori* and neither is it fixed. In such applications the best equilibrium is typically chosen to maximise economic benefit and/or minimise cost. Traditionally this is implemented under a framework known as real-time optimisation (RTO), where an upper layer calculates the best operating point, and an underlying MPC layer drives the plant to that condition.

Recently, there has been interest in letting MPC directly determine the best operating condition. Initially, this was termed dynamic real-time optimisation (dRTO), but has come to be labelled as *economic* MPC (EMPC). MPC is labelled economic when its cost function relates directly to the cost of operating the plant, rather than targeting a particular steady state [67]. This was first examined in the context of *unreachable setpoints* in [68], wherein it is noted that such an economic stage cost can yield better performance than traditional *tracking* MPC in cases where the true cost is of economic form.

In EMPC therefore, the stage cost is related to the true economic cost. Tracking MPC typically uses the cost function as a Lyapunov function, as described in Subsection 2.4.3. However, with an economic stage cost, this need no longer needs to be the case. As such, much of the recent literature around economic MPC has been concerned with stability. The key assumption that fails in economic MPC in the general case is that $0 = l(x_s, u_s) < l(x, u)$ where x_s is the optimal steady state and u_s is the corresponding steady input. An important property that appears in much of the economic MPC literature is that of *dissipativity* [69, 70]. If there exists a $\lambda : \mathcal{X} \to \mathbb{R}$ such that for a supply rate $s : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$,

$$\lambda(x_{k+1}) - \lambda(x_k) \leq s(x,u) \tag{2.68}$$

then the system is said to be dissipative with respect to s. λ is known as a *storage function*. In [71], a rotated stage cost

$$L(x_k, u_k) = l(x_k, u_k) + \lambda(x_k) - \lambda(x_{k+1}) - l(x_s, u_s) \tag{2.69}$$

is introduced, where $l(\cdot,\cdot)$ is the economic stage cost. Along with assumptions regarding controllability and strong duality of the steady state problem, a Lyapunov function can be formed using this rotated stage cost instead of the economic stage cost and, subject to terminal conditions, stability of the closed loop can be proved. In terms of performance, [72] shows that on average, performance of EMPC is no worse than performance at the steady state.

Most works in the EMPC literature make some assumptions about terminal conditions in their proofs of closed loop stability and performance. An interesting pair of works, [73] and [74], concern themselves with discarding the terminal conditions and attempting to recover these proofs. The latter makes some particularly useful advances in terms of replacing terminal conditions with a turnpike assumption. Optimal trajectories of a closed loop system are said to exhibit a turnpike if they move towards an equilibrium and stay there for some time before finally departing again [75]. The name turnpike derives a road traffic analogy, where a time-optimal vehicle route for a long journey comprises using a highway (*turnpike*), only leaving at the end of the journey, using local roads to reach the destination. The existence of a turnpike can be proven under certain conditions, most important of which is a dissipativity assumption [69].

2.4.6 Numerical Methods for Optimal Control

The formulation of MPC problems in both the linear and nonlinear cases has been examined, where the resulting optimisation programmes to be solved are quadratic (QP) and nonlinear (NLP) respectively. The solution of the programme in these two cases is now discussed.

Solution of quadratic programmes

In the case of a QP (2.32), due to the constraints imposed on P, Q and R in (2.31), then $\Gamma \succ 0$ in (2.32). This gives the strict convexity property of the optimisation problem that is required for the problem to be a quadratic programme. The reader is referred to [76] for a full treatment of convex optimisation. The effort expended to retain this property pays off when it comes to solving the optimisation to find the optimal input set $u^\star$. The increasing popularity of MPC in recent years has driven a significant body of research on efficient solution of the resulting problem. In the main, modern QP solvers for MPC fall into one of two methods: active set and interior point. Both methods are described here, whilst a comparison of their use for MPC applications is given in [77].

Active set methods begin with a feasible but suboptimal solution. The active set $\mathcal{A}_0$ is the set of constraints which hold with equality at the current solution point. These are

the inequality constraints (2.29c) which are binding. The equality constraints (2.29b) always remain in the active set. For a general phrasing of constraints as in (2.29c) a constraint i is active at x if $g_i(x) = 0$ and inactive if $g_i(x) > 0$. From a starting solution and active set z_k and $\mathcal{A}_k$, the method iterates as shown in Algorithm 2.1.

Algorithm 2.1 Active set methods

1: Solve the equality constrained QP for $\mathcal{A}_k$ and z_k resulting in a new solution z'_k and new set of Lagrange multipliers μ'_k
2: **if** solution is feasible **then**
3: **if** $\mu'_k \geq 0 \; \forall \; k$ **then**
4: Terminate
5: **else**
6: remove the constraint with the most negative Lagrange multiplier μ'_k from $\mathcal{A}_k$ to generate $\mathcal{A}_{k+1}$ and let and go to 1. Otherwise, if all $\mu'_k \geq 0$, terminate
7: $z_k \leftarrow z'_k$
8: Go to Step 1.
9: **end if**
10: **else**
11: perform a line search to move as close as possible to the solution without violating constraints by solving $\max_\alpha H(z_k + \alpha z'_k) \leq h$
12: let $z_{k+1} \leftarrow z_k + \alpha z'_k$
13: Add the new constraint to the active set. Go to Step 1.
14: **end if**

Active set methods have seen significant development, particularly in the direction of tailoring to the programme structures that appear in predictive control applications. Some examples are the *primal-dual* version [78], and others, including [79] and [80]. Examples of its use for predictive control are numerous, including [81] and [82] for linear MPC.

One particularly successful recent variant of this algorithm was developed and implemented in the qpOASES software [83], known as *online* active set. The underlying concept is to improve performance by calculation of the mp-QP for the current state x, which produces the optimal solution $u^\star(x)$ such that under the assumption of no changes in active set, the solution to subsequent MPC problems is trivially computed. When an active set change does occur, the mp-QP solution is recalculated and used until the next active set change. This has proved extremely effective in practice, yielding performance improvements of an order of magnitude or better in real-world applications.

Interior point methods are a class of barrier methods. By this it is implied that for cost and inequality constraint functions

$$\begin{aligned} \underset{x}{\text{minimise}} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0 \; \forall i, \end{aligned} \tag{2.70}$$

the barrier function is formulated as

$$J(x,\tau) = f(x) - \tau \sum_{i=1}^{m} \log\Big(h_i(x)\Big), \tag{2.71}$$

where τ is a parameter. Initially, τ is set to a very large number which means the initial solution is the geometric mean of the constraints, and gradually add in the cost function by letting $\tau \to 0$, as shown in Figure 2.8. The path traced out in these iterations is known as the *central path*. Each step in the path (the *search direction*) is determined by solution of a linear system of equations. A possible termination criterion for IP algorithms is that the cost function is increasing in all feasible directions (i.e., those that would not violate constraints). The use of a logarithmic barrier is ubiquitous in applications of interior point algorithms.

Interior point methods have seen marginally fewer applications in MPC, but are still popular for large problems (many constraints), since they can be faster under these circumstances. Primal-dual methods are again popular, for instance in [84] [85], with some work on parallel implementations of the optimiser [86], including an FPGA implementation [87]. Applications of MPC employing interior point solvers can be seen in [88, 89].

Whilst it is true that these two methods represent the vast majority of optimisers used in MPC applications, there has been interest in other methods, driven in principle by evolving computation platforms. Nesterov's fast gradient method (FGM) was overlooked for MPC for a long time since it is superficially less efficient than either active set or interior point methods. However, its amenability to parallelisation has therefore provoked some interest in recent years, thanks to the falling costs of FPGA and GPU platforms for computation. Examples are given in [90, 91, 92], where the former two works display implementation in configurable hardware.

Solution of nonlinear MPC problems

Solution of the optimisation programme resulting from the nonlinear MPC formulation has seen much attention in the field of optimal control, and techniques are roughly divided into two fields: *direct transcription methods* discretise the problem simply by solving the ordinary differential equations (ODEs) over the prediction horizon, computing the state and input trajectory in this manner for each input sequence. Conversely, *indirect transcription methods* solve the problem by solving the necessarily optimality conditions. These two methodologies are colloquially known as *discretise then optimise* and *optimise then discretise* respectively. Direct transcription methods have been favoured for optimal control applications, due to the difficulty in deriving the necessary

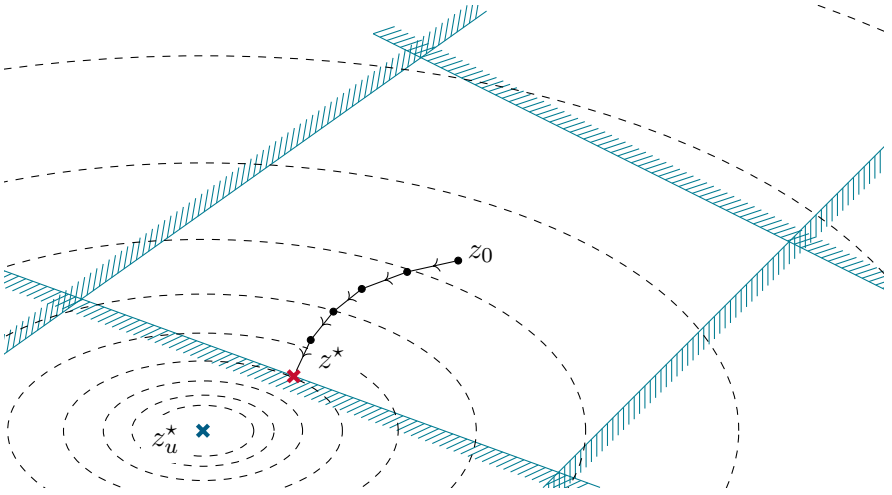


FIGURE 2.8: Interior point methods trace out the *central path* from the starting guess z_0 at the geometric mean of the constraints (hatched) towards the unconstrained optimum $z_u^\star$ (blue cross) of the cost function (dashed), but is constrained to the best feasible solution $z^\star$ (red cross).

conditions for indirect methods [93], so the current discussion is limited to direct methods. A continuous time plant model is assumed, since this is the case for most physically derived models.

Construction of the optimisation programme

Casting the optimal control problem (2.66) into a nonlinear programme (NLP) can be achieved by several methods; those most popular in the MPC field are described here. For a complete description of numerical methods for optimal control, the reader is referred to [93].

Direct single shooting methods begin by discretising the prediction horizon into a mesh of N intervals, and a set of basis functions for the input signal $u(t)$, which are most often chosen to be constant, i.e. $u(t) = q_k$ for $t \in [k, k+1)$, as shown in Figure 2.10. The resulting optimisation variable is

$$z = \Big[q_0, \quad q_1, \quad \ldots, \quad q_N.\Big] \tag{2.72}$$

The system equations (2.63) are then integrated from the measured current plant state x_0 over the prediction horizon from t_0 to t_f, applying the relevant q_i in each interval i, after which the cost function is evaluated and a new set of q_i are chosen to improve the cost. Accordingly, this method belongs to the *sequential* class.

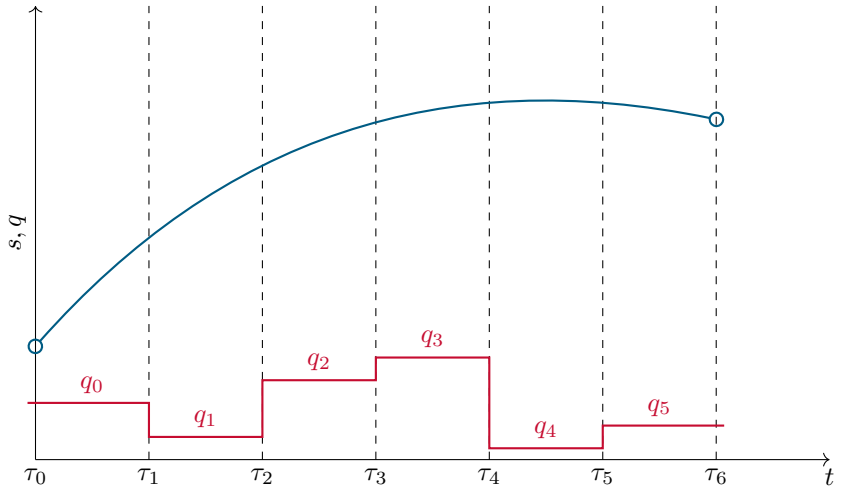


FIGURE 2.9: Direct single shooting integrates the states $s_0 = x_0$ to s_N under the piecewise constant controls q_k.

Direct multiple shooting is in fact preferable to single shooting in most practical applications. The key difference in this case is that the state trajectories are also discretised over the prediction horizon, resulting in a decision variable

$$z = \Big[q_0, \quad s_1, \quad q_1, \quad \ldots, \quad s_N,\Big] \tag{2.73}$$

wherein N new constraints are added to the problem to ensure consistency of the state trajectories, i.e. $s_{i+1} = \int_{t_i}^{t_{i+1}} f(s_i, q_i)\,\mathrm{d}t$. These are typically known as defect constraints. Since the algorithm solves for consistency of state trajectories and optimality of inputs at the same time, this is a *simultaneous* method. The main advantage of direct multiple shooting is that, despite the increased problem size in terms of decision variables and constraints, the integrators run over single intervals of the mesh grid only, meaning that stiff and unstable solution trajectories are numerically better conditioned. Intuitively, this can be viewed as distributing the the nonlinearity equally over the nodes [94]. An additional advantage is the potential for parallelisation of integration in these intervals.

Direct collocation methods are another simultaneous method, but instead of integration of the system equations (2.63), an approximation is made by use of a polynomial function which satisfies the solution of a two-point boundary value problem (TP-BVP) defined by the state trajectories between two points in time (*collocation points*) and their derivatives at those points (if these points coincide with the grid mesh points the method is termed a *Lobatto* method). An example is the trapezoidal method in which the state at the end of the interval is described by the following TP-BVP:

$$x_{k+1} = x_k + \frac{h}{2}\Big[f(x_k, u_k) + f(x_{k+1}, u_k)\Big] \tag{2.74}$$

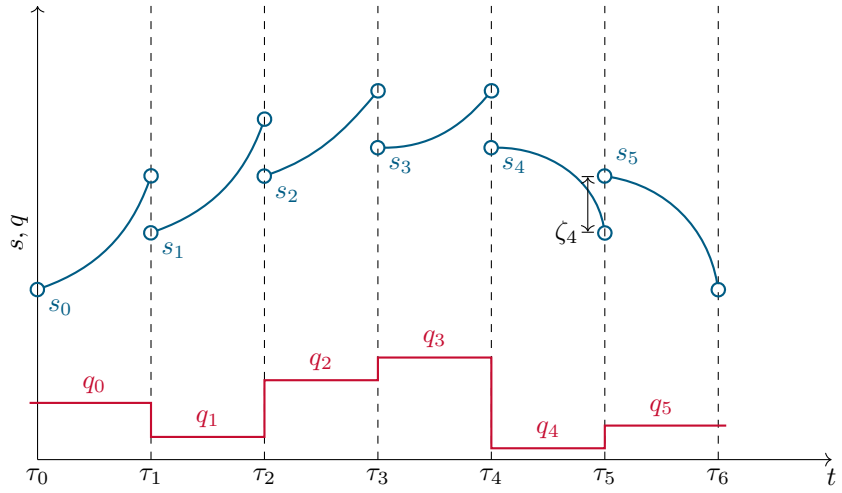


FIGURE 2.10: Direct multiple shooting integrates the states s_k to s_{k+1} under the piecewise constant controls q_k. Constraints enforce $s_0 = x_0$ and the defects $\zeta_k = 0$ to ensure consistency of state trajectories.

where h is the integration step size and is defined, assuming the initial and final times are fixed, as $h = \Delta\tau = \tau_i - \tau_{i-1}$. u_k appears in both f terms due to the assumption that the input is fixed during the interval. The defect constraints exist to ensure the consistency of the state trajectories between points on the mesh grid, as shown in Figure 2.10. From (2.74), the defects are:

$$\zeta_k = x_{k+1} - x_k - \frac{h}{2}\Big[f(x_k, u_k) + f(x_{k+1}, u_k)\Big] \tag{2.75}$$

and the constraints $\zeta_k = 0$, $\forall k$ are added as equality constraints to the NLP. Trapezoidal collocation methods, a first order scheme, appear often in the literature, as do higher order polynomial schemes. The repeated solution of the transcribed OCP is now required:

$$\underset{z}{\text{minimise}} \quad J(z) := \sum_{k=0}^{N-1} l(s_k, q_k) + E(s_N) \tag{2.76a}$$

$$\text{subject to} \quad s_0 = x_0 \qquad \text{(initial condition)} \tag{2.76b}$$

$$\zeta_k = s_{k+1} - F(s_k, q_k) = 0 \qquad \text{(defect constraints)} \tag{2.76c}$$

$$g(z) = 0 \qquad \text{(equality conditions)} \tag{2.76d}$$

$$h(z) \leq 0 \qquad \text{(inequality conditions)} \tag{2.76e}$$

where F is the solution to the initial value problem (IVP) associated with the system of ODEs (2.63) at the following time step $k + 1$:

$$F(x_k, u_k) = x_{k+1} = \int_{t_0}^{t_0+H} f(x(t), u(t))\,\mathrm{d}t, \tag{2.77a}$$

$$x(t_0) = x_k. \tag{2.77b}$$

The optimal input is given by:

$$u_0^\star = \begin{bmatrix} \mathbb{I}_{n_u}, & 0, & 0, & \ldots, & 0 \end{bmatrix} z^\star \tag{2.78}$$

where $\mathbb{I}_n$ represents an $n \times n$ identity matrix. Note that equality constraints can be added to the NLP with the given formulation. It is important to note that the transcription process guarantees only pointwise-in-time constraint satisfaction, i.e. at the mesh nodes, rather than the continuous constraint satisfaction that the original continuous time OCP (2.66) demands.

Solution of nonlinear programmes

A (highly structured, in the case of NMPC) NLP of the following form is to be solved:

$$\begin{aligned} \underset{z}{\text{minimise}} \quad & f(z) && \text{(2.79a)} \\ \text{subject to} \quad & g_i(z) = 0 \ \forall i && \text{(2.79b)} \\ & h_j(z) \leq 0 \ \forall j && \text{(2.79c)} \end{aligned}$$

Any solution of this optimisation programme must satisfy the Karush-Kuhn-Tucker (KKT) conditions. Defining the Lagrangian of the problem (2.79) as

$$\mathcal{L}(z, \lambda, \mu) := f(z) - \lambda^T g(z) - \mu^T h(z) \tag{2.80}$$

there exist Lagrange multiplier vectors $\lambda^\star$ and $\mu^\star$ (with $\star$ denoting values at the optimum) such that the KKT conditions hold [94]:

$$\begin{aligned} \nabla_z \mathcal{L}(z^\star, \lambda^\star, \mu^\star) = 0 \qquad & \text{(stationarity)} && \text{(2.81a)} \\ g_i(z^\star) = 0 \ \forall i \qquad & \text{(primal feasibility)} && \text{(2.81b)} \\ \mu_j^\star \geq 0 \ \forall j \qquad & \text{(dual feasibility)} && \text{(2.81c)} \\ h_j(z^\star)\mu_j^\star = 0 \ \forall j \qquad & \text{(complementary slackness)} && \text{(2.81d)} \end{aligned}$$

All Newton-type optimisation algorithms attempt to find solutions satisfying these KKT conditions. The major difference is due to the treatment of the complementary slackness condition (2.81d) [94], which comes from the inequality constraints[1]. In particular, interior point algorithms treat this condition by replacing (2.81d) with $h_j(z^\star)\mu_j^\star = \tau \; \forall j$ and let $\tau \to 0$.

Sequential quadratic programming (SQP) is one of the most often used methods in the NMPC literature for solving the nonlinear programme (2.79). This is perhaps because of the significant body of research that has gone into fast solution of a QP for MPC purposes. The method works by repeatedly linearising all equations in (2.81) and noting that the result is the KKT conditions for a QP, where its solution is a search direction in x. QPs are solved successively until convergence to the optimum is obtained. The SQP method iterates according to

$$z_{k+1} = z_k + \beta_k \Delta z_k \tag{2.82}$$

where Δz_k is the solution to the quadratic programme

$$\begin{aligned}
\underset{\Delta z_k \in Z_k}{\text{minimise}} \quad & \nabla F(z_k)\Delta z_k + \frac{1}{2}\Delta z_k^T B_k \Delta z_k && \text{(2.83a)}\\
\text{subject to} \quad & G(z_k) + \nabla G(z_k)^T \Delta z_k = 0 && \text{(2.83b)}\\
& H(z_k) + \nabla H(z_k)^T \Delta z_k \leq 0 && \text{(2.83c)}
\end{aligned}$$

where ∇F, ∇G and ∇H are the Jacobians of the cost, equality and inequality constraint functions respectively. B_k is the Hessian of the Lagrangian (2.80), where $B_k := \nabla_z^2 \mathcal{L}(z, \lambda, \mu)$. The Hessian B_k is usually not evaluated exactly (though this is possible for some systems where an analytical Hessian can be stored and evaluated online). More often, the famous Broyden-Fletcher-Goldfarb-Shanno (BFGS) update rule is used. In the special case of a sum-of-squares cost function, the Gauss-Newton Hessian approximation can be used instead. See [94] for descriptions of these two methods. Where the Hessian is not computed exactly, the method is termed a *quasi*-Newton method.

Generation of the Jacobians can be implemented in one of several ways:

- Analytical: For small differentiable systems it is possible to store the analytical solutions to the Jacobian and evaluate them as necessary.

[1]This condition is known as complementary slackness in the literature surrounding the theory of optimisation, particularly that regarding duality. Intuitively, it can be seen as a requirement that if an inequality constraint is binding, then its Lagrange multiplier must be positive. A non-binding (inactive) equality constraint must have a Lagrange multiplier equal to zero.

- Numerical: The Jacobians can be computed by a finite difference computation. Relies on some continuity assumptions of the cost and constraint functions.
- Sensitivity integration: In general, the solutions of ODEs can depend on parameters and initial conditions. Some modern schemes generate the numerical Jacobians by integration of the so-called sensitivity equations alongside integration of the system equations themselves. There is typically some redundancy that can be exploited to make this process faster than the separate evaluation of the two sets of equations [95].

One approximate NMPC approach based on SQP is real-time iterations (RTI) [96], in which a single SQP step is performed at each plant sample time (the controls are therefore suboptimal but approach optimality during runtime). Optimisation of the controls occurs alongside the run time of the plant. Of particular note in this algorithm is the division of the SQP step into *preparation* and *feedback* phases, with the preparation step being significantly longer but performed in the period between plant sampling instants. The feedback phase is relatively fast and thus allows minimal time between plant state measurement and controls being applied.

Numerical integration of ODEs is typically required in the implementation of NMPC methods. In the direct single and multiple shooting frameworks, integrators are required to evaluate the state consistency conditions (2.76d), since this involves solution of the system equations (2.63). There has recently been interest in implicit integration methods, since they are known to have a larger region of stability than explicit methods and and thus better suited to stiff and unstable systems [97]. In particular for NMPC applications, there have been recent advances in integrators with inbuilt sensitivity generation; [98] is one example based on implicit Runge-Kutta (IRK) methods.

In terms of existing software packages for solution of MPC problems, there are several tailored towards QPs with structures that appear in MPC problems (2.32). Two very recent active set methods are online active set, implemented in the qpOASES package [83], and dual Newton strategy implemented by the qpDUNES package [99]. The former of these exploits ideas from explicit MPC where the QP is solved multiparametrically, generating solutions almost instantaneously when the active set does not change from one time step to the next. The dual Newton strategy attempts to combine the structure exploitation of interior point methods with the warm-starting capabilities of active set methods. One of the most widely used interior point based solvers is FORCES [100], which is based on a primal-dual interior point algorithm[2]. For NLPs, IPOPT [101] is one of the most well known, using a primal-dual interior point algorithm to directly

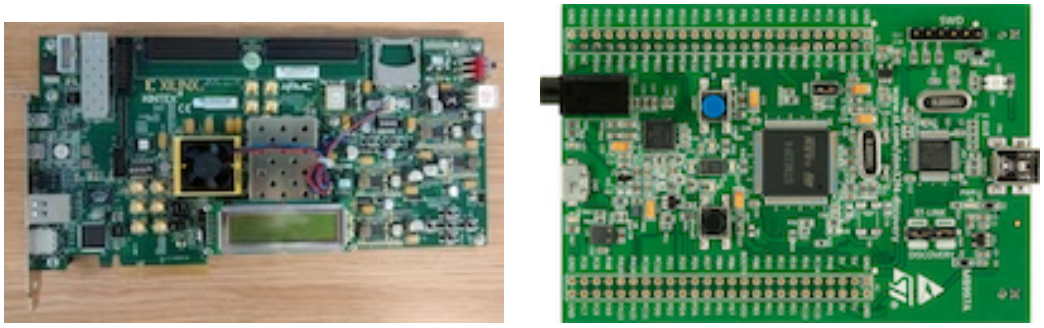
[2]Primal-dual refers to the algorithm simultaneously searching for solutions to both the primal and dual problems, in order to improve convergence speed. This formulation also allows to use the value of the duality gap as an indication of convergence or as a termination condition.

solve the NLP. The solvers mentioned are supplied with both C and MATLAB interfaces, permitting direct used for simulation on a desktop computer as well as compilation for other platforms. Whilst not directly deployable to an FPGA, the code structure could be ported to these devices to ease the work of designing an FPGA QP solver from scratch. The ACADO [102] and ICLOCS [103] packages represent the forefront of frameworks for solution of NMPC problems. They both permit the expression of (possibly nonlinear) system dynamics in continuous time as well as general objective cost and constraint functions. They carry out the transcription process described in Subsection 2.4.6, with ICLOCS recommending IPOPT for solution of the resulting NLP, whilst ACADO uses an SQP strategy and is bundled with qpOASES for solution of the underlying QPs. It is worth mentioning that both support direct multiple shooting, with ICLOCS relying on the SUNDIALS toolbox (specifically, `cvodes`) to achieve the numerical integration with forward sensitivity generation, whilst ACADO does this internally via an implicit Runge-Kutta (IRK) scheme [98].

2.4.7 Hardware platforms for MPC

The solution of the finite horizon OCP (2.76) is computationally complex and is the fundamental obstacle to implementation of MPC controllers on real-time platforms. Two parts of the algorithm are typically responsible for the computational complexity. Firstly, the optimisation at the heart of MPC is an operation which is inherently complex, but also has no guaranteed runtime before the solution is found. This causes issues in online applications since there is typically limited computational resource available to solve the problem in a timeframe which is already uncertain. Secondly, if integration of the system dynamics is required, this is also computationally expensive (particularly if sensitivity equations are also to be solved).

With respect to the hardware on which the controller will be based, there are again two competing options. The traditional embedded microprocessor has seen enormous leaps in performance and decreases in cost in recent years. The microprocessors typically found in vehicle aftertreatment control systems are not comparable to the central processing units (CPUs) found in desktop hardware. They typically have significantly lower clock rates and, perhaps more importantly, far less fast random access memory (RAM). Desktop CPU architectures include several mechanisms for speeding up computation (e.g. *caches*), knowledge of which can be exploited to maximise the amount of computation time available on the CPU. However, typical microprocessor systems do not have such features. One the most promising devices in this area is the ARM CORTEX M4F unit, a development board for which is shown in Figure 2.11b. This unit contains a digital signal processing (DSP) core which supports highly vectorised computation. However, there are currently very few examples of MPC running natively on a microprocessor. One of the few is shown in [104], which uses an ARM Cortex-M microcontroller, albeit a



(a) Xilinx Kintex KC705. (b) ARM Cortex M4.

FIGURE 2.11: FPGA and microprocessor development boards.

fairly simple plant model. It should be noted however, that the clock speed of this unit is significantly lower than can be attained nowadays, and also lacks the aforementioned DSP core. Explicit MPC, thanks to its lack of online optimisation, is well suited to microprocessor implementation – this is discussed in [105], with one example given by [106].

The alternative platform is the field programmable gate array (FPGA), an example of a development board for which is shown in Figure 2.11a. These devices consist of configurable blocks, which can implement functions from memory to complex combinatorial logic, providing rapid prototyping functionality for application specific integrated circuits (ASICs). One of the principal attractions of FPGAs is that they offer the potential for parallelisation of an algorithm (if the structure of the algorithm permits it). An example of exploiting this to develop a QP solver for a constrained linear MPC control application using the interior point method is given in [107]. It makes direct use of a technique known as multiplexed MPC (MMPC), in which the MPC problem is solved for several subsystems independently and sequentially [108]. For a complex system, exploitation of opportunities for parallelism in the MPC algorithm, and their implementation on a parallel computation platform, may provide the key to satisfactory performance of the controller. It is noted in [109] that the opportunity for parallelisation is entirely dependent on the algorithm, since data dependencies can preclude it, implying that the formulation of the problem must be completed with parallelism in mind. Examples of MPC on FPGAs are rather more numerous [90, 58, 110], but it is noteworthy that it is typically the optimiser which is the subject of the authors' effort to speed up the algorithm. A recent development promising to ease the rapid prototyping of MPC on FPGA hardware is given in [111], which solves a condensed QP using a fast gradient method. A stochastic approach using particle swarm optimisation, which is highly amenable to parallelisation, is examined in [112].

2.5 Automotive Applications of Predictive Control

Examples of the application of model predictive control in automotive environments have become more and more frequent in the past few years. Whilst mostly in an academic setting, the focus of making these controllers suitable for industrial and consumer applications has begun to increase. Naturally, the areas in which effort has thus far been expended have been those systems whose characteristics are known to suit a predictive control framework, coupled with those in which classical control systems cannot attain the desired performance. One particular example of this is in active suspension control: the ability of MPC to respect actuator constraints in addition to incorporating feedforward information from sensors (in this case, road preview information) has been an important factor [59, 113], through these works involve simulation only rather than real-time application. Additionally, the active suspension has only one actuator, sampling times are relatively forgiving, and the system is not safety critical.

2.5.1 Powertrain Controls

As computational power continues to increase, research is moving into using predictive control to regulate operation of the internal combustion engine itself. These typically require faster sampling times, and more importantly, are more safety critical. Positioning of an EGR valve has been proven with MPC in [114], which uses a 3-state, 1-input model of the DC actuator motor, and demonstrates improved valve positioning accuracy. More advanced applications in this area involve management of the engine air path in general, in which MPC appears to demonstrate improved control over PID due to its natural expansion to multiple-input multiple-output (MIMO) systems. Applications have included control of an inlet throttle and EGR [115], EGR and variable geometry turbocharger (VGT) [116, 117] (the former of which applies the C/GMRES MPC scheme), and dual-loop EGR, VGT and exhaust throttle (ET) [118]. For spark-ignition engines, 6.5% fuel savings in simulation have been demonstrated by the application of MPC to indicated mean effective pressure (IMEP) tracking whilst minimising fuel consumption [119]. A two stage turbocharged gasoline air path control application has been demonstrated to be real-time capable by using the SQP NMPC approach [120].

There has been consideration of directly optimising for emissions, which has been termed integrated emissions management (IEM); a paradigm shift in the design of automotive control systems. The idea is to govern the engine, powertrain and aftertreatment with one overarching supervisory controller, which determines operating points in order to optimise a particular cost function. Underneath this trajectory-planning controller are low level controllers which are designed to regulate actuators, and estimators or filters to deal with sensors. The design of the cost function is the key parameter in the construction of IEM control systems. In fact, the cost function is typically varied online such

that power delivery and emissions are acceptable on average. A simulation approach showing the benefits of IEM is presented in [121], which demonstrates a 2% fuel saving over the baseline strategy. This is complemented by a second paper [122] which applies the derived controller to a real HDD engine and shows that the fuel efficiency gains can be realised in practice.

More recently, control of hybrid electric vehicles (HEVs) has been a promising area for the application of MPC. The application of predictive control to the power split in a HEV has been investigated and proved to work well [123, 124]. Conversely, several works examine state-of-charge and battery lifetime optimisation of an electric vehicle [125, 126, 127]. Economic MPC approaches have also been considered in this vein, and those which consider other hybrid power sources, such a fuel cells and supercapacitors [128]. Supervisory MPC has also been explored for drive cycle optimisation, with [129] demonstrating a given state of charge at the end of a journey, and [130] examining how MPC can be added to control retrofitted HEV equipment in a supervisory role. The latter makes useful comparisons to rule-based industry standard approaches. However, all of these works are simulation only and do not attempt to solve the MPC problems in real time.

2.5.2 Aftertreatment Controls

The competence of predictive control in the management of complex chemical processes has been well documented, not least due to the fact that the development of the technique began in this field. Nevertheless, there have been relatively few attempts in published literature to apply predictive control to the SCR process. One such approach was made by [131], in which the authors initially build a full chemical and thermal model of the SCR system, and then derive a reduced order parameterised linear model for use in the predictive controller optimisation. The parameters are calculated online as the system moves through various states, and are used to construct the linearised model for the current timestep. The authors compare this technique to the simple successive linearisation technique discussed in Section 2.4, concluding that the parameterised model technique is computationally less expensive and is therefore preferred. The timestep used in this work is 1000ms, of which the controller requires 31ms on a desktop (and estimated 115ms on an embedded microprocessor system). An interesting notation is introduced in that they use the NO_x conversion *inefficiency* rather than the efficiency as the controlled output and the primary performance metric – it is stated that this is more appropriate since emissions regulations are stated in terms of the amount of NO_x released to the environment rather than the amount converted by the aftertreatment system. Finally, it is noted that the combination of their derived predictive controller along with the inclusion of a feedforward controller (as is often currently used in vehicles)

could improve performance still further. This is something to be examined in the course of this research project.

A linearised SCR model is also used by [132], again derived from a nonlinear model, and is given in (10) in this paper. The engine-out NO_x model used is a lookup table with a lag to account for thermal effects. The implementation of the controller is trivial since a canonical linear model of the system with no further augmentation is available. However, the results prove worthwhile, with average NO_x conversion efficiency above 93% and ammonia slip lower than 10ppm. This work does not compare the designed controller against the original nonlinear SCR model, so the performance sacrifice due to the use of the linear model is unknown. There is also no data on the computational effort required for this implementation, although it could be estimated by benchmarking a QP solver with a similar cost function, number of constraints and states.

Conversely, more recent attempts have used a nonlinear plant model, perhaps under the assumption that computational power and advances in numerical optimal control are bringing real-time NMPC closer to reality [132, 133, 134]. The results appear to demonstrate very high conversion efficiencies: up to 96.6% in one case, though it should be noted that these are simulation results only. However, these works typically do not consider the real-time implementation of the NMPC scheme, but instead focus on model-in-the-loop implementation, and no details of the computational demand or solution time are given. The details of the solution of the NMPC problem and the route to practical implementation in a vehicle are left to future researchers.

A quite complete work on the model-based optimisation of an entire diesel engine is given in [32], which includes the EGR rate as well as the urea dosing rate in the control problem. This work makes no attempt to simplify the SCR model, using a physically inspired nonlinear model. As a result, each optimiser run takes 20-60 seconds with a 1 second timestep, which precludes online implementation. Some investigation as to why this is the case is required; it is likely to be a combination of the nonlinear problem formulation and the optimiser in use. With careful tailoring of the algorithm and implementation on fast enough hardware, it may be possible to run the full nonlinear model developed in this paper fast enough for online operation. However, the required speed increase of nearly two orders of magnitude is likely to require that some compromises be made. It is noted in this paper that the use of the predictive controller in combination with lookahead information, such as road speed and junction located (e.g. from a satellite navigation system) could lead to further performance improvements. The author notes that optimisation of cold start emissions is not examined, and this should be included in future work for vehicle applications, since incorrect control of the SCR in these conditions could cause unwanted ammonia slip.

Adaptive MPC schemes have also appeared in the literature with application to SCR, particularly because urea doser calibration is one of the major causes of plant-model

mismatch in model-based control schemes for SCR [28]. This is implemented by including a term in one of the linearised model parameters which is updated with a gradient based parameter estimation law. The results indicate that the inclusion of this adapted parameter is a significant improvement over the baseline MPC controller. Subsequently, this has been attempted using NMPC, also including the doser calibration as an adaption parameter [135]. The adaptive element is included to counteract ageing effects in both the catalyst and the urea doser unit, and it is shown that such online adaptation can improve the performance of the control system significantly, due to the profound effects that such mismatch can have on the closed loop performance.

Thermal management of the exhaust and aftertreatment system has begun to appear in the literature, sometimes coupled with SCR control [136]. There have been approaches to air path control that are optimised for catalyst heating [137]. This has also been investigated with regards to HEVs with aftertreatment with the aim of optimising for minimum catalyst warm up period [138].

Chapter 3

Emissions Data Collection

This chapter details a project that was undertaken with the Department for Transport (DfT) and Southampton City Council, in which the council has been given funding by the DfT to report on NO_x emissions in the city, which are presumed to be at intolerably high levels. Transport for London (TfL) has undertaken research to understand the causes of urban NO_x pollution, and has found that buses are responsible for a significant fraction [139]. Whilst Euro VI buses are significantly better in this regard [140], the age and composition of bus fleets typically mean that replacing the entire fleet is infeasible [141]. Southampton exceeded the annual NO_2 limit in 2013 and is predicted to remain non-compliant until 2020 under the current strategy [142]. The work reported here is in collaboration with First Group. Euro V buses underwent emissions testing under real world driving conditions, in order to quantify the efficiency of the factory-fitted SCR system, and to explore the potential of improving this via the retrofitting of additional technology. A portable emissions measurement system (PEMS) unit as shown in Figure 3.2a was used to sample the tailpipe emissions whilst the vehicles were loaded with ballast and driven over typical routes through the city at peak times. The PEMS equipment measured tailpipe NO_x, CO, CO_2 and O_2 as well as exhaust flow rate and temperature, via a tailpipe extension as shown in Figure 3.2b. Separate sensors were added to the vehicle to measure tailpipe ammonia (since the PEMS unit cannot measure this). Additionally, a custom telematics unit was constructed to measure key parameters from the engine diagnostics bus. Constraints on the packaging of the measurement systems meant that measurement of upstream NO_x, i.e., before the SCR, was not possible; however, it is shown that this can be adequately estimated by combining information from the PEMS and telematics units.

The data collected in this chapter provides information on important aspects of modelling an SCR catalyst. In particular, it demonstrates the requirement for thermal management in order to have SCR catalysts operate optimally. It also indicates a desire to bring control of thermal management systems under the umbrella of the aftertreatment controller, in the vein of the IEM approaches discussed in Section 2.5. Allowing the

(a) Wrightbus Streetlite DF

(b) Engine bay, catalyst bottom left

FIGURE 3.1: The vehicles whose real-world emissions were measured in this project.

aftertreatment system to manage the thermal management strategy will allow a combined control objective of emissions reduction, rather than unsynchronised controllers with differing, and often conflicting, priorities.

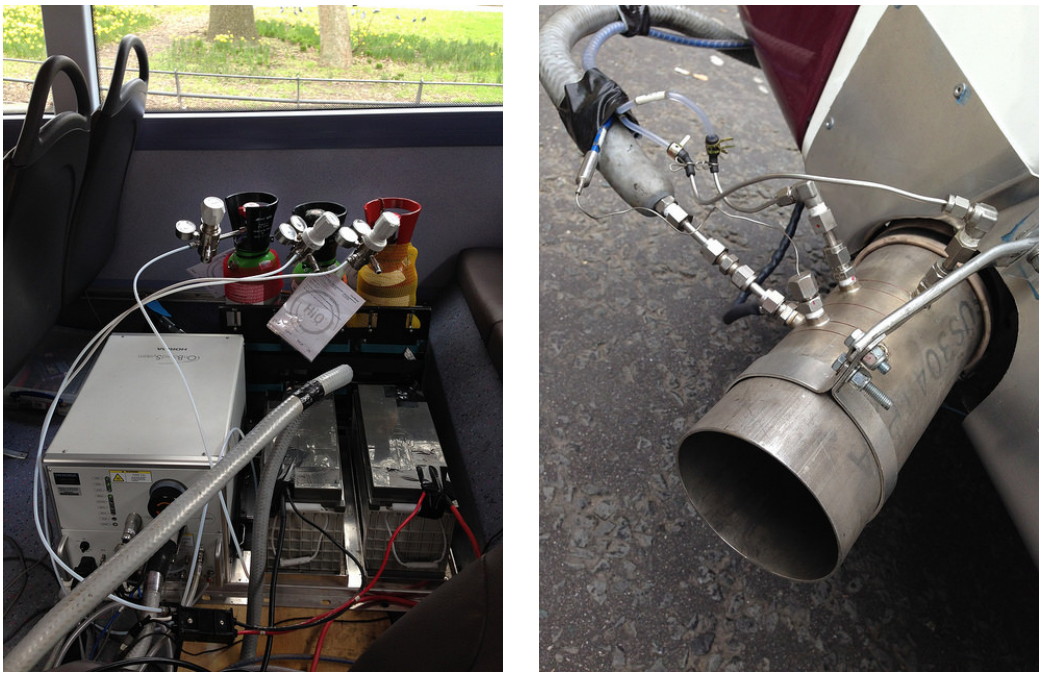
3.1 Instrumentation

Instrumentation of the vehicles was considered in view of the dual aims of examining off-cycle emissions, and understanding the conditions which lead to suboptimal performance and thus potential mitigation routes. The final configuration comprised two principal elements: direct recording of exhaust emission concentrations of key species, and monitoring of engine control system and vehicle diagnostic parameters. WrightBus StreetLite Door Forward (DF) vehicles employ SCR in order to meet their Euro V NO_x emissions limits (shown in Figure 3.1a), so two of these were selected for testing. This system is factory-fitted. The compact engine bay layout and ease of access, as shown in Figure 3.1b, permitted simple installation and inspection of the testing equipment.

3.1.1 Exhaust Emissions Measurement

The vehicles were fitted with the Horiba OBS-2200 portable emissions measurement system (PEMS). The PEMS unit samples tailpipe gas immediately before its release to the environment, and is capable of recording key emissions including CO, CO_2, H_2O and NO_x at a sample rate of 10 Hz. Notably, the vehicles were not fitted with NO_x sensors before the SCR, implying that pre-SCR NO_x levels were estimated by the aftertreatment control system based on other measured parameters. This was a key factor in the decision to additionally monitor engine parameters, as discussed below.

The in-cabin equipment, comprising the OBS-2200, batteries and gas bottles, was fitted in place of the rearmost pair of passenger seats. A custom baseplate was used, as shown



(a) OBS-2200, batteries and gas bottles (b) Tailpipe attachment

FIGURE 3.2: Installation of the Horiba OBS-2200 PEMS equipment on the vehicles to be tested.

in Figure 3.2a, to retain the PEMS equipment, gas bottles and batteries. The OBS-2200 collects samples of the exhaust gas via a tailpipe attachment shown in Figure 3.2b. Additionally, this attachment carries a pair of pitot tubes for flow rate measurement, and a thermocouple to monitor exhaust gas temperature. These data are all fed into the main OBS-2200 unit, which is in turn connected to a laptop computer via Ethernet for setup, calibration and data logging.

The PEMS uses a heated chemiluminescent detector (HCLD) to measure NO_x, as well as having a heated sample pipe to bring a sample of the exhaust gases to the main unit (around 5 metres) in order to reduce the likelihood of condensation of water vapour in the exhaust sample, which could damage the analysers. Consequently, the OBS-2200 requires a 24V power supply capable of up to 40 Amperes. It was considered impractical to draw power from electrical systems on the bus, since the significant electrical load would lead to increased fuel consumption and therefore impact emissions. Correcting for this influence would be extremely difficult and would thus lead to error in the data. Instead, pairs of 12 Volt 120 Ah sealed lead acid (SLA) batteries were used, with each pair permitting around three hours of run time. These can be seen in the bottom right of Figure 3.2a. In the rear of this image are three gas canisters and regulators that the OBS-2200 requires: a calibration gas (*span gas*) with known proportions of detectable substances, synthetic air and hydrogen (H_2).

The exhaust sample pipe, twin pitot tubes and exhaust thermocouple were routed from the exhaust tailpipe attachment into the cabin. This was achieved by removing an access

panel in the cabin floor which is usually used for engine maintenance. To prevent ingress of engine and road noise, as well as exhaust fumes from the engine bay, a custom panel was manufactured, as shown in Figure 3.3.

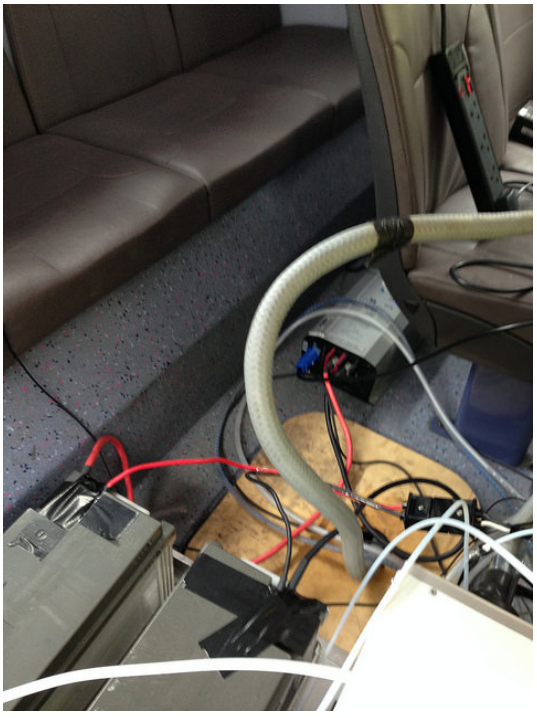


FIGURE 3.3: The in-cabin engine access panel was replaced with a custom panel allowing connections to the tailpipe attachment to be routed into the cabin whilst preventing engine noise and exhaust fumes from entering.

3.1.2 Engine Telematics

Engine operating conditions are used in the onboard engine-out NO_x estimator and the SCR control system. As such, a telematics unit was built to request and record relevant parameters over the OBD-II diagnostics port in the vehicle. These data are listed in Table 3.1. Two desirable parameters for characterising the catalyst would be engine output NO_x (estimated, in the case of these vehicles) and urea injected before the SCR

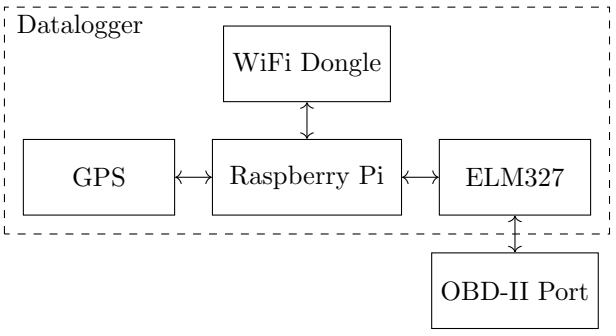


FIGURE 3.4: Schematic of the OBD-II datalogger unit.

catalyst by the aftertreatment control system. Unfortunately, due to the closed-source nature of the system, it was not possible to retrieve these parameters over the diagnostics bus.

The unit was based around a Raspberry Pi single-board computer, as shown schematically in Figure 3.4 along with the complete unit shown in Figure 3.5. The vehicles use the SAE J1939 standard for on-board diagnostics, which is one of the modes supported by the ELM Electronics 327 CAN/OBD device, which are cheaply available with USB-Serial connectivity, such that commands can be sent and responses easily parsed by the Raspberry Pi software. A global positioning system (GPS) receiver was also included in the setup, both for real-time clock (RTC) purposes and to record vehicle latitude, longitude and altitude on the route. These positional data were necessary in order to map emissions to key points on the route, such as inclines and bus stops. A wireless adapter was included such that the datalogger was able to upload the stored data when connectivity was available, specifically at the bus depot where the vehicle was deposited overnight.

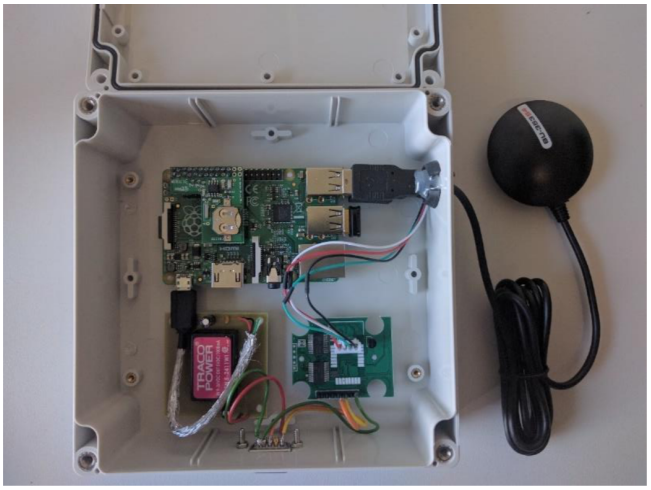


FIGURE 3.5: Telematics unit internal view, showing the Raspberry Pi (internal, top), ELM327 (internal, bottom right), power supply (internal, bottom left) and GPS (external, right).

The software written for the telematics unit simply requested a pre-defined list of parameters, as shown in Table 3.1, from the vehicle on a regular basis after ensuring the engine was running. Data were compressed and stored locally until such time as internet connectivity was available, whereafter a separate process managed transmission of the data to a server. The server parsed and inserted the data into the relevant table of a MySQL database, along with metadata such as the Vehicle Identification Number (VIN) which uniquely identifies each vehicle.

Table 3.1: Parameters logged by the telematics unit and their units.

Parameter	Units
Accelerator pedal position	%
Engine torque	%
Maximum available torque	%
SCR inlet temperature	°C
SCR outlet temperature	°C
Brake switch	Integer
Clutch switch	Integer
Coolant temperature	°C
Demand torque	%
Fuel flow rate	l/h
Intake mass air flow rate	kg/h
Intake manifold temperature	°C
Engine load	%
Engine speed	rpm
Turbocharger intake pressure	kPa
Brake power	kW
Transmission gear current value	Integer

3.2 Test Procedure

Since the aim was to measure real-world emissions, each test comprised the vehicle running its usual route during the normal times of day. The vehicle was loaded with 700kg of sacks of sand, which along with the test team and measurement equipment, simulated a bus partially loaded with passengers. Each test comprised two hours of running the normal route, as driven by a PSV-licensed operator appointed driver, and stopping at all usual bus stops. Risk and safety assessments were conducted prior to the start of the experiments, and before the equipment was set up on board the bus (at the depot) and at the beginning of each trial, a safety briefing was given to all those involved in the tests.

Each of the two vehicles underwent a total of four tests, split over two days. On each day, the vehicle left the bus depot at 6.30am and ran for two hours, which encompasses the city rush hour. The two hour duration was chosen since it allowed at least two complete circuits of the route. The vehicle then returned to the depot for half an hour, to allow a change of driver and exchange of PEMS batteries, before setting out for the second two hour test at 9am. Vehicle 1 was tested in February, resulting in a mean ambient temperature over the test of 8.7°C, whereas Vehicle 2 was tested in May, with a mean ambient temperature of 16.1°C. The weather conditions were otherwise similar, with little or no rain and negligible wind.

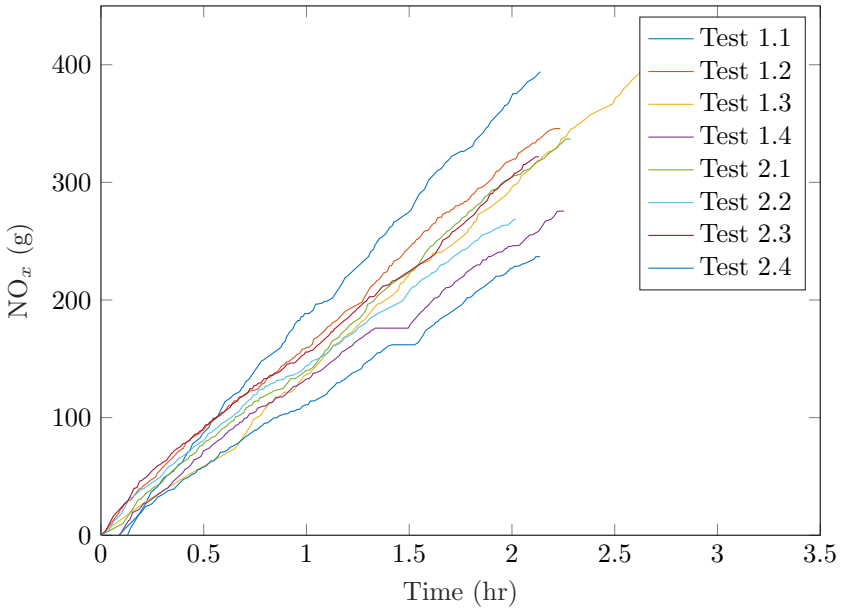


FIGURE 3.6: Cumulative NO_x output during each of the 8 tests, where Test $x.y$ refers to Vehicle x, Test Number y in Table 3.2.

3.3 Results and Analysis

Table 3.2 shows the NO_x emissions from each vehicle for each test as well as the mean exhaust temperature, while Figure 3.6 displays the cumulative NO_x output over each test. Since the ambient temperature was lower for Vehicle 1 tests, the average exhaust temperature is lower. The effect of ambient temperature on NO_x output is shown in Table 3.2. It should be noted that the second test on each day (Tests 1.2, 1.4, 2.2 and 2.4) have higher average temperatures than the first test on each day (Tests 1.1, 1.3, 2.1 and 2.3) since some heat is retained in the engine and exhaust system during the pause in testing. This results in these second tests producing less NO_x. It should be noted that even in the most favourable conditions (second test of the day in warmer weather – i.e., tests 2.2 and 2.4), the vehicles emit NO_x in excess of three times the Euro V heavy-duty limit determined in steady state and transient testing (at 2.0 g/kWh).

TABLE 3.2: NO_x emissions from the two vehicles on each of the four tests.

Vehicle	Test Number	Mean Exhaust Temperature (°C)	NO_x (g/kWh)
1	1	137	13.5
1	2	151	7.9
1	3	148	13.2
1	4	169	7.4
2	1	212	8.4
2	2	224	6.6
2	3	214	9.0
2	4	220	6.4

A GPS trace of the route during a particular run is shown in Figure 3.7. The trace is colourised by the NO_x emissions per kilometre over the route, with the lowest in green, denoting 0 g/km, and the highest in red, denoting 1.7 g/km. It is clear that the city centre is prone to significant NO_x emissions, due to low vehicle speed thanks to congestion and junctions, and frequent bus stops. This causes the SCR to cool significantly, leading to higher NO_x emissions. By contrast, the residential areas see less NO_x emissions. Stops are fewer in this portion of the route, which helps to keep the catalyst at operating temperature (see Figure 3.8).

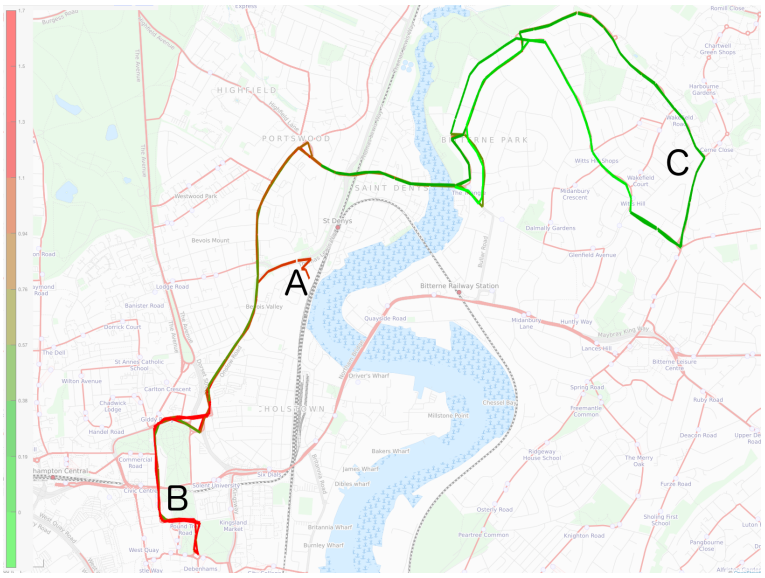


FIGURE 3.7: The vehicle position trace recorded by the telematics unit on Test 2.2. The trace colour denotes NO_x from 0 g/km (green) to 1.7 g/km (red). Labelled are the bus depot (A), the main pedestrian thoroughfare in the city centre (B) and the hilly residential area (C). Map data ©OpenStreetMap contributors [143].

3.3.1 Low Speed Operation

Short stops, such as those made at bus stops, traffic lights or junctions, are typically of a duration of between ten and sixty seconds. During testing, these occur most frequently in the city centre, denoted (B) in Figure 3.7. For instance, Figure 3.9a is a segment of Test 2.1 which shows the vehicle stopping for around sixty seconds at a bus stop, then moving off again without coming to rest again within more than three minutes. The temperature of the catalyst falls significantly during the stationary period, as seen in Figure 3.9a. It is interesting to note that whilst the vehicle is stopped and idling, the exhaust gas has a very low flow rate and temperature. As such, when the vehicle moves off at around 2,420 seconds, this cooler gas is pushed through the catalyst, resulting in yet further cooling. Even with three minutes of sustained engine load, the temperature does not recover. The end result of this is that when the vehicle moves off again, NO_x conversion is significantly worse, as seen in Figure 3.9b: the SCR achieves only 14% NO_x reduction during the period shown.

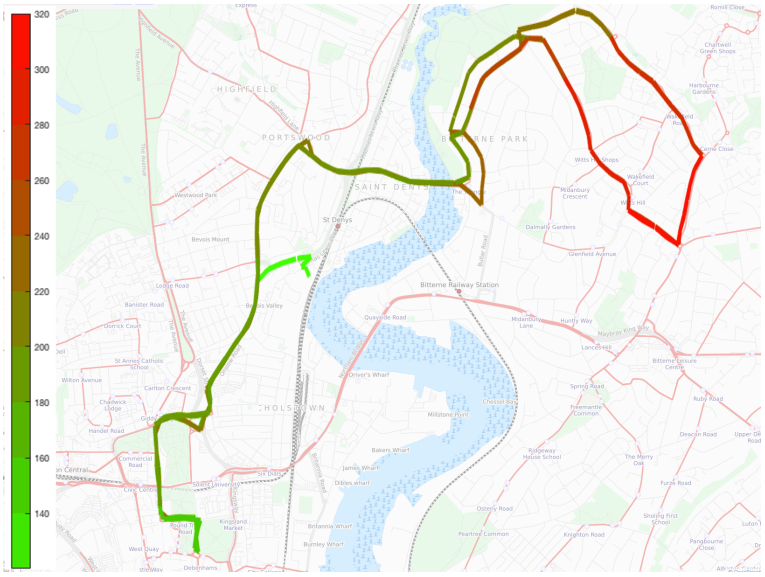


FIGURE 3.8: The vehicle position trace recorded by the telematics unit on Test 2.2. The trace colour denotes catalyst gas outlet temperature from 140°C (green) to 320°C (red). Map data ©OpenStreetMap contributors [143].

3.3.2 Nominal Temperature Operation

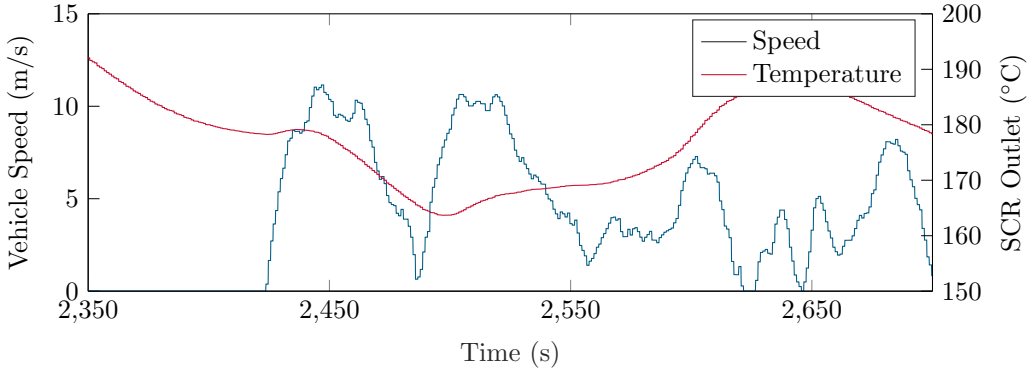
Figure 3.9c shows the catalyst outlet temperature during a period of the journey in which 30 metres of elevation is gained during the first 200 seconds, resulting in maximum torque demand for multiple durations. This is evident in the increasing exhaust temperature, reaching nearly 280°C, meaning the SCR was well within normal operating temperature.

Further evidence of this is seen in the scatter plot shown in Figure 3.10, which plots the mean SCR outlet temperature versus the mass of NO_x emitted in each kilometre of the route for one particular test. The inverse correlation of NO_x emitted with gas temperature is clear.

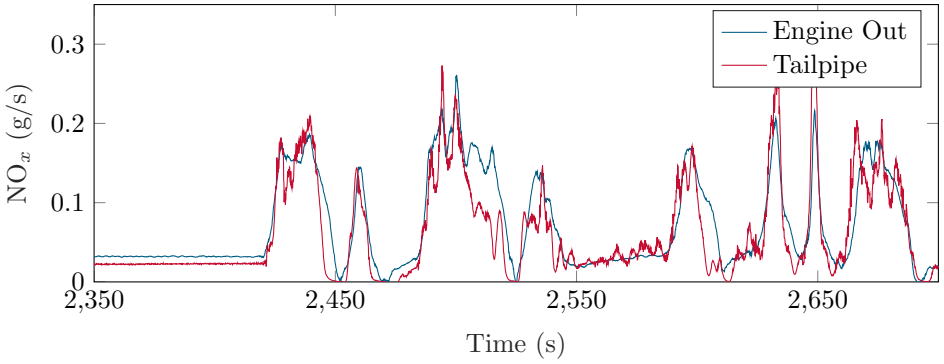
Remark 3.1. During a particular test, a period of just over ten minutes occurred during which an SCR fault was present and the catalyst was non-operational. As such, the tailpipe NO_x data recorded during this period is equivalent to the catalyst inlet NO_x. A neural network with 20 input delay taps and 20 feedback delay taps, and 3 hidden layers was trained using this data and four parameters recorded from the CAN bus: engine fuelling rate, engine speed, engine intake mass air flow and SCR catalyst inlet gas temperature. The results are shown in Figure 3.11.

3.4 Conclusions

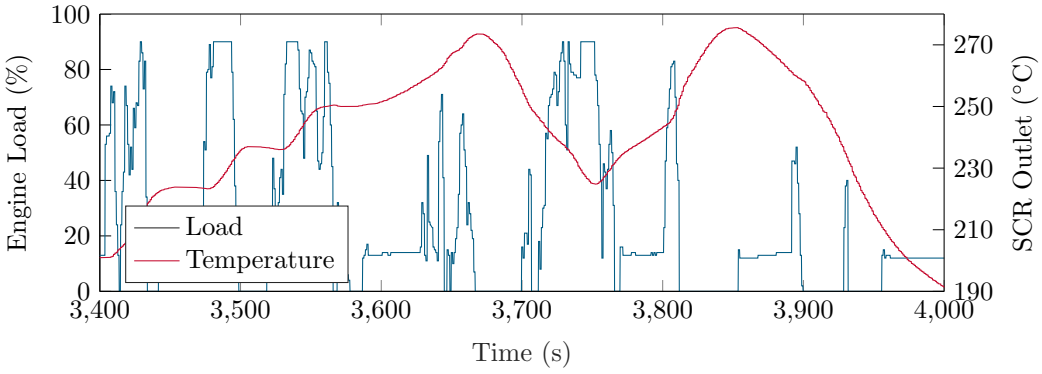
Despite a focus on urban NO_x concentrations and a resulting uptake in aftertreatment technology such as SCR in recent years, city centre environments continue to experience



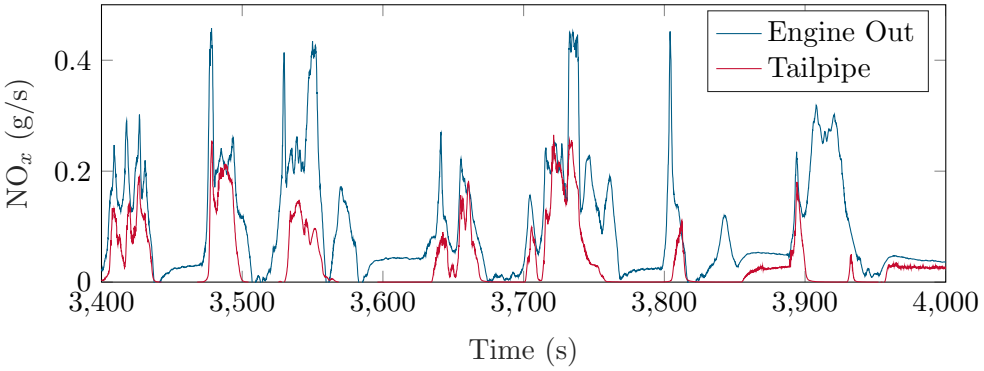
(a) Speed and catalyst temperature in a low load period



(b) Tailpipe NO_x in a low load period



(c) Engine load and catalyst temperature in a high load period



(d) Tailpipe NO_x during a high load period

FIGURE 3.9: Examples of the relationship between vehicle speed/load, SCR outlet temperature, and tailpipe NO_x. Figures 3.9a and 3.9b show a low speed/load period, whilst Figures 3.9c and 3.9d show a high one.

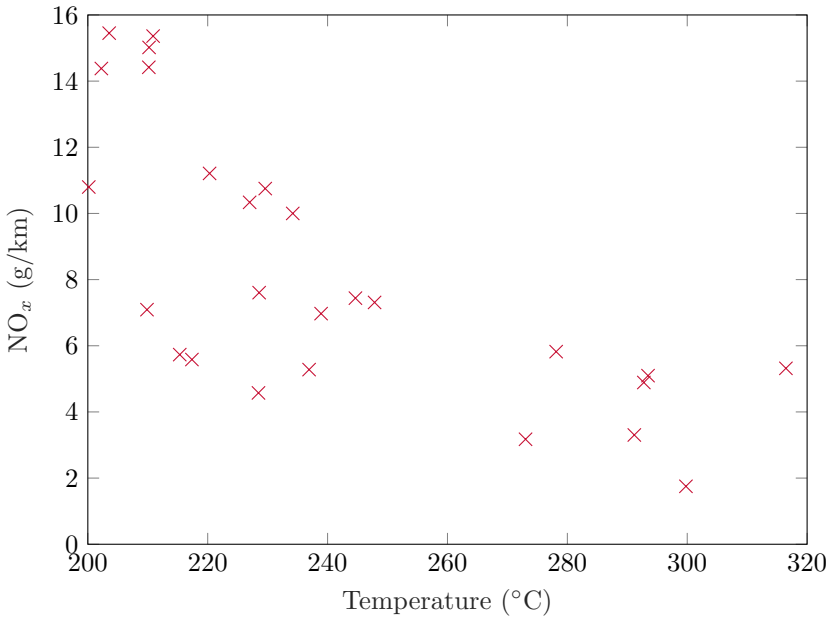


FIGURE 3.10: Mean SCR outlet gas temperature vs mass of NO_x emitted per kilometre on Test 2, Vehicle 2.

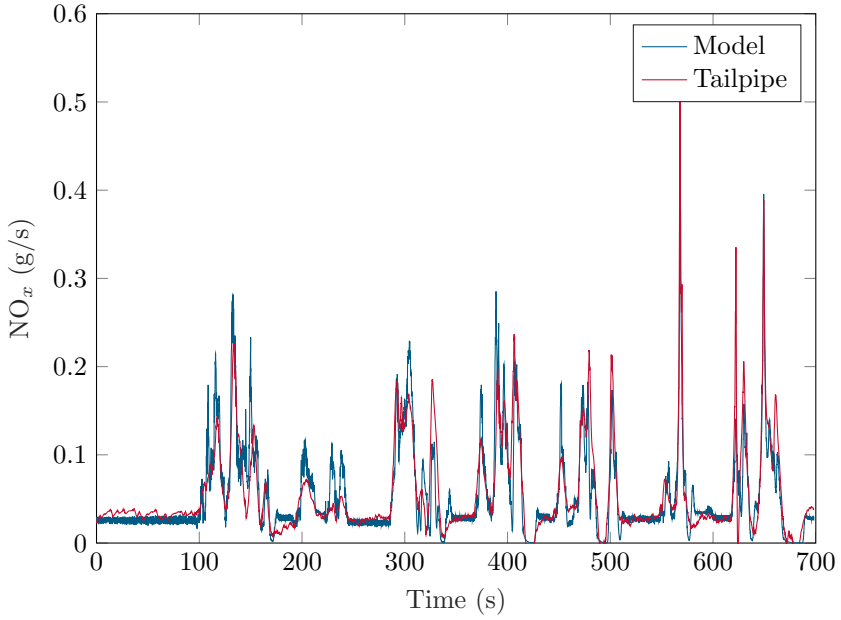


FIGURE 3.11: Training a neutral network to predict engine-out NO_x, with an R-value of 0.864.

concentrations above safe exposure levels. In terms of abatement technology, SCR has proven extremely effective, and thus popular, for the control of NO_x from heavy duty vehicles. In this chapter the results of real-world emissions testing of Euro V buses factory fitted with SCR were presented. It was found that the NO_x emissions were well above their design levels. The temperature of the catalytic surface was identified as having a profound effect on the performance of the catalyst. Whilst the catalyst remains in the operating condition where reaction rates are temperature limited, NO_x conversion is necessarily suboptimal. Urban environments are currently of key concern for NO_x levels, yet it is precisely these circumstances which require frequent stopping and starting of vehicles, which have a deleterious effect on catalyst temperature and therefore emissions. This is particularly true of public transport vehicles like buses, which are mandated to regularly stop even without congestion and junctions, meaning they may be incapable of meeting the emissions limits for which they were intended.

The issue of vehicles exceeding their NO_x emissions limits reported in this chapter does not imply that diesel engines should no longer be considered for this application. Indeed, these problems can be addressed, at least to some extent, by improved control and estimation schemes. The control frameworks proposed in this thesis are designed with the aim of improving efficiency of SCR, providing a pathway to ensuring that vehicles meet their emissions limits not only at design time but throughout their operational life.

Chapter 4

Modelling of SCR

This chapter deals with the structure and calibration of models of the SCR catalyst and surrounding aftertreatment systems. It examines both high fidelity models, which, once calibrated, are used in lieu of a real catalyst for rapid prototyping of controllers, and lower fidelity models which are optimised for execution speed. These fast models are typically employed as the basis of model based controllers, estimators and diagnostics.

4.1 High Fidelity Models

High fidelity models are required for model-based control development. In developing these, a model of the real plant is created which incorporates as many real world effects as possible and thus makes the fewest assumptions about the physical and chemical interactions occurring. Speed is not a concern in this case, instead prioritising fidelity. The models are one dimensional, composed of a series of zero dimensional models, where the degree of axial discretisation is determined by the user. The output of the final element is the output of the complete SCR model. The overall structure is shown schematically in Figure 4.1, where it can be seen that the SCR model consists of three coupled submodels – catalyst (chemical), thermal and pressure drop. This is common to both the SCR and DPF models; the SCR model was generated by taking the relevant elements from the other models. The model is vectorised with the axial elements, and also with a vector of flow rates of each of the chemical species, which includes NO_x:

$$\begin{bmatrix} NO, & NO_2, & O_2, & NH_3, & CO, & CO_2, & \ldots \end{bmatrix} \tag{4.1}$$

The catalyst model implements the equations given in [144] for mass transfer by diffusion between the solid and the gas phases for each of the species (4.1) in each of the elements:

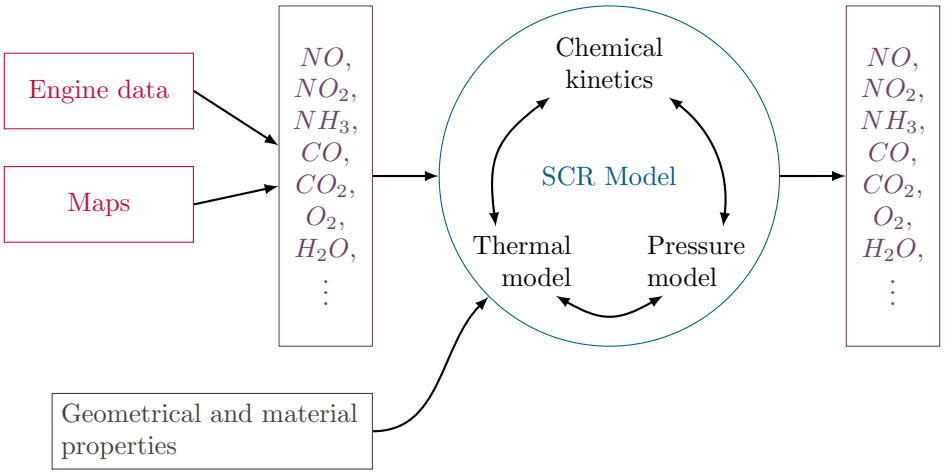


FIGURE 4.1: Top level schematic of SCR model, showing structure, inputs and outputs

$$\begin{aligned} \epsilon\frac{\partial C_g}{\partial t} + \nu_g\frac{\partial C_g}{\partial z} &= -k_m G_a(C_g - C_s) \\ (1-\epsilon)\frac{\partial C_s}{\partial t} &= k_m G_a(C_g - C_s) - Nr \end{aligned} \tag{4.2}$$

where C_g and C_s are the gas and surface phase concentrations respectively, and r is the reaction rate for each species, N is the number of moles of the species, ϵ is the void fraction of the monolith, ν_g is the exhaust gas velocity, k_m is the mass transfer coefficient and G_a is the surface area of the reactor per unit volume. Via some approximations (namely neglecting the term in $\frac{\partial C_g}{\partial t}$), both equations in (4.2) can be rearranged to ordinary differential equations and implemented in the model. The reaction enthalpies of the reactions implemented by this model are derived from published literature, and the catalyst submodel outputs the total heat of reaction for inclusion in the thermal model. It also contains functionality to adjust the storage capacity of the SCR catalyst as a function of temperature; an example model of this is given in [33, Equation 27]. Since the model includes a urea injector, the catalyst submodel also includes functionality implementing the urea decomposition into ammonia as described in Subsection 2.1.4.

The actual reactions implemented in the SCR model are a subset of those given in Subsection 2.1.3 and Subsection 2.1.4. The rates of reactions R in (4.2) are mostly taken from literature, and typically adhere to Eley-Rideal kinetics which are appropriate when one species adsorbs onto a catalyst surface and the other reacts with it directly from the gas phase, as in the case of NH_3 and NO_x respectively. The rate is given by:

$$r = kC_S C_B \frac{K_1 C_A}{K_1 C_A + 1} \tag{4.3}$$

where C_A is the concentration of the species that adsorbs, C_B is that of the one that does not, and C_S is the concentration of active sites on the catalyst. K_1 and k are Arrhenius temperature dependencies, given by:

$$k = Ae^{E_a/RT} \tag{4.4}$$

where A is known as the pre-exponential factor and E_a is the activation energy of the reaction. One notable exception to the Arrhenius type rate is the ammonia adsorption and desorption, since these are physical rather than chemical reactions and therefore depend on the coverage of ammonia on the surface of the catalyst.

The constants A and E_a in (4.4) need to be determined for each reaction in the model using the test data, which has constituted the bulk of the modelling work. Gas stand test data were provided for a small SCR sample, which has a much smaller volume than the real unit. Data was adjusted to have the same space velocity (defined as volumetric flow rate divided by volume of the catalyst). The SIMULINK parameter estimation tool `spetool` was used to carry out the calibration; given the input and outlet catalyst data, the optimiser will adjust parameters via a gradient descent based method (such as interior point) to minimise the RMS error between the inlet and outlet. One of the biggest problems encountered was isolating a particular equation for calibration – this was made somewhat easier by some gas stand runs containing (for example) no NO_x, such that the ammonia storage equation (2.11) is isolated.

The thermal model predicts heat transfer through the monolith due to convection and conduction, which are caused by heat input from incoming exhaust gas and the reactions occurring in the catalyst. The dynamic model is based on that given in [145], which can be summarised as:

$$\begin{aligned} \bar{T_g} &= \bar{T_s} + (T_{g,0} - \bar{T_s})\epsilon_s \\ \bar{T_s} &= \bar{T_g} + (T_{s,0} - \bar{T_s})\epsilon_t \end{aligned} \tag{4.5}$$

where $\bar{T_g}$ and $\bar{T_s}$ are the average gas and surface temperatures in an element respectively, which are then used to calculate the rate of heat transfer to and from the surface. A term in conduction is included according to Fick's second law of diffusion:

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial z^2} \tag{4.6}$$

as well as the heat generated by the reactions, which is provided at each timestep by the catalyst submodel.

The pressure drop model is given by [146], which derives an expression for the pressure drop across a DPF. The DPF is a wall flow filter whereas the SCR catalyst has open ended channels, so the terms in the pressure drop expression for wall flow are neglected for use in the SCR model.

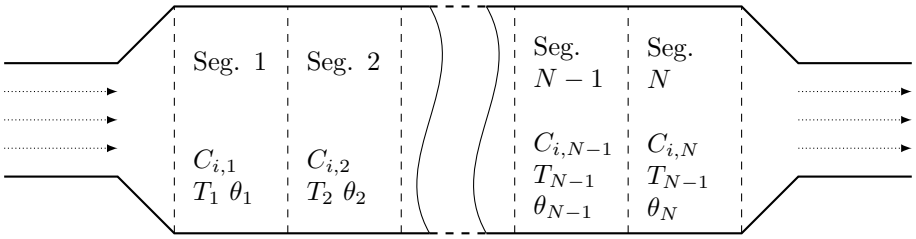


FIGURE 4.2: Multisegment CSTR catalyst model with N segments.

The dynamics of (4.2) are very stiff – a small change in the $C_g - C_s$ term has a large effect on the left hand side of the equation due to the $k_m G_a$ factor being very large. This causes performance problems since it often takes a long time for the solver to converge to a solution, especially during transient events. Originally, the model operated with a fixed step solver, such as SIMULINK's `ode5`, whose step size had to be on the order of the time constant of the fastest dynamics in the model (around 10 ms). A fairly significant amount of work has been undertaken in this area to identify parts of the model that rely on a fixed step solver and reimplement them in a manner amenable to a variable step solver. This has resulted in roughly an order of magnitude speedup, since the solver can now take large steps during periods of steady state operation.

4.2 Calibration and Validation

Calibration of the SCR model is primarily divided into two main areas. Firstly, the geometric and material properties of the catalyst need to be known. Physical dimensions are usually easy to acquire, and material properties such as specific heat capacity, ammonia storage capacity per unit volume, are typically known from the manufacturer of the substrate and its datasheets. Secondly and more problematic are the constants A and E_a in (4.4). These are typically found experimentally [31].

4.3 Control Oriented Modelling

It is clear that the simpler the model the better for online execution in a model based control environment. Since the calculation of each control step requires the solution of the model potentially very many times, refining the model to the minimum elements and fidelity required for acceptable control is of key importance to the performance of the resulting controller. The catalyst is modelled as a Continuously (Ideal) Stirred Tank Reactor (CSTR), which is an ideal reactor model often used to represent chemical processes in which the composition of the reactor output is assumed to be equivalent to the composition inside the reactor (i.e., the reactor is ideally mixed). To account for axial variation in temperatures and concentrations, it is possible to create a multisegment model, as shown in Figure 4.2, in which each segment is modelled as a CSTR. The CSTR

implementation used to model the SCR catalyst in this work is a modification of the CSTR approach in [34].

Each differential equation in the CSTR model accounts for the rate of change of one species of interest. The rate of change of species i is:

$$\frac{dN_i}{dt} = F_{in} - F_{out} + V\nu_i r_i \tag{4.7}$$

where N_i is the number of moles of species i, F_{in} and F_{out} are the molar flow rates into and out of the reactor respectively, V is the volume of the reactor, ν_i is the *stoichiometric coefficient*, and r_i is the reaction rate of species i.

Assumption 4.3.1. By inspection of (4.7), the following conditions are imposed on $r_{i,j}$, where j denotes a particular reaction, for a well conditioned CSTR model:

1. $r_{i,j} = 0$ when $C_i = 0$ for all j where $\nu_{i,j} < 0$,
2. Where $\nu_{i,j} > 0$ for some j, $r_{i,j} \to 0$ when $C_i \to 0$ for all C_i on the left hand side (reactants) of reaction j.

In reality, this assumption is very mild, since it simply restricts the model to physically feasible operating conditions. Specifically, providing bounded inlet concentration $C_{i,in}$, these conditions assure that concentrations are non-negative and are upper bounded by the sum of $C_{i,in}$ and the production of species i by chemical processes in the reactor. The second condition bounds production of species i by availability of reactants. The reaction rate expressions r_1 for standard SCR (2.4) and r_2 for fast SCR (2.5) are given by [33] as:

$$r_1 = k_1\, C_{NO}\, C_{O2}\, \theta\, \Phi\, V^2 \tag{4.8}$$

$$r_2 = k_2\, C_{NO}\, C_{NO2}\, \theta\, \Phi\, V^2. \tag{4.9}$$

From [33], the rate for NO oxidation is:

$$r_3 = k_3 \left(C_{NO}\sqrt{C_{O2}} - \frac{C_{NO2}}{K_{eq}} \right) \tag{4.10}$$

and ammonia storage (2.11) is:

$$r_{ads} = k_{ads}\, C_{NH3}\,(1-\theta) \tag{4.11}$$

$$r_{des} = k_{des} \exp\left(-\frac{E_{des}}{RT}(1-\gamma\theta)\right)\theta. \tag{4.12}$$

In these rates, concentrations (C_i) are in moles per cubic metre, θ is the ammonia coverage ratio (i.e., proportion of the catalyst filled with ammonia), Φ is the storage capacity in moles, V is the catalyst volume, C_{NH3} is the concentration of ammonia in the reactor, R and T are the universal gas constant and temperature respectively, K_{eq} is an equilibrium constant that is given for various NO:NO_x ratios, and the set $\{k_{ads}, k_{des}, E_{des}, \gamma\}$ are constant parameters to be calibrated. For reactions (4.8)-(4.10), the k_i ($i = \{1,2,3\}$) terms are assumed to be Arrhenius type expressions as in (4.4).

Using the CSTR model (4.7), the set of differential equations that govern the evolution of the species of interest in the SCR can be constructed. Following the process detailed in [34], the species NO, NO_2 and NH_3 are of interest, so these will comprise the state vector. The coverage ratio θ appears in all of the rate expressions and cannot be directly measured, so this is appended to the state vector, giving $x = \begin{bmatrix} C_{NO} & C_{NO2} & C_{NH3} & \theta \end{bmatrix}^T$. In contrast to [33], it will be assumed the quantity of ammonia injected ($C_{NH3,in}$) can be directly controlled since the urea decomposition model is quite separate from the SCR process. The resulting nonlinear state space model is:

$$\begin{bmatrix} \dot{C}_{NO} \\ \dot{C}_{NO2} \\ \dot{C}_{NH3} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} -r_1 - \frac{1}{2}r_2 - r_3 + S(C_{NO,in} - C_{NO}) \\ -\frac{1}{2}r_2 + r_3 + S(C_{NO2,in} - C_{NO2}) \\ r_{des} + r_{ads} + S(C_{NH3,in} - C_{NH3}) \\ \frac{V}{\Phi}(r_{ads} - r_{des}) - r_1 - r_2 \end{bmatrix} \tag{4.13}$$

where $C_{NO,in}$ and $C_{NO2,in}$ are the incoming NO and NO_2 concentrations which are regarded as external disturbances. $C_{NH3,in}$ is regarded as the control input. With F as the exhaust volumetric flow rate, $S = \frac{F}{V}$ is known as the space velocity.

The rate equations taken from [31] have published parameters, which are used directly in the model. The standard and fast SCR reactions remained to be calibrated, and this was completed using NEDC emissions cycle data from RICARDO. The overall model was verified against this data using one of the simplest possible controllers for an SCR system. This is a purely feedforward controller based on the principle that, as can be seen from (2.4) and (2.5), there is always a 1 : 1 stoichiometry between NO_x and ammonia. As such, the ideal quantity of ammonia to inject is the same as that of NO_x. This is known as a fixed Ammonia-to-NO_x Ratio (ANR) controller. Whilst the ideal theoretical ANR is 1, the actual optimal value can vary from this due to the fact that not all ammonia reacts with NO_x, and the ammonia slip limit must be respected.

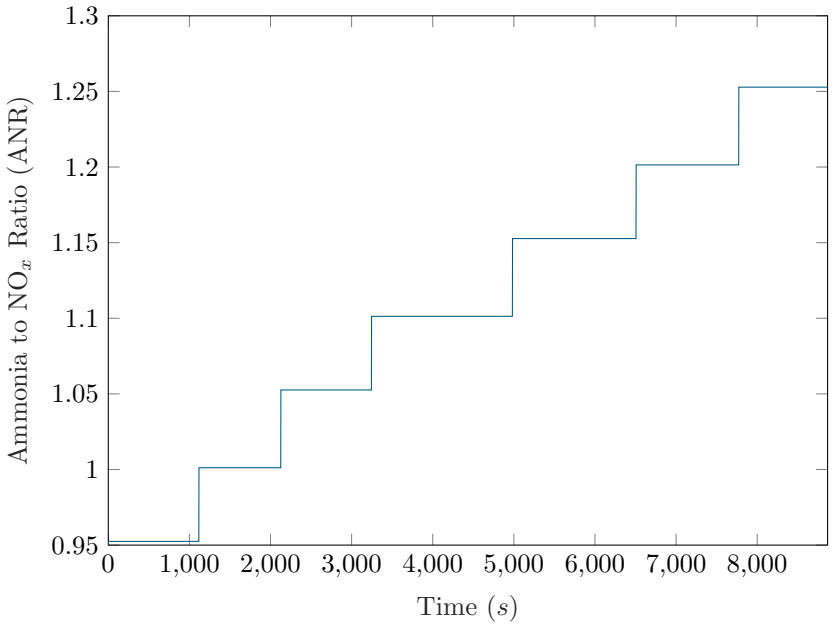


FIGURE 4.3: Ammonia-to-NO_x ratio sweep targets during the catalyst characterisation.

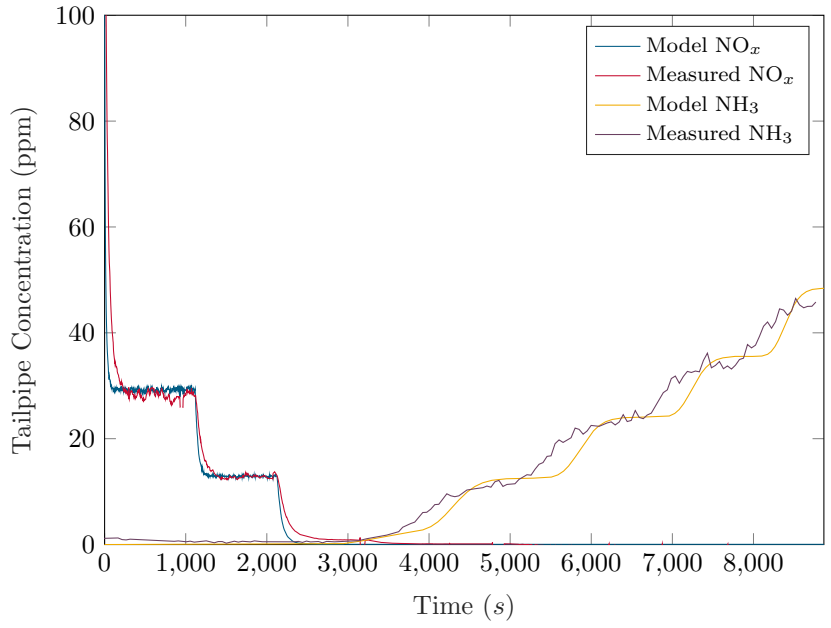


FIGURE 4.4: Model versus measured tailpipe NO_x and NH_3 during ANR sweep characterisation.

A CSTR model of an SCR comprising sixteen segments was calibrated using testbed data. The testbed data comprises exhaust NO_x and temperature being held constant, whilst the catalyst inlet ammonia was swept over a range of ammonia-to-NO_x (ANR) values, as shown in Figure 4.3. The model parameters were then adjusted using a parameter optimisation routine to minimise the discrepancy between predicted and measured outlet NO_x and ammonia, with the optimal calibration shown in Figure 4.4. The resulting ammonia storage fraction in each segment is shown in Figure 4.5.

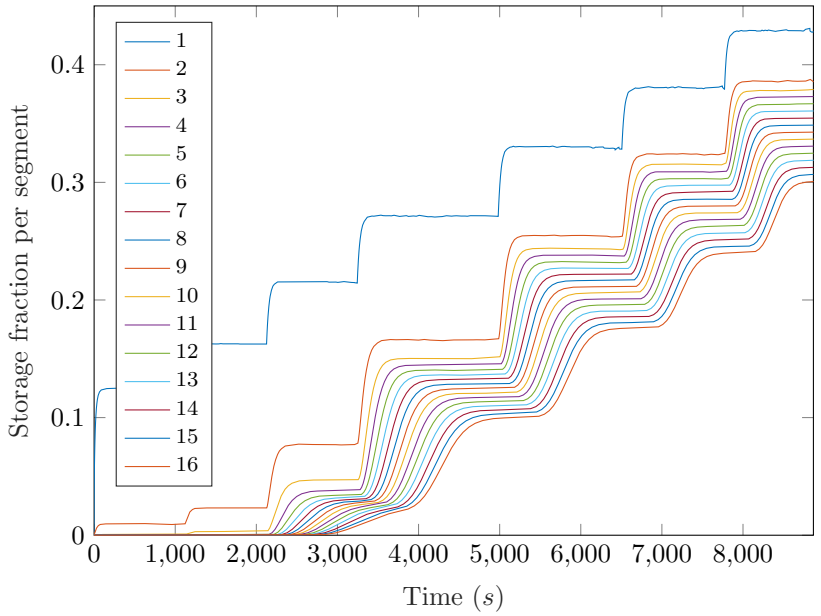


FIGURE 4.5: Storage fraction in each of the 16 segments during the ANR sweep calibration.

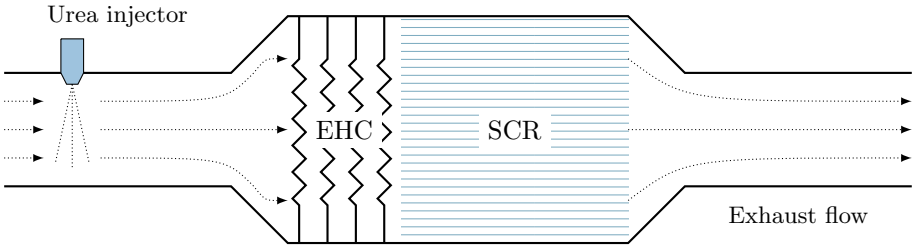


FIGURE 4.6: Schematic of an SCR catalyst fitted with an electric catalyst heater (EHC).

4.4 Electric Catalyst Heater

Modelling of the EHC is relatively simple and is just a first order gas flow model:

$$\Delta T = \frac{P_{ehc}}{\dot{m} c_p} \tag{4.14}$$

where P_{ehc} is the power supplied to the heater in Watts, $\dot{m}$ is the mass flow rate, c_p is the specific heat capacity of the exhaust gas at constant pressure and ΔT is the temperature rise from heater inlet to outlet. This relatively simple model is easy to include in embedded control applications. This model assumes 100% efficient conversion of electrical power into heat, and assumes that all of the generated heat is transferred to the gas stream.

4.5 Concluding Remarks

This chapter examined the modelling frameworks that have been used in the literature for SCR, both high fidelity offline models and approximate models that have been used for model-based control. The CSTR framework has been examined and a model has been developed and calibrated using data acquired from a real SCR on a testbed. It is phrased in such a manner so as to be applicable to most SCR chemistries, just requiring calibration either from an engine test bed or from gas stand catalyst samples.

In the next chapter, these models will be used to build model-based control schemes, both for the internal controller model and as plant models when benchmarking performance. It will be demonstrated how the CSTR framework permits significant flexibility in the model, both in terms of the computational complexity versus fidelity, and including and excluding reactions as appropriate.

Chapter 5

Reference Governors for SCR

Current industrial strategies for automotive SCR are typically based on driving the ammonia storage in the SCR to a particular value, which is typically a function of catalyst temperature and gas flow rate through the SCR. This is typically stored in a map which is calibrated at engine design time using a testbed. Since the constrained quantities, i.e., pollutant emissions and actuator inputs, are not included in the controller, the map is typically highly conservative, and provides no runtime guarantee of constraint satisfaction.

Typically, the controller is implemented to drive the ammonia storage to its target value of the PID type. One concept which has proved highly popular in the automotive field for adding constraint handling and thus permitting plants to operate closer to optimal setpoints is the reference governor. One of its principal advantages is that it can be added in a supervisory capacity to an existing PID controller, only intervening where constraint violation is predicted, or where a modified setpoint would allow closer-to-optimal operation.

In this chapter, the reference governor concept is introduced and how it can be added to existing SCR controllers is discussed. Its advantages over standard controller topologies are demonstrated by simulation. Finally, the reference governor is implemented in hardware-in-the-loop using an automotive rapid prototyping platform, showing that this scheme is feasible for online implementation on typical automotive grade hardware.

5.1 Review of the Reference Governor

One of the principal disadvantages of classical control strategies is their lack of constraint handling, since many systems across the control field are subject to constraints on their inputs, process variables and outputs. One option for handling pointwise-in-time constraints which has seen success in many fields including automotive is the *reference*

governor. Under this framework, a conventional control scheme is augmented with a supervisory layer which may adapt the closed loop reference such that constraints are met. Typically, the implementation is such that the reference governor will act only to modify the closed loop behaviour if constraint violation is foreseen; otherwise the reference should be allowed to be as close as possible to the conventional controller demand [147]. In this chapter, a model-based reference governor for an SCR application is designed and benchmarked. The decoupling of tuning of the primal controller simultaneously for constraint satisfaction and performance is shown to lead to a simpler calibration process. Furthermore, reference governor is implemented on a rapid prototyping platform for vehicle ECUs and demonstrate that it is capable of running in real time.

Consider a nonlinear continuous time dynamic system

$$\dot{x}(t) = f(x(t), u(t)) \tag{5.1}$$

$$y(t) = h(x(t), u(t)) \tag{5.2}$$

where $x \in \mathcal{X} \subseteq \mathbb{R}^n$ is the state, $u \in \mathcal{U} \subseteq \mathbb{R}^m$ is the input and $y \in \mathcal{Y} \subseteq \mathbb{R}^p$ is the output. It is assumed that f and h are smooth and continuous across the state and input spaces.

A conventional PID controller is used to bring this system under closed loop control and bring the output $y(t)$ to the reference $r(t)$. Define the set of constant references for which the output y is constraint admissible throughout the prediction horizon H_p

$$\mathcal{W}_y = \{r : y \in \mathcal{Y}\}, \tag{5.3}$$

and similarly for the control input u

$$\mathcal{W}_u = \{r : u \in \mathcal{U}\}, \tag{5.4}$$

which are aggregated to find the set of references that respect all constraints

$$\mathcal{W} = \mathcal{W}_u \cap \mathcal{W}_y \tag{5.5}$$

If $r(t) \notin \mathcal{W}$, the reference governor will define a modified reference $w(t)$ for the conventional controller

$$w(t) := \underset{w \in \mathcal{W}}{\arg\min} \quad \|r(t) - w\| \tag{5.6}$$

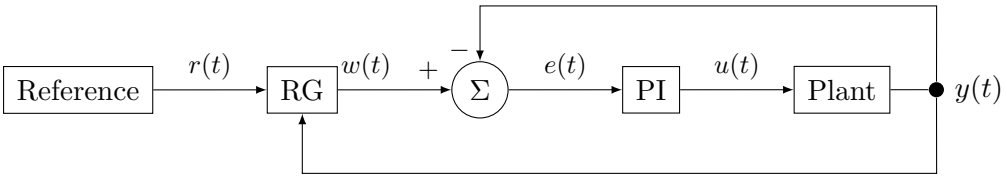


FIGURE 5.1: Reference governor schematic

Examples of automotive application of reference governors include that of managing boost pressure in a diesel engine with a gradient descent algorithm [148] and control of aftertreatment temperature [149].

5.2 Ammonia Slip Catalyst

Unreacted ammonia exiting the tailpipe is damaging to both living organisms and the environment. As such, constraining the so-called *ammonia slip* is one of the key targets of SCR control. To permit higher concentrations of ammonia in the catalyst in order to promote adsorption (and hence NO_x reduction), an *ammonia slip catalyst* is often placed downstream of the SCR in the exhaust system. This device is tailored to exhibit high selectivity towards (2.12). A representative example of the temperature dependent ammonia conversion efficiency is shown in Figure 5.2.

5.3 Conventional Controllers

Industry standard practice for the control of SCR involves controlling of the ammonia stored in the catalyst. During vehicle calibration, a lookup table is created that defines the storage target ϕ_{ref} for each monolith temperature and space velocity. An observer is used to estimate the current storage of ammonia ϕ in the catalyst, and ammonia injection is scheduled to bring that to the target, i.e. to minimise $\|\phi - \phi_{ref}\|$. For this investigation, a proportional-integral (PI) controller was configured with the controller gains $k_p = 7 \times 10^{-5}$ and $k_i = 1 \times 10^{-8}$.

5.4 Reference Governor Design

Since reaction rates are functions of the monolith temperature T_{mon} whereas only the upstream exhaust temperature $T_{exh,in}$ is measured, a simple first order thermal model is utilised. A parameter k_M models the convectional heat transfer from the exhaust gases to the monolith, whilst a second parameter k_A models heat loss from the monolith to the environment

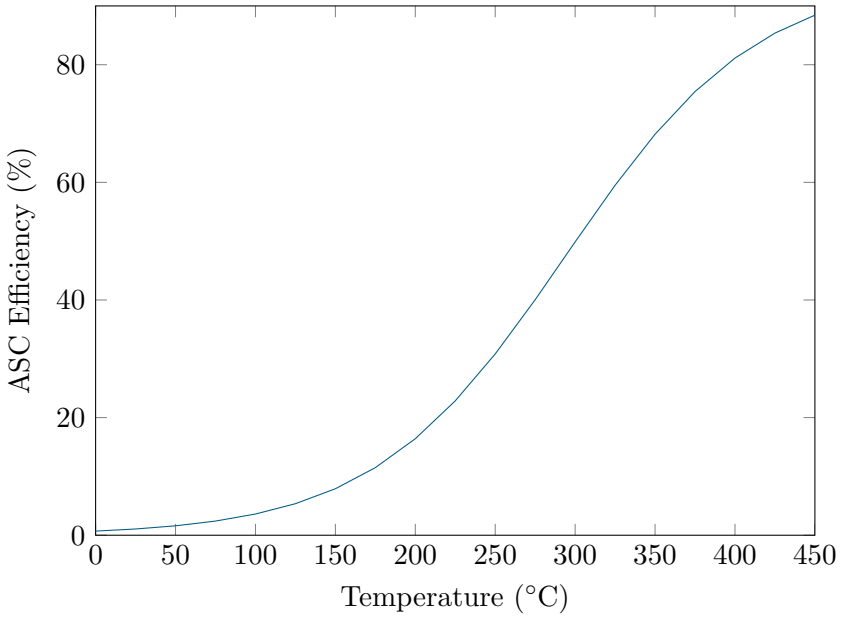


FIGURE 5.2: Ammonia slip catalyst conversion efficiency

$$\dot{T}_{mon} = k_M(T_{exh,in} - T_{mon}) - k_A(T_{mon} - T_{amb}) \tag{5.7}$$

Define the vector of state variables

$$x = \begin{bmatrix} C_{NO_x} & \phi & C_{NH_3} & T_{mon} \end{bmatrix}^T \tag{5.8}$$

along with the control input vector

$$u = \begin{bmatrix} NH_{3,inj} \end{bmatrix} \tag{5.9}$$

and the disturbance input vector

$$d = \begin{bmatrix} NO_{x,in} & T_{exh,in} \end{bmatrix}^T. \tag{5.10}$$

Using the equations and reaction rates above, it is possible to model the SCR reactor using a continuously stirred tank reactor (CSTR) framework. This is an idealised reactor model in which the contents are assumed to be mixed perfectly and instantaneously, hence the output concentrations are equal to the reactor concentrations. The CSTR equation is

$$\frac{dN_i}{dt} = F_{in}(t) - F_{out}(t) + V\nu_i r_i(t) \tag{5.11}$$

where F_{in} and F_{out} are the molar flow rates of the species in and out of the reactor respectively, V is the reactor volume, ν_i and r_i are the stoichiometric coefficient and reaction rate of species i respectively. The SCR reactor model can now be phrased as

$$\begin{bmatrix} \dot{C}_{NO_x} \\ \dot{\phi} \\ \dot{C}_{NH_3} \\ \dot{T}_{mon} \end{bmatrix} = \begin{bmatrix} \phi R_{oxi} - C_{NO_x}(\phi R_{red} + F) \\ R_{ads} C_{NH_3} \Phi - \phi(R_{ads} C_{NH_3} + R_{des} + R_{red} C_{NO_x} + R_{oxi}) \\ \phi R_{des} - C_{NH_3}(R_{ads}(\Phi - \phi) + F) \\ -k_M T_{mon} - k_A(T_{mon} - T_{amb}) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{V} \\ 0 \end{bmatrix} u(t) + \begin{bmatrix} \frac{1}{V} & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & k_M \end{bmatrix} d(t) \quad (5.12)$$

The model (5.12) is very stiff, since examination of the reaction rates demonstrates that the C_{NO_x} and C_{NH_3} states are around 3 orders of magnitude faster than the ammonia storage state ϕ. To increase the execution speed of the predictive model in the reference governor, neglect the dynamics of the two species, i.e. assume they are constant during each time step of the model:

$$\dot{C}_{NO_x}(t) = \dot{C}_{NH_3}(t) = 0 \quad (5.13)$$

resulting in a 2-state model which can be executed with a longer sampling period. Instead, compute the predicted ammonia slip at the time instants at which constraints are to be enforced

$$C_{NH_3}(t) = \frac{\frac{u(t)}{V} + \phi(t) R_{des}(t)}{R_{ads}(t)\Big(\Phi(T) - \phi(t)\Big) + F(t)} \quad (5.14)$$

Transferring of the computational burden from the state equation (5.1) to the output equation (5.2) in this manner is of great utility if, as is the case in this work, constraints are to be enforced less frequently than the sample period of the model (5.12), since (5.14) needs to be evaluated only at the constraint enforcement points.

5.4.1 Cost Function

Following the cost function study in [149], the cost function is designed to include a weighed sum of the constraint violations over the prediction horizon:

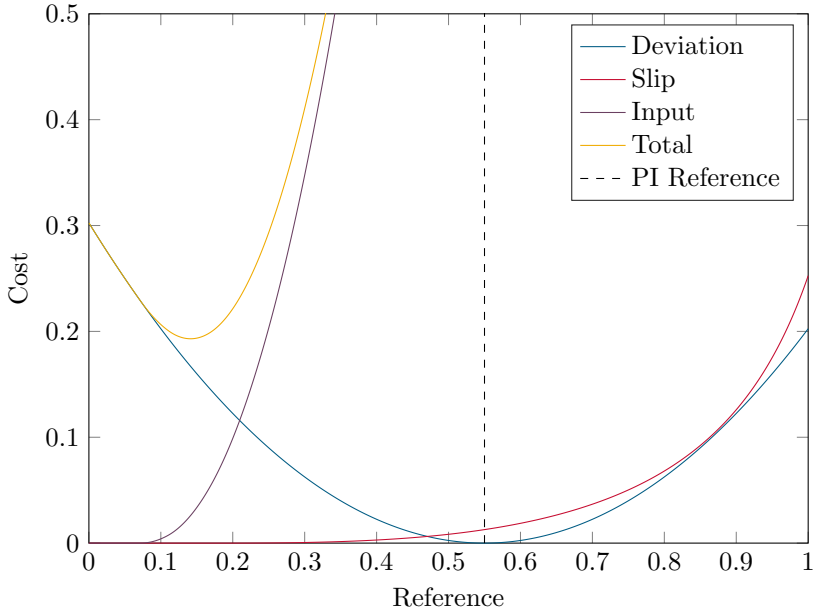


FIGURE 5.3: The three components of the cost functions

$$J(r,w) = \|r-w\|^2 + \rho_x \left[\sum_{i=0}^{N} \max\left(\hat{c}(k+i|k) - \bar{c}, 0\right)\right]^2 + \rho_u \left[\sum_{i=0}^{N} \max\left(\hat{u}(k+i|k) - u_{max}, 0\right)\right]^2 \quad (5.15)$$

where ρ_x and ρ_u are the penalties applied to state and input constraint violations respectively.

The square of the total constraint violation is weighted (rather than just the violation) for two reasons. Firstly, very small violations of ammonia slip and injection quantity limits are tolerable. Secondly, this is one method of guaranteeing strict convexity of the cost function, thereby generating an optimisation programme that has well behaved properties. As in typical reference governor design, a term is added to penalise the deviation of the adjusted reference $w(t)$ from the existing reference $r(t)$. This is visualised in Figure 5.3 where the deviation cost, the ammonia slip cost and the input violation cost are plotted for $t = 0$ in the simulation results in Section 5.5.

5.4.2 Optimiser

Important considerations in selection of the solver were the computational expense and dimensionality of the cost function (5.15). Since evaluation of the predictive model is the most expensive step, a bisection method [150] was adopted where evaluation

points were chosen according to the golden section method. This has the benefit of a guaranteed interval reduction at each iteration, reducing the possibility of ill conditioning of the algorithm, and providing a deterministic upper bound on the runtime for a given termination tolerance. The golden section rule implies that since triples of search points maintain the same interval spacing, only one evaluation of the objective function is made during each iteration. A further advantage of this type of search is that only one preceding interval is maintained, meaning the memory requirements do not scale with the quantity of iterations. This means the memory requirements are quantifiable and fixed, making selection of implementation hardware and safety guarantees possible.

The initial search bracket is defined as

$$[w_a, w_b] = [-0.5, 1.5] \tag{5.16}$$

since although $\theta \in [0, 1]$ due to physical constraints, the reference governor may wish to place the reference outside this range in transient operation. The algorithm searches for a minimum $w^\star$ of the function $f(w)$ as described in Algorithm 5.2.

Algorithm 5.2 Golden section search

$\Phi = \frac{-1+\sqrt{5}}{2}$
$w_c = w_b + \Phi(w_a - w_b)$
$w_d = w_a + \Phi(w_b - w_a)$
while $|f(w_a) - f(w_b)| > \epsilon$ **do**
    if $f(w_c) < f(w_d)$ **then**
        $w_b \leftarrow w_d$
        $w_d \leftarrow w_c$
        $w_c = w_b + \Phi(w_a - w_b)$
    else
        $w_a \leftarrow w_c$
        $w_c \leftarrow w_d$
        $w_d = w_a + \Phi(w_b - w_a)$
    end if
end while
$w^\star = \frac{w_a+w_b}{2}$

5.5 Results

Defining the space velocity

$$SV := \frac{F}{V}, \tag{5.17}$$

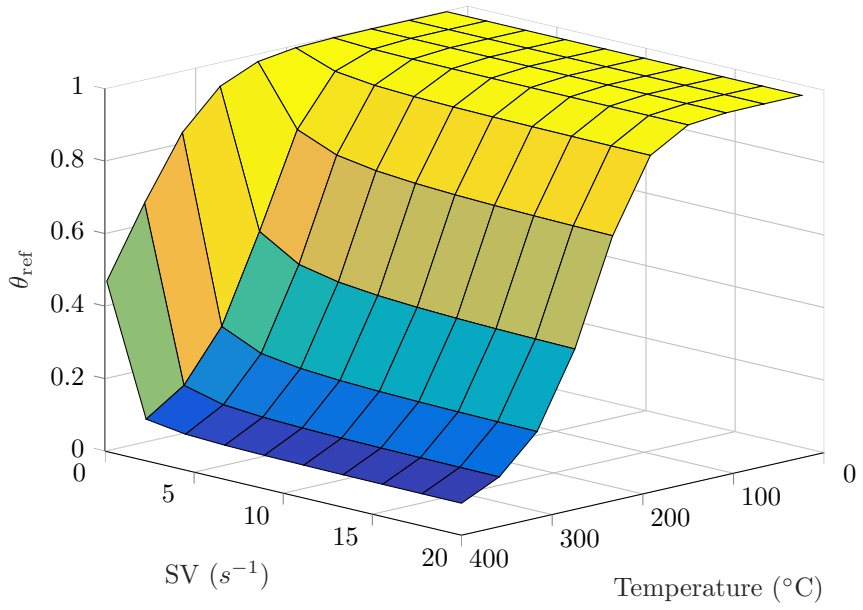


FIGURE 5.4: Target ammonia storage as a function of monolith temperature and space velocity to achieve 5 ppm ammonia slip in steady state.

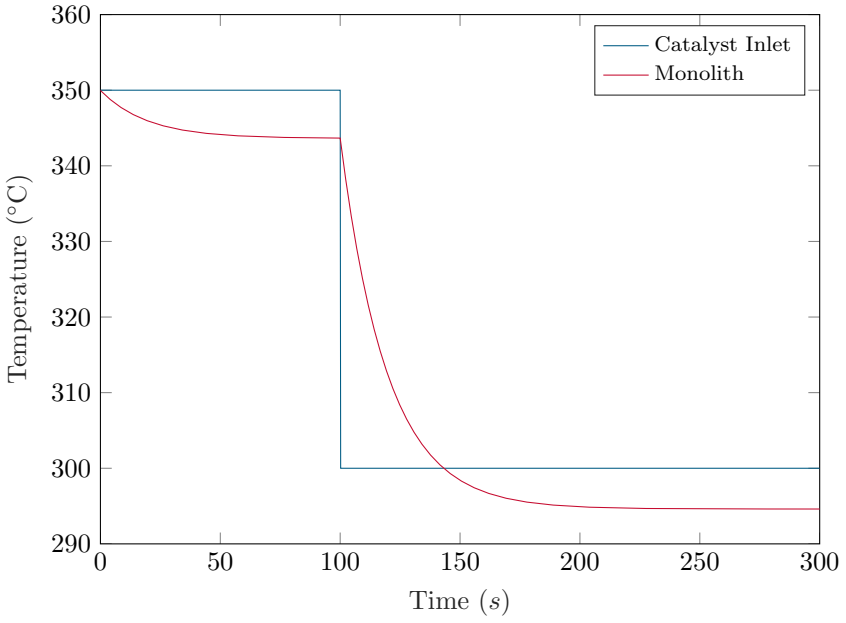
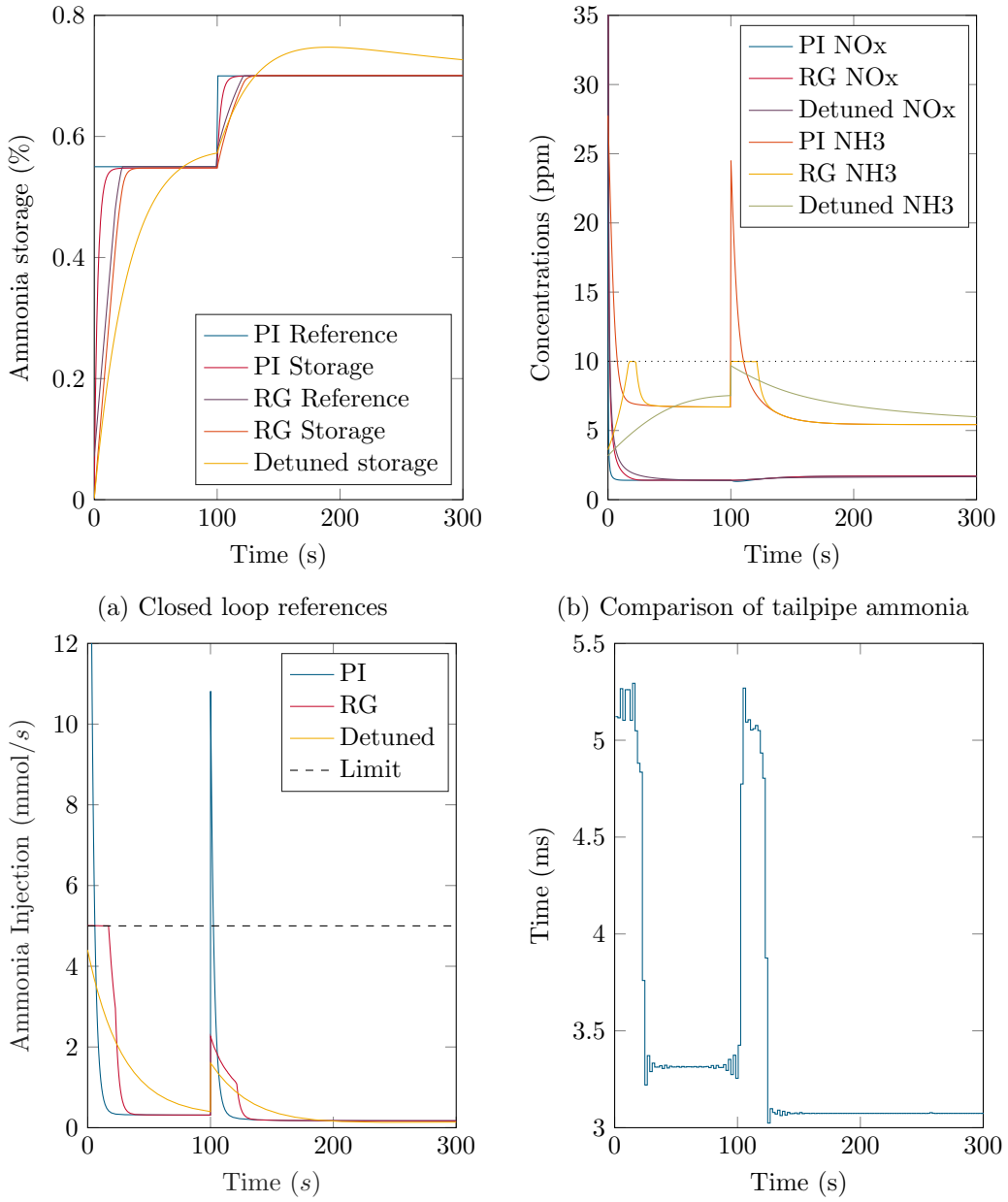


FIGURE 5.5: Catalyst monolith temperature, showing a 50°C fall at $t = 100$s

a storage target map $\theta_{ref}(T_{mon}, SV)$ was generated in order to maintain a tailpipe ammonia slip of around 5 ppm in steady state, as displayed in Figure 5.4. In order to present a challenging scenario for the controller, a simulated empty catalyst at 350°C was to be brought to an ammonia storage target of 55%, followed 100 seconds into the simulation by a fall in exhaust temperature of 50°C, as shown in Figure 5.5, at which ϕ_{ref} was 70%. Constraints are to be enforced at a tailpipe ammonia slip of 10ppm and ammonia injection of 5mmol/s.



(a) Closed loop references

(b) Comparison of tailpipe ammonia

(c) Comparison of ammonia injection

(d) Control computation time

FIGURE 5.6: Results of the reference governor to the step change in exhaust gas temperature.

The original and modified references are shown in Figure 5.6a (blue and yellow respectively). The ungoverned controller violates both the ammonia slip limit, as seen in Figure 5.6b, and the input limit, as seen in Figure 5.6c. Conversely, the ammonia injection is limited within the $0 - 20$ seconds period by the injection limit, and elsewhere in the simulation by the ammonia slip limit, both of which can be seen to be respected.

A detuned PI controller was also designed ($k_p = 8 \times 10^{-6}$ and $k_i = 5 \times 10^{-9}$) to demonstrate the loss of performance with a conventional controller tuned to respect the specific constraints. Note that whilst constraints are respected in this example, they

FIGURE 5.7: The Ricardo rCube2 rapid prototyping electronic control unit.

were not systematically handled and thus no guarantees of constraint satisfactions can be made for the closed loop system. This loss of performance is evident in Figure 5.6a in the poor tracking of the ammonia storage target.

5.6 Embedded Application

Applicability of the reference governor designed in this work highly depends on its computational requirement compared to that available on commercial electronic control units used in aftertreatment control. To demonstrate that the reference governor is suitable, it is implemented in hardware-in-the-loop (HiL) on the Ricardo rCube2 rapid prototyping electronic control unit, shown in Figure 5.7. This platform comprises a pair of Infineon Tricore TC1797 microprocessors each running at 150MHz. The reference governor HiL demonstration uses one processor for the controller, and runs the plant model on the other, with the master processor also responsible for communication and calibration via the universal measurement and calibration protocol (XCP), which is defined by ASAM (Association for Standardisation of Automation and Measuring Systems). This is implemented over an Ethernet connection to the device, as shown in Figure 5.8. The hardware implementation of model based controllers is explored further in Chapters 8 and 9.

The time required by the reference governor for computation of the adjusted reference $w(t)$ is shown in Figure 5.6d. With a maximum and mean time of 5.4 ms and 3.4 ms respectively, this is well below the control timestep of 100 ms, meaning the reference governor is feasible for real-time application. As expected, periods where imposed constraints are active (e.g. the ammonia slip constraint in Figure 5.6b and the input constraint in Figure 5.6c) correspond to longer computation times.

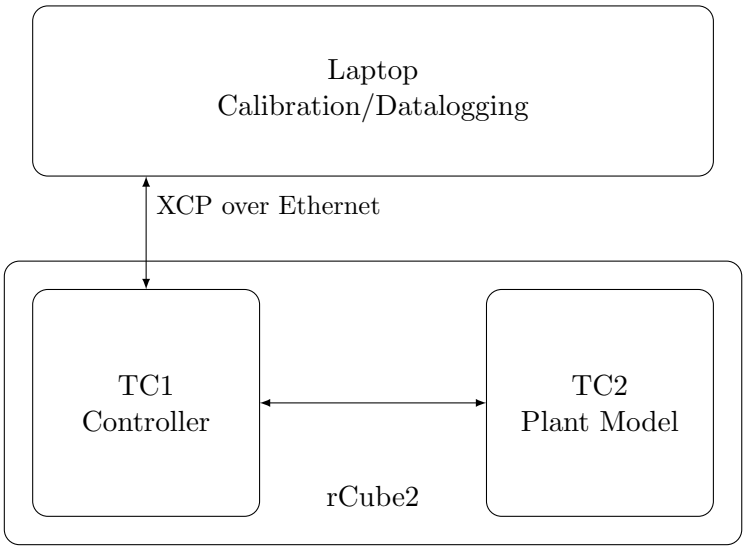


FIGURE 5.8: HiL setup comprising the rCube2 rapid prototyping platform with two Tricore microprocessors (TC1 and TC2) and calibration/datalogging machine

As explained earlier, the implementation of reference governor outlined in this work does not require variable amounts of runtime memory: a constant 408 bytes despite varying computation times and number of active constraints.

5.7 Test Cycle Verification

To demonstrate the performance of the reference governor, testbed data was acquired for a heavy duty engine over the non-road test cycle (NRTC). This included NO_x flow rate, exhaust mass flow rate, and SCR inlet temperature. The reference governor primal PI controller was tuned as in Section 5.3, using the space velocity and temperature map in Figure 5.4 to determine the storage reference. The catalyst was simulated as initially empty ($\theta = 0$). The static map reference, modified reference, and true storage are shown in Figure 5.9.

Keeping the ammonia slip limit at 10 ppm, Figure 5.10 shows that the reference governor has to regularly intervene to respect this limit. It is apparent that the NO_x conversion is poor during the first 200 seconds, due to the controller attempting to build up ammonia storage in the catalyst without violating the input constraint, as shown in Figure 5.11.

5.8 Conclusions

Management of ammonia slip whilst maintaining conversion efficiency in the face of disturbances is one of the key challenges in improving the working efficiency of urea SCR systems. The model based reference governor presented in this chapter provides a

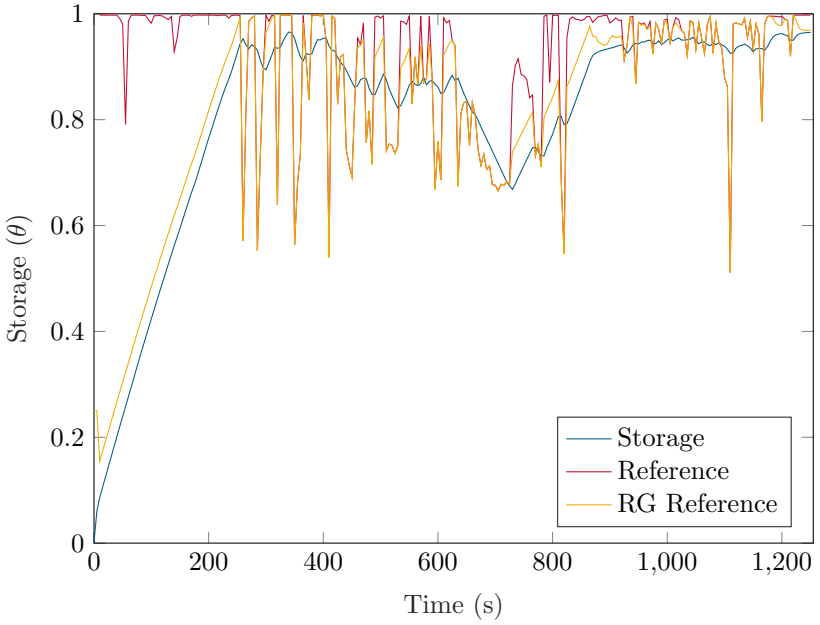


FIGURE 5.9: The reference governor continually modifies the storage reference during the NRTC cycle.

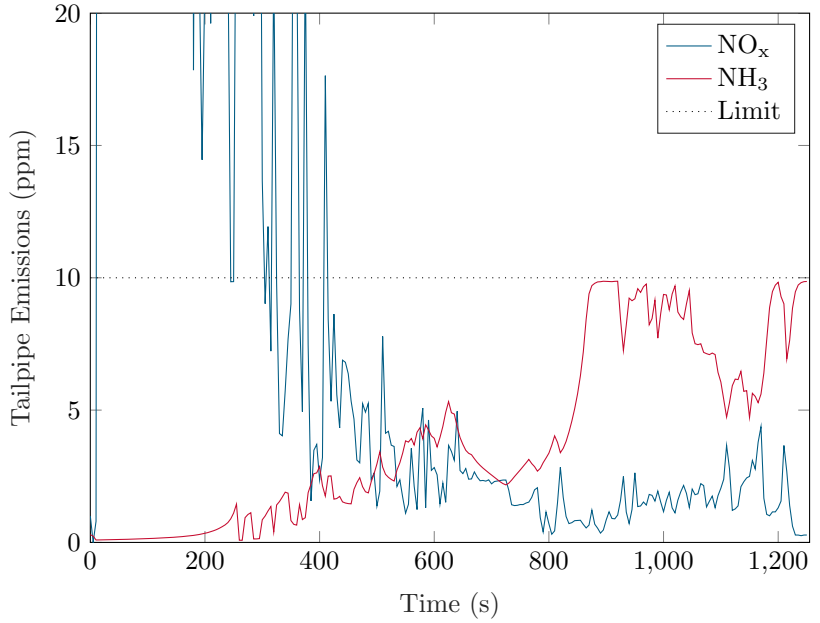


FIGURE 5.10: The reference governor keeps the tailpipe ammonia slip within the 10 ppm limit.

method for adding constraint handling into the controller in a systematic way, providing simple integration into existing control strategies. It thus permits more aggressive tuning of the low level controller, resulting in improved catalyst efficiency and reduced tailpipe emissions. It also decouples tuning the controller for performance and for constraint adherence, leading to simpler calibration.

This chapter has demonstrated that such a model based reference governor is suitable

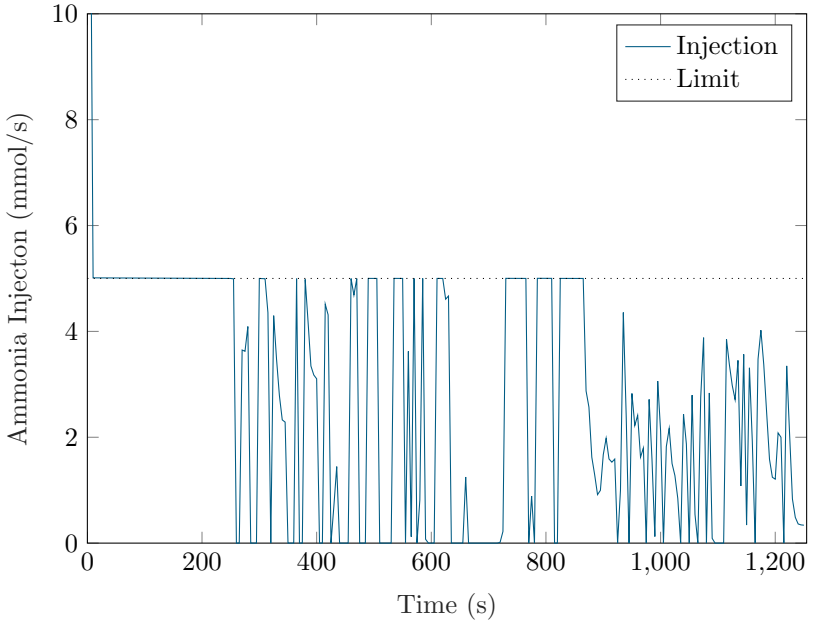


FIGURE 5.11: The reference governor respects the input limit of ammonia injection demand no greater than 5 mmol/s, and thereby avoids wind up of the integrator in the primal controller.

for real-time implementation on automotive electronic control units. Such a route to production is vital for such controllers to become accepted and integrated into industrial practice. The quantifiable upper bound on execution time and memory requirements aid hardware selection and controller certification.

Chapter 6

Predictive Control of SCR

Model predictive control (MPC) holds several advantages compared to the reference governor. Not only does it remove the requirement to design and tune a primal controller, but it allows natural extension to multiple-input, multiple-output (MIMO) control systems. Reference governors for MIMO systems suffer the same problems as interacting PID loops, wherein tuning is difficult due to loop interactions. Instead, MPC directly optimises a cost function to calculate the optimal inputs to the plant.

In this chapter, an economic cost function for SCR is formulated in order to minimise pollutant emissions from the catalyst. Implementation is carried out with both a linear and nonlinear model, and their performance is benchmarked using test cycles. The use of the nonlinear model is shown to be necessary in order to operate the plant over the required temperature range. The economic cost function presents two distinct advantages over the storage tracking MPC for SCR that exists in the literature: namely improved closed loop performance and avoidance of calculating the steady state conditions. Furthermore, two key properties of the nonlinear economic MPC formulation are examined: stability and feasibility. Finally, a catalyst heater is added to the MPC formulation and it is demonstrated that this improves NO_x conversion significantly when exhaust temperatures are low.

6.1 Linear Predictive Control of SCR

An SCR controller determines the quantity and frequency of urea injection over a catalyst, with the objective of minimising the NO_x emitted from the catalyst. This is done against the constraints of actuators, as well as the constraint of a maximum ammonia tailpipe concentration which is given by legislation.

Recall that a model of the SCR controller was developed in Chapter 4, and this will be used as the basis of the MPC schemes designed in this chapter. The thermal model is also

employed at the end of this chapter, where the thermal management of the catalyst is considered in order to mitigate the problems that were identified with operating efficiency in Chapter 3.

The MATLAB Model Predictive Control toolbox is used to construct a linear MPC controller. Typically, an SCR plant has two NO_x sensors, one before the catalyst and one at the tailpipe, sometimes alongside a tailpipe ammonia sensor, as in Figure 1.2. In this chapter all states are available from measurements, although this is not the case in reality. In practice, state estimation is necessary to acquire this information, as discussed in Chapter 7.

6.1.1 Linearised Model

A linear MPC controller requires a linearised model of the plant, which it uses to predict the state of the plant over the prediction horizon. However, recalling the model from Chapter 4, it is obvious that the model (4.13) is nonlinear due to products of states in (4.8)-(4.12). For instance in (4.9), the rate is a function of the product of two state variables, C_{NO} and C_{NO2}. As such, to obtain a linear model, an operating point is chosen at which the plant is linearised. Since operating point should be representative of typical operating conditions of the SCR, it is chosen based on the mean NEDC cycle data, resulting in 50 parts per million (ppm) inlet of each of NO and NO_2, an exhaust gas volumetric flow rate of $1.5\text{m}^3\text{min}^{-1}$ and an exhaust oxygen concentration of 12%.

The linearisation is carried out using the MATLAB `linmod` tool, with verification of the linearised model demonstrated by plotting of step responses to each control and disturbance input. Since the linearised plant directly affects the choice of control input, it is important in predictive controller design to ensure that the linearised model accurately reflects the characteristics of the plant, otherwise poor controller performance and damage to the plant may occur. The resulting linear state space model is:

$$A = \begin{bmatrix} 2.1 \times 10^{-5} & 9.9 \times 10^{-5} & 6.0 \times 10^{-5} & 5.0 \times 10^{-4} \\ 2.0 \times 10^{-5} & -6.5 \times 10^{-5} & 1.0 \times 10^{-4} & -4.7 \times 10^{-4} \\ 3.0 \times 10^{-4} & -2.9 \times 10^{-4} & -2.7 \times 10^{-4} & 7.1 \times 10^{-3} \\ 0.043 & -0.041 & -0.038 & 0.99 \end{bmatrix} \tag{6.1a}$$

$$B = \begin{bmatrix} -7.7 \times 10^{-9} & -7.3 \times 10^{-9} & 2.3 \times 10^{-7} & 1.6 \times 10^{-5} \end{bmatrix}^T \tag{6.1b}$$

$$C = \begin{bmatrix} 2.3 \times 10^4 & 0 & 0 & 0 \\ 0 & 2.3 \times 10^4 & 0 & 0 \\ 0 & 0 & 2.3 \times 10^4 & 0 \end{bmatrix} \tag{6.1c}$$

$$D = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T \tag{6.1d}$$

The inlet NO_x concentrations, $C_{NO,in}$ and $C_{NO2,in}$ are regarded as external disturbances.

6.1.2 Controller Design and Tuning

As discussed in Subsection 2.4.1, linear MPC is the term applied when the cost function is quadratic, and the constraints and model are linear. Accordingly, the optimisation programme is a quadratic programme, solution of which is discussed in Subsection 2.4.6. As such, a quadratic cost as given by (2.28) is required. It therefore remains to choose the weighting matrices Q and R in the stage cost, and the terminal cost. The Euro 6 NO_x limit for passenger vehicles is 0.08 g/km and it can be determined that around 60% NO_x reduction is required over the NEDC cycle data, which amounts to an average tailpipe NO_x concentration of 20ppm, to meet this limit. It can be seen that by letting $x_{ref} = u_{ref} = 0$ in (2.67), the reference becomes unreachable except in the trivial case where $CNO, in = C_{NO_2,in} = 0$. This implies that the MPC scheme can now be termed *economic*, as introduced in Subsection 2.4.5. Since we also neglect the input weighting R (as we prioritise controller performance here), the stage cost can be written as:

$$l(x, u) = x^T Q x \tag{6.2}$$

with Q as:

$$Q = \begin{bmatrix} Q_1 & 0 & 0 & 0 \\ 0 & Q_2 & 0 & 0 \\ 0 & 0 & Q_3 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{6.3}$$

where the final row and column are both zero since no weight is put on the ammonia storage θ. It will be shown later that the model and the cost Q uniquely imply an optimal storage $\theta^\star$ which is a function of the states, inputs and disturbances.

A prediction horizon of $t_f = 10$ seconds and a sampling time of $t_s = 0.5$ seconds are chosen in order to capture those parts of the plant dynamics which have longer time constants. Constraints are applied to the state and input values in ppm:

$$x \in \mathcal{X} \Rightarrow \begin{bmatrix} -\infty & -\infty & -\infty & -\infty \end{bmatrix} \leq x \leq \begin{bmatrix} \infty & \infty & 10 & \infty \end{bmatrix} \tag{6.4a}$$

$$u \in \mathcal{U},\ \mathcal{U} := [0, 1000]. \tag{6.4b}$$

The terminal cost $E(\cdot)$ in (2.28c) is typically used as a stability guarantee. By constraining the state to be in a suitable terminal set by the end of the prediction horizon, it

TABLE 6.1: NO_x conversion performance and ammonia slip for the fixed-ANR, PI and MPC controllers

Controller	Mean NO_x Reduction (%)	Mean Ammonia Slip (ppm)	AUS Used (ml)
Fixed ANR (0.8)	61.0	0.2	3.25
Fixed ANR (1.0)	68.0	0.7	4.00
Fixed ANR (1.2)	71.4	1.9	4.71
PI	84.4	9.5	5.09
MPC	77.0	0.8	3.77

is clear that feasibility of the resulting optimisation programme implies stability. However, terminal cost is neglected in this application since the terminal state is nontrivial in general due to the unreachable setpoint. As avoiding calculation of this terminal state is a priority, the associated cost is neglected.

6.1.3 Controller Performance

In order to compare the MPC controller to commonly used controllers, two other controllers were tested with the model (4.13). The first is a fixed ANR controller, with ANR values of 0.8, 1.0 and 1.2. The second controller is a classical controller of PI type with the back-calculation anti-windup method enabled. This was hand tuned to provide a representative example of feedback control for SCR. The resulting gains were $k_p = 100$ and $k_i = 10$.

The controllers are benchmarked over a portion of the NEDC cycle data, comprising part city and part highway driving conditions. This is chosen to be the latter part of the cycle, since the catalyst is at operating temperature by this point and a comparison of the warm start performance of the controllers is the subject of this section. The available NEDC test data do not include the NO:NO_x ratio at the SCR inlet, so for the purposes of benchmarking this is set to be 0.5, which emulates the ideal conditions at the SCR inlet when a DOC is used to balance the NO and NO_2 concentrations. The temperature data for the monolith were generated by estimating the heat capacity of the monolith using the pre- and post-SCR gas temperatures in the test cycle data. For this testing, a warm-start sequence was chosen with an initial monolith temperature of 200°C. The cycle emissions and temperature data are shown in Figure 6.1.

The results of all three controllers, including three different ANR target values, are summarised in Table 6.1, for a catalyst initially containing no adsorbed ammonia and a maximum ammonia injection quantity of 250ppm. As expected, the best NO_x reductions typically accompany the highest ammonia slip. The fixed-ANR controllers achieve very low ammonia slip, due to their filling up the catalyst very slowly, but as a consequence do not achieve optimal NO_x reduction. The PI controller, as expected, has a mean ammonia

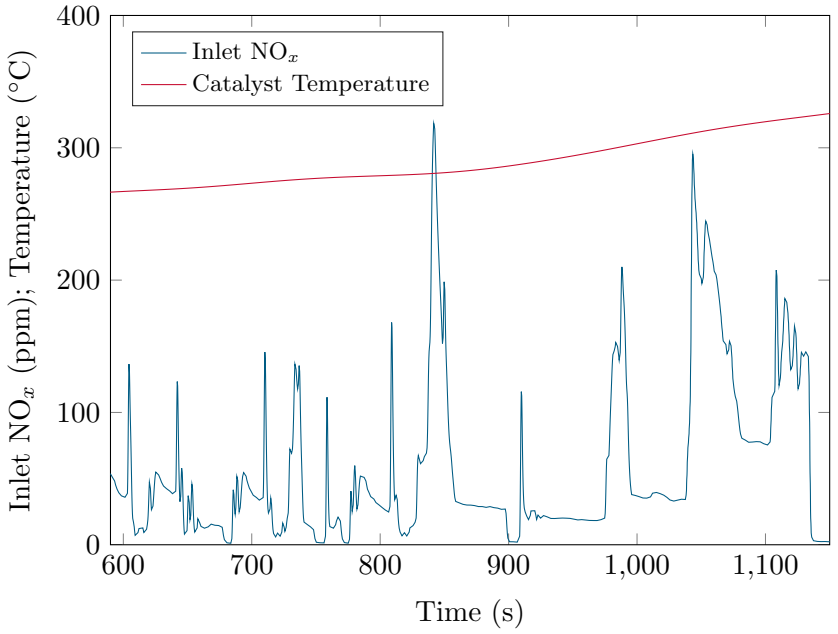
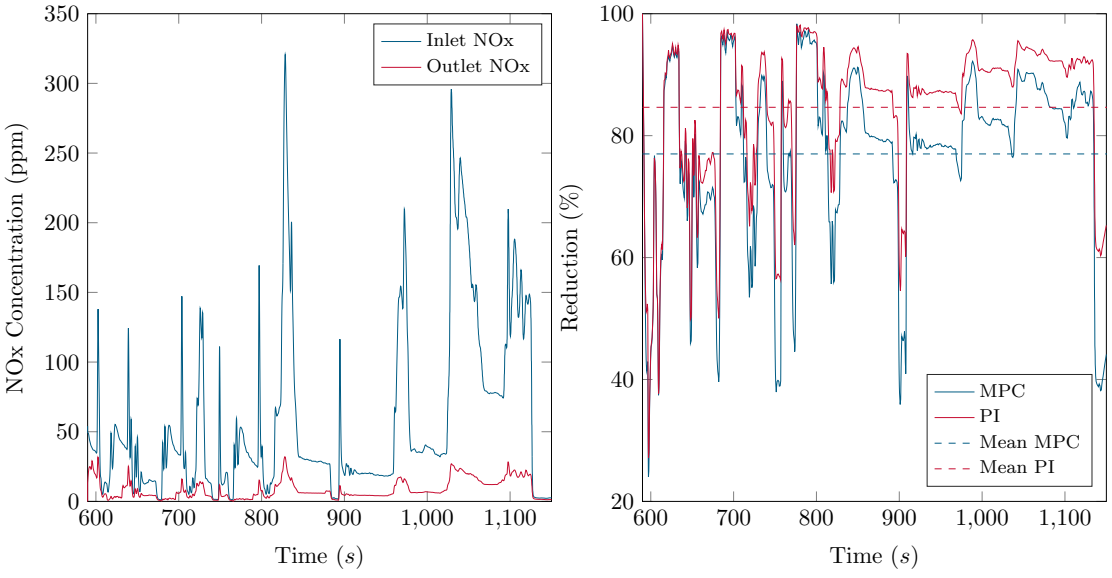


FIGURE 6.1: Portion of NEDC cycle used for controller benchmarking

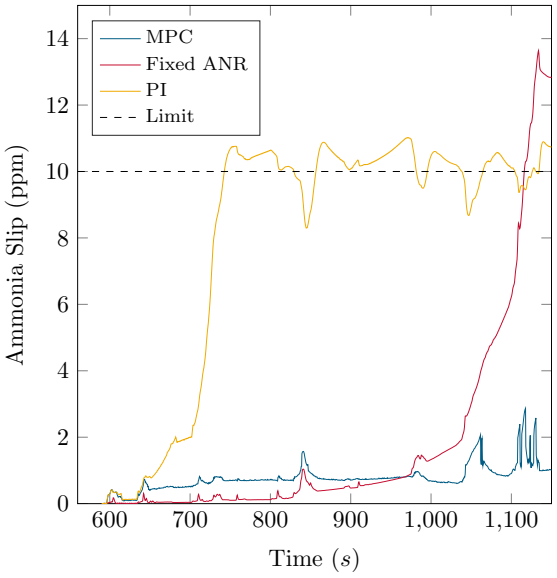
slip very close to its target of 10ppm, indicating correct controller tuning. The MPC controller finds an excellent compromise, with only 0.1ppm greater ammonia slip than the best fixed-ANR controller, but achieving 9% better NO_x reduction. The optimisation and tuning of weights has shown that trying to improve NO_x emissions even further would result in significant additional ammonia slip. Herein lies the benefit of a predictive controller to SCR – in periods where the chemical reactions are limited by temperature or by availability of NO_x rather than the amount of ammonia available, the predictive controller foresees ammonia slip and can react to minimise or eliminate it altogether. The MPC controller also achieves the additional objective of minimising consumption of AUS, beaten only by the lowest ANR controller whose NO_x conversion performance is significantly worse. Therefore, the predictive controller achieves a satisfactory balance between NO_x conversion and ammonia slip whilst additionally minimising ammonia usage, instead of trying to improve NO_x conversion performance by one or two percent at the expense of higher ammonia emissions.

The NO_x concentrations at the catalyst inlet and outlet during the operation of the MPC controller and the corresponding NO_x conversion efficiency are shown in Figures 6.2a and 6.2b respectively. The initial conversion efficiency is low since it is only adsorbed ammonia that takes part in the NO_x reactions, yet the catalyst is initially empty. The efficiency increases once the controller has injected enough ammonia to coat the catalyst. It is noteworthy that during the steady state periods of engine operation, NO_x reduction is around $80-90\%$, while the worst NO_x conversion efficiency occurs when the inlet NO_x suddenly varies and there is too much or too little ammonia on the catalyst surface. The fixed-ANR and PI controllers produce similar NO_x reduction results, as can be seen from Table 6.1; it is much more instructive to examine the ammonia slip and injected ammonia

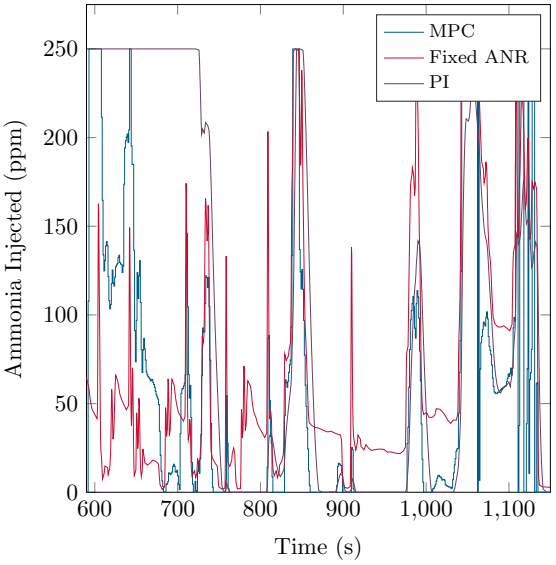


(a) Input and output NO_x

(b) Reduction in tailpipe NO_x



(c) Tailpipe ammonia slip



(d) Ammonia injection quantity

FIGURE 6.2: Comparison of fixed-ANR, PI and MPC controllers over the test cycle.

(which acts as the control input), as shown in Figures 6.2c and 6.2d respectively. In these plots, the fixed-ANR controller has a target ANR of 1.2 which was found to give the best performance over the portion of the NEDC cycle, as shown in Table 6.1. This is due to the fact that the catalyst was initially empty of stored ammonia ($\theta_0 = 0$), such that the excess ammonia is used in building up stored ammonia on the catalyst substrate.

At the start of the simulation, the ammonia slip is low since none has yet been injected and there is none stored in the catalyst. This causes the PI controller to inject the maximum ammonia during the first 150 seconds, which contributes to its high resulting ammonia slip in the rest of the cycle. The fixed-ANR controller does not model the

storage of ammonia in the catalyst, so does not act any differently at the beginning of the cycle even though the catalyst is empty and it could safely inject more ammonia. Towards the end of the cycle, it violates the ammonia slip limit quite considerably and may continue to do so on a longer test cycle. The predictive controller, on the other hand, captures the SCR trait that an atypically high injection is required at the beginning of the cycle such that the ammonia coverage ratio θ rises enough above 0 for the SCR reactions to take place. It therefore injects much more ammonia than the fixed-ANR controller, but not so much as the PI controller (which causes undesirable ammonia slip). These results are comparable to those in [32], but a linear predictive controller, whose internal model is very cheap to execute and is therefore more amenable to embedded implementation in a vehicle environment, has been employed here.

6.1.4 Concluding Remarks

In this section a linear economic MPC controller was developed to control the ammonia injection of an SCR catalyst for NO_x reduction of internal combustion engine exhausts. The linear model was derived from a calibrated nonlinear model around an operating point typical of the NEDC cycle data on which it was benchmarked. The model performs well in this region of state space, providing excellent SCR performance on the cycle data. The benefits of predictive control to SCR are clear. The economic formulation allows the controller to succinctly capture the essence of the SCR control problem in finding the balance between maximising NO_x reduction while minimising ammonia slip and ammonia consumption – an objective which is difficult to formulate in a classical control framework. Indeed, the feedback PI controller has difficulty dealing with such a plant, due to the inherent transport delays and lags in the system. The SCR plant is affected by disturbances, including exhaust gas volumetric flow rate, temperature, and NO_x concentrations, some of which can be measured and utilised by the predictive controller for feedforward compensation. Again, this is difficult to achieve under the framework of classical control. The ANR map controllers currently used in industry require all these variables to be accounted for in the feedforward ANR controller if its performance is to be acceptable. The maps therefore require storage space and time-consuming extensive offline calibration. In contrast, the predictive controller requires tuning that is both less time consuming and more intuitive.

6.2 Nonlinear Predictive Control of SCR

Although it has been demonstrated in Section 6.1 that linear MPC performs well when applied to SCR, the nonlinearities in the model (4.13), particularly those of the Arrhenius temperature relations, are not well captured by the linear MPC controller due to the restriction of a linear internal model. The case in which this deficiency becomes most

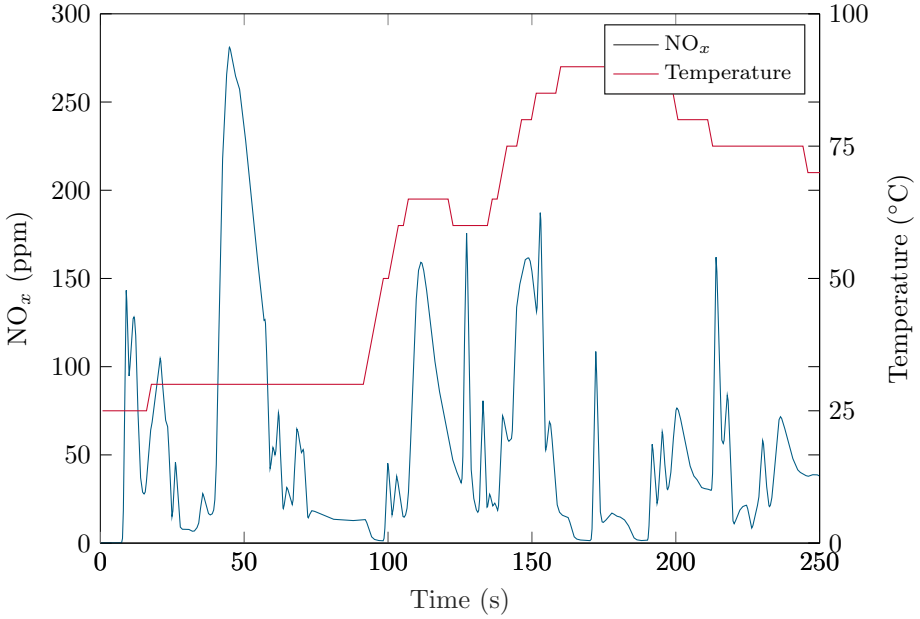


FIGURE 6.3: Inlet NO_x and SCR temperature during the cold start portion of the NEDC cycle

noticeable is in low temperature regions, which, as demonstrated in Chapter 3, is of great importance in SCR control. A nonlinear MPC (NMPC) controller was designed in order to compare the two approaches. The `iclocs` software was to used to build the controller [103], which implements the direct transcription method described in Subsection 2.4.4. The resulting NLP (2.76) is solved using the `ipopt` utility [101].

The NMPC controller was designed with the same cost function and weights as the linear MPC controller, with the stage cost $l(\cdot,\cdot)$ in (2.65) of the form given in the linear MPC design (6.2) and Q is as given in (6.3). Again, the terminal cost $E(\cdot)$ was discarded since calculation of a nontrivial terminal set is computationally expensive. The input and state constraints were as described in (6.4). The prediction horizon was chosen to be $t_f = 10$ seconds such that the fast storage dynamics (4.11)-(4.12) are captured, and a sufficient period of the slower dynamics are also captured. A sampling time of $t_s = 0.5$ seconds was found to be the maximum before controller performance began to degrade significantly. This fixes $N = 20$ in (2.73).

The nonlinear MPC controller developed in this section is tested against the linear MPC controller from Section 6.1 over two portions of the NEDC test cycle. The first is the hot start cycle used in Section 6.1, whilst the second is an earlier part of the same data: specifically, the first 250 seconds, since the catalyst is significantly below typical operating temperature during this period. The NO_x inlet to the SCR and the monolith temperature are shown in Figure 6.3.

The results for both controllers on both cycle segments are displayed in Table 6.2. The hot cycle data in the first two rows shows that the NMPC and linear MPC performance

is comparable when the catalyst is at operating temperature, since the linearised model is obtained around the operating temperature for linear MPC. The NMPC controller achieves marginally better NO_x reduction at the expense of slightly higher ammonia slip; this can be attributed to the higher fidelity of the nonlinear model permitting operation slightly closer to the ammonia slip constraint. However, the differences between the controllers at this temperature are minimal.

TABLE 6.2: Controller comparison over hot and cold start sequences

Cycle	Controller	NO_x Reduction (%)	NH_3 slip (ppm)
Hot	Linear	83.18	1.79
Hot	Nonlinear	86.45	3.13
Cold	Linear	8.74	13.28
Cold	Nonlinear	7.14	0.91

Conversely, the performance of the two controllers differs significantly when tested using the cold start cycle data depicted in Figure 6.3. The results are shown in the last two rows of Table 6.2. The injected ammonia and resulting ammonia slip for the two controllers are shown in Figure 6.4 and Figure 6.5 respectively. As expected, the NO_x conversion is relatively low, since the low catalyst temperature permits very little reduction to take place. Figure 6.6 corroborates this by showing almost identical tailpipe NO_x despite the profoundly different ammonia injection profiles in Figure 6.4. This demonstrates that the reactions are temperature limited in this cold start region. Even so, it is noteworthy that some conversion is possible even in these conditions.

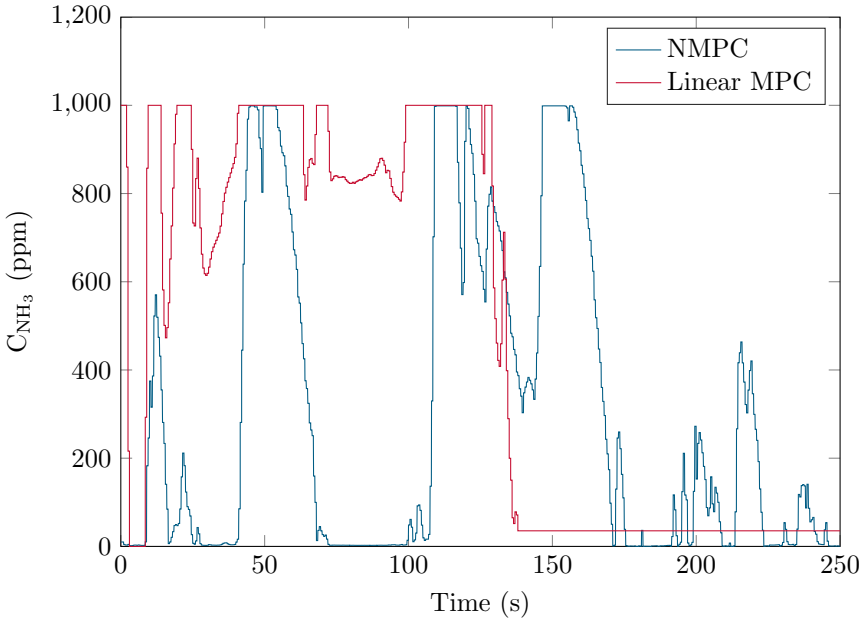


FIGURE 6.4: Cold start control inputs from the linear and nonlinear MPC controllers

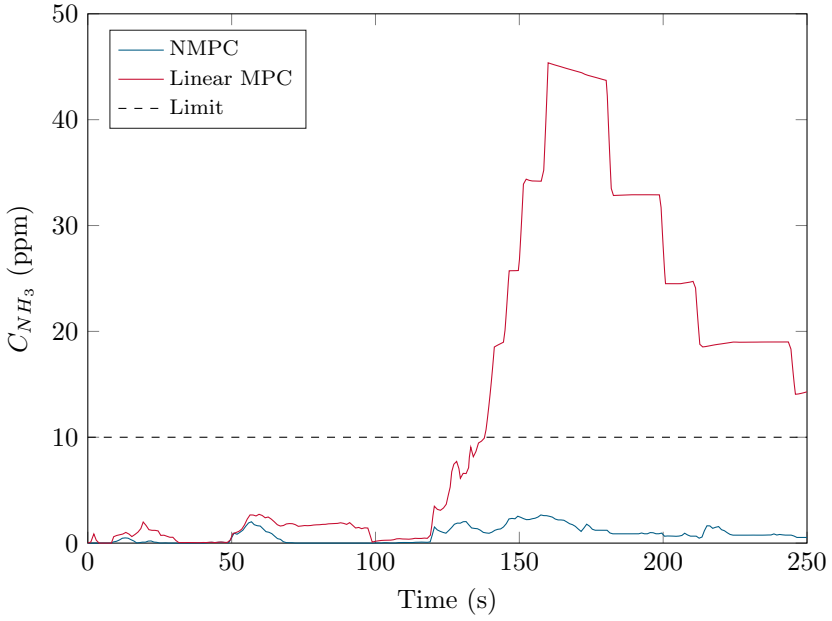


FIGURE 6.5: Cold start ammonia slip when controlled by the linear and nonlinear MPC controllers

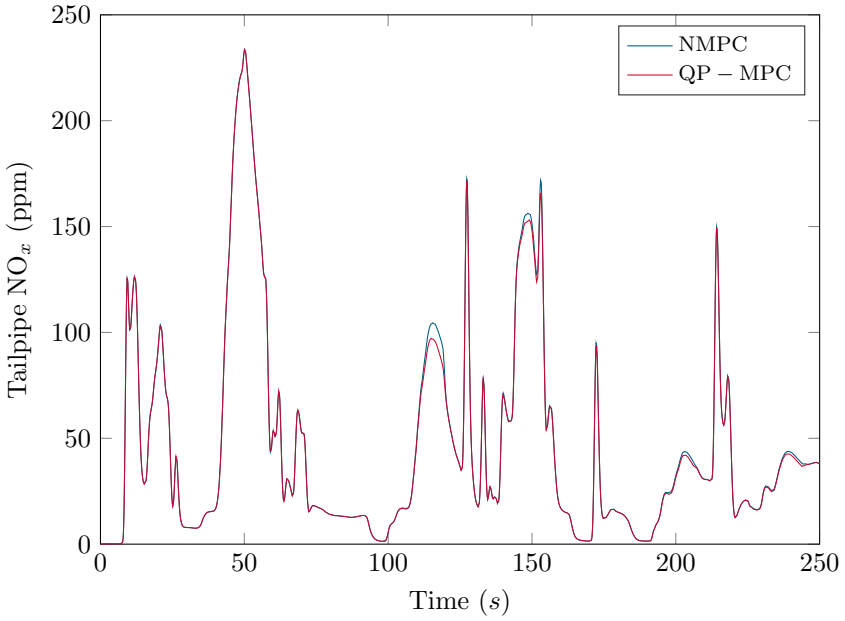


FIGURE 6.6: Cold start tailpipe NO_x when controlled by the linear and nonlinear MPC controllers (note: these are almost indistinguishable)

It is clear that the linear MPC controller is incapable of respecting the ammonia slip limit, with an average of over 13ppm and a maximum around 46ppm. Conversely, the NMPC controller achieves the required balance between NO_x conversion and ammonia slip, even in the face of adverse temperature conditions. During cold start, the limiting factor to NO_x conversion is the catalyst temperature. Since the linear MPC controller

model is linearised around operating temperature, the model over-predicts conversion of NO_x, leading to the high ammonia slip. Thus, the nonlinear temperature dependence of the reaction rates on temperature (4.4) is crucial to the model used by the controller, but is only captured well in the locality of the temperature at which the linear MPC model is linearised.

6.2.1 Computational Feasibility

Over the test cycle data in Figure 6.3, the maximum and average time to solve the NLP to a tolerance of 10^{-8} were 0.598 and 0.340 seconds respectively. These figures are generated on an INTEL Core i7 desktop computer.

The average time to solution is below the sampling time of 0.5 seconds, implying that real time implementation is feasible for this plant. However, the worst case exceeds the available time, so some work is required to remedy this. ICLOCS does not prioritise speed, so some improvement would be possible by carrying out the transcription process in a compiled language and using the relevant IPOPT interface. Additionally, different methods for solving the NLP may produce faster solutions. The benefits brought about via use of NMPC instead of linear MPC are principally due to the former accounting for the highly nonlinear temperature dependence of the reactions, which has a significant effect on the controller action. A simpler model of the SCR that still takes into account this temperature dependence is possible to decrease the computation effort. For example, a LPV model parameterised over temperature may provide performance acceptably close to the NMPC case whilst allowing solution using a linear MPC framework.

Desktop computers are not representative of vehicle hardware. To achieve these sampling times in a vehicle, other techniques will need to be exploited. One of the most promising is the opportunity for parallelisation of the solution algorithm, for example, solution of the NLP by direct multiple shooting in an RTI framework. For a vehicle platform, a multicore microprocessor or FPGA would permit such facilities to be made available.

6.2.2 Feasibility and Stability

It was noted at the beginning of this chapter that the terminal cost $E(\cdot)$ in the cost function (2.28) is typically included to give a guarantee of stability. However, in the case of the economic MPC formulation designed in Section 6.2, this is not feasible, since to guarantee stability, x_N must be constrained to be the optimal steady state x_s. However, calculation of x_s is nontrivial and would add to the computational demand of the controller.

Instead, some recent results are leveraged in order to show practical asymptotic stability of the closed loop in the absence of terminal conditions. Begin by collecting the vector of disturbances:

$$d = \begin{bmatrix} d_1 & d_2 & d_3 & d_4 \end{bmatrix}^T = \begin{bmatrix} C_{NO,in} & C_{NO_2,in} & S & T_{mon} \end{bmatrix}^T \tag{6.5}$$

Definition 6.1 (States, inputs and disturbances of the model)**.** The model (4.13) has finite dimensional state, input and disturbance spaces, defined as:

$$x \in \mathcal{X} \subseteq \mathbb{R}^{n_x}_{\geq 0} \tag{6.6}$$

$$u \in \mathcal{U} \subseteq \mathbb{R}^{n_u}_{\geq 0} \tag{6.7}$$

$$d \in \mathcal{D} \subseteq \mathbb{R}^{n_d}_{\geq 0} \tag{6.8}$$

where n_x, n_u and n_d are the numbers of states, inputs and disturbances respectively.

The rates r_l are (in some cases) products of state variables, and the Arrhenius factor (4.4) is an exponential function of temperature. As such, it is clear that the SCR plant model dynamics are highly nonlinear. This model (4.13) is the same as that used in [151], which in turn is a modification of that constructed in [33] using ideal reactor principles.

Remark 6.2. For any $d \in \mathcal{D}$, if $u(t) = 0$, then $x_s = \begin{bmatrix} d_1 & d_2 & 0 & 0 \end{bmatrix}^T$ is the only equilibrium of the system with region of attraction $\mathcal{X}$.

The model conforms to a continuous time nonlinear description:

$$\begin{aligned} &\dot{x} = f\big(x, u, d\big), \qquad x(0) = x_0 \\ &x \in \check{\mathcal{X}} \subseteq \mathcal{X} \subseteq \mathbb{R}^{n_x} \\ &u \in \check{\mathcal{U}} \subseteq \mathcal{U} \subseteq \mathbb{R}^{n_u} \\ &d \in \mathcal{D} \subseteq \mathbb{R}^{n_d} \end{aligned} \tag{6.9}$$

where $f : \mathcal{X} \times \mathcal{U} \times \mathcal{D} \to \mathcal{X}$. Further, x is the state vector, x_0 is the initial state, u is a vector of controlled inputs and d is a vector of disturbances. With $\check{\mathcal{X}}$ and $\check{\mathcal{U}}$ are denoted the sets of feasible states and inputs.

The terminal cost $E(\cdot)$ was neglected from (6.2) since computation of the terminal state x_s for a particular $d \in \mathcal{D}$ is nontrivial. Whilst allowing the constraint be that x_N resides in some terminal set $\mathcal{X}_N \equiv \check{\mathcal{X}}$ is possible, this would not adequately address stability concerns, which are therefore discussed in the sequel. Note that θ (x_4) is not weighted in (6.2) as there is no preferred ammonia storage fraction, only the concentrations of pollutants emitted by the reactor.

The optimisation process to find $u^\star$ is repeated at each sampling time in a receding horizon setting. This raises important questions about the resulting optimisation programme, including its feasibility and uniqueness of its solutions. Furthermore, a guarantee of asymptotic stability of the closed loop is required.

Feasibility

Assumption 6.2.1. The model (4.13) incorporates some disturbances $d \in \mathcal{D}$. The system is analysed for a constant value of d since it is held constant during the prediction horizon of the MPC scheme. Hence, define a new system $\dot{x} = f_d(x, u)$.

Definition 6.3 (Equilibrium set)**.** The equilibrium set is the set of equilibrium pairs $\mathcal{Z} := \{(x, u) : 0 = f_d(x, u)\}$. The projection of $\mathcal{Z}$ onto $\mathcal{X}$ is denoted $\mathcal{X}_Z$.

From the stage cost (6.2) it is clear that the controller will try and force the plant towards $\|x\|_i = 0, i = \{1, 2, 3\}$. However, this setpoint is unreachable except in the trivial case $d_1 = d_2 = 0$. In the general case therefore, $\min_{(x,u)\in\mathcal{Z}} l(x, u) \neq 0$. Using Remark 6.2, a property of the plant (4.13) on which the following analysis will be based can be demonstrated.

Lemma 6.4. *For the system (4.13), for all* $u \in \mathcal{U}$, $\exists\, x_s : 0 = f_d(x_s, u)$.

This implies that any $u \in \mathcal{U}$ will stabilise the plant at some corresponding steady state $x_s \in \mathcal{X}$, which may not be constraint admissible (i.e. not necessarily in $\check{\mathcal{X}}$). In other words, the plant has an infinite number of feasible equilibria and it is the function of the controller to repeatedly target that which minimises the cost whilst obeying constraints. It is assumed henceforth that $f_d(\cdot, \cdot)$ and $l(\cdot, \cdot)$ are Lipschitz on x and u. Whilst it is not in general necessary to compute $\mathcal{Z}$, an illustration of the possible equilibria and their associated costs is given for one particular value of d in Figure 6.7. Given that the optimal value of the cost function (2.65) depends on the disturbances d, make the following assumption:

Assumption 6.2.2. For every $d \in \mathcal{D}$ there exists a non-empty set

$$\mathcal{Z}^\star \subseteq \mathcal{Z} := \{(x^\star, u^\star) : l(x^\star, u^\star) \leq l(x, u)\ \forall (x, u) \in \mathcal{Z}\}, \tag{6.10}$$

The implication of Assumption 6.2.2 is that any set of disturbances yields at least one cost-optimal equilibrium pair, but that this pair is not necessarily unique. The non-uniqueness can be proven by examination of the stage cost (6.2); an increase in one variable can be countered by a decrease in another, resulting in the same cost for a different state.

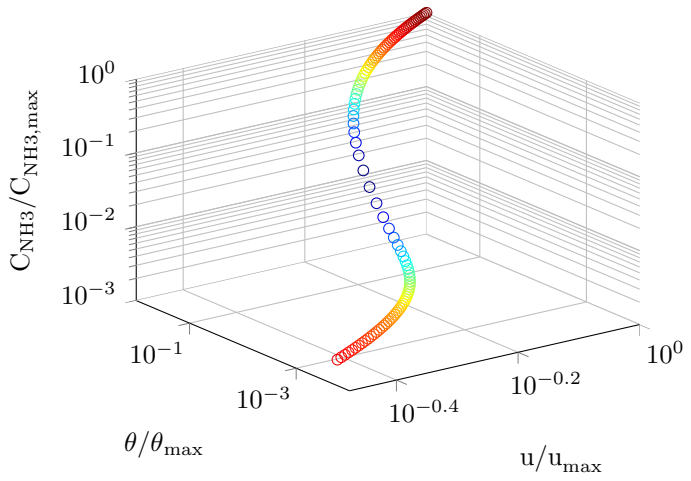


FIGURE 6.7: Section of $\mathcal{Z}$ in the $u-\theta-C_{NH_3}$ dimensions. Each point is an equilibrium coloured by its normalised log-cost $\ln l(x,u)$.

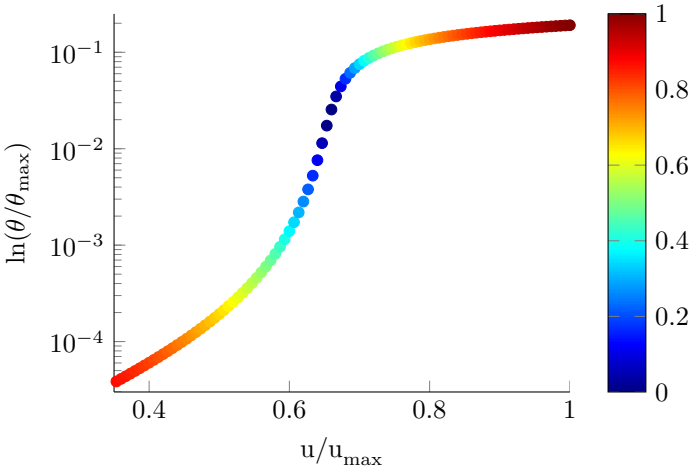


FIGURE 6.8: Section of the $u-\theta$ plane in $\mathcal{Z}$. Each point is an equilibrium coloured by its normalised log-cost $\ln l(x,u)$.

Remark 6.5. The dynamics of the SCR system (4.13) guarantee that the map $u_s \to \theta_s$ is one-to-one, as demonstrated by Figure 6.8 monotonically increasing. As such, there cannot exist two pairs of $(x_s, u_s) \in \mathcal{Z}$ with identical costs. Therefore, $\mathcal{Z}^\star$ is a single point in $\mathcal{Z}$-space for this system.

Remark 6.6. For the system (4.13), ensuring that $u_s \in \breve{\mathcal{U}}$ can be achieved by ensuring that $\sup_u \breve{\mathcal{U}} \geq |d_1| + |d_2|$. This is evidenced by Remark 2.1.

Definition 6.7 (Controlled invariant subset)**.** Define the set of initial states for which there is at least one input trajectory that keeps $x(t) \in \breve{\mathcal{X}} \subseteq \mathcal{X}$ for all t:

$$\widetilde{\mathcal{X}} := \left\{ x \in \breve{\mathcal{X}} : \exists\, u(t) \in \mathcal{U} : x_u(t) \in \breve{\mathcal{X}}\ \forall\, t \in [0, \infty) \right\} \tag{6.11}$$

Feasibility of the closed loop trajectories requires that $x(t) \in \check{\mathcal{X}}\ \forall t$. However, due to input constraints, it is not guaranteed that $\check{\mathcal{X}}$ is a controlled invariant set. To see this, consider the dynamics of C_{NH_3} in (4.13) under $u(t) = \min u \in \check{\mathcal{U}} = 0$ when $\theta = 1$, i.e. the catalyst ammonia storage is maximum:

$$\begin{aligned} \dot{C}_{NH_3} &= r_{des} - SC_{NH_3} \\ &= Vk_{des} \exp\left(\frac{-E_{des}}{RT}(1-\gamma)\right) - SC_{NH_3} \end{aligned} \tag{6.12}$$

Letting $\bar{C}_{NH_3} = \max_{x \in \check{\mathcal{X}}} C_{NH_3}$,

$$\begin{aligned} 0 < \dot{C}_{NH_3} \\ < Vk_{des} \exp\left(\frac{-E_{des}}{RT}(1-\gamma)\right) - S\bar{C}_{NH_3} \end{aligned} \tag{6.13}$$

which, whilst dependent on parameters and $\bar{C}_{NH_3}$, is typically true at high temperatures. $\dot{C}_{NH_3} > 0$ will cause the state $x(t)$ to immediately leave $\check{\mathcal{X}}$. $\square$

The following relationship therefore holds:

Lemma 6.8. *Recalling the definition of $\mathcal{X}_Z$, it is clear that*

$$\mathcal{X}_Z \subseteq \widetilde{\mathcal{X}} \tag{6.14}$$

This can be proved by considering the definition of $\mathcal{Z}$, since for any x_0 in $\mathcal{X}_Z$, a feasible input is $u(t) = u_s\ \forall\, t \in [0, \infty)$. The implication is thus that $\mathcal{Z} \not\equiv \emptyset \Rightarrow \mathcal{X}_Z \not\equiv \emptyset \Rightarrow \widetilde{\mathcal{X}} \not\equiv \emptyset$. As a result, safely make the following assumption:

Assumption 6.2.3. Our analysis is restricted to cases where $x_0 \in \widetilde{\mathcal{X}}$.

By implication, existence of a feasible trajectory $u(t)$ implies existence of a feasible $u_0^\star$ as the first step of this trajectory. It has so far been proven that a feasible input trajectory $u(t)$ exists providing $x_0 \in \widetilde{\mathcal{X}}$. It is desirable to comment on the objective function of the optimisation programme (2.66). To do this, assure the following two lemmata:

Lemma 6.9. *For the system (4.13), the map $u(t) \to x_u(t)$ is unique. There exists no other $u(t)$ which will result in the same state trajectory.*

Proof of this lemma derives from Lipschitz assumptions on $f_d(\cdot,\cdot)$ and $l(\cdot,\cdot)$ and by considering remark 6.5. By implication, then, the sequence of stage costs over a trajectory and thus the value function (2.65) are unique for a given x_0 and $u(t)$. The main result of this section can therefore be stated:

Theorem 6.10. *For an initial state $x_0 \in \widetilde{\mathcal{X}}$, the minimiser $u_0^\star = \mu_N(x_0) = \arg\min_u V_N(x_0)$ exists, is constraint admissible and is unique.*

Lemma 6.8 proves existence of at least one feasible input at time t_0, whilst lemma 6.9 proves that there exists an input trajectory whose cost is strictly better than all other input trajectories. □

Stability

Receding horizon optimal control often employs a terminal cost $E(\cdot)$, as in (2.65), to certify closed-loop stability. However, determining $E(\cdot)$ will require the knowledge of x_s which is not always possible. As mentioned earlier in eMPC, stabilizing the exact equilibrium x_s is not always the target. Whilst we could let the constraint be that x_N resides in some terminal set $\mathcal{X}_N \equiv \check{\mathcal{X}}$, this would not guarantee convergence to x_s. Therefore, we will discard terminal conditions and costs in this application. Instead, we will show that for a given disturbance d, in addition to existence of a feasible u, $x(t)$ will converge asymptotically to a small set around x_s, thereby proving practical closed-loop asymptotic stability [75]. In order to achieve this, the terminal conditions are replaced by an approximate turnpike property [74, 75].

Definition 6.11 (Approximate turnpike property)**.** The solutions of the OCP exhibit an approximate turnpike property with respect to some steady state $x_s \in \check{\mathcal{X}}$ if there exists some function $\nu \in \mathcal{K}_\infty : (0, \infty) \to [0, \infty)$ such that for all $x_0 \in \check{\mathcal{X}}$ and all $T > 0$ we have:

$$\mu[\Theta_{\epsilon,T}] < \nu(\epsilon) \;\; \forall \epsilon > 0, \tag{6.15}$$

where $\mu[\cdot]$ is the Lebesgue measure on the real line and

$$\Theta_{\epsilon,T} = \{t \in [0, T] : \|x_{u_N^\star} - x_s\| > \epsilon\}. \tag{6.16}$$

In this SCR control problem, the continuity of $f(\cdot)$ and $l(\cdot, \cdot)$ is given by the dynamics of the system (4.13) and the choice of l, respectively. The two functions are also locally Lipschitz around x_s and u_s. In this problem which can be seen as a case of tracking MPC, following the approach used in [74], we use the optimal value function

$$V_N(x) = \inf_{u_N \in \mathcal{U}^N} \sum_{k=0}^{N-1} l(x_{u_N}^k, u_N^k) \tag{6.17}$$

as a Lyapunov function, with $u_N \in \mathcal{U}_N$, and closed-loop stability can thereby be proven. However, in our case, recalling that we have defined Q as given in (6.3, the stage cost (6.2) is positive semidefinite, and thus $V_N(x)$ is also positive semidefinite, and therefore cannot be used to infer asymptotic stability of the closed loop. Instead, we will formulate an alternative value function that is ensured to be positive definite, thus ensuring closed-loop asymptotic stability of the system.

Definition 6.12 (Strict dissipativity)**.** [152] A system is strictly dissipative with respect to some supply rate $s(x,u)$ if for some $\alpha \in \mathcal{K}_\infty$,

$$\lambda(x_0) := \sup_{N,\mu(x_0)} \sum_{k=0}^{N-1} -\Big(s(x_k,u_k) - \alpha(\|x_k - x_s\|)\Big) < \infty, \tag{6.18}$$

for all x_0.

Taking a supply rate $s(x,u) = l(x,u) - l(x_s,u_s)$, we require

$$\sup_{N,u\in\mathcal{U}} \sum_{k=0}^{N-1} \Big(\alpha(\|x_k - x_s\|) + l(x_s,u_s) - l(x_k,u_k)\Big) < \infty \tag{6.19}$$

for all x_0. From the model (4.13) and the restrictions on the reaction rates $r_{i,j}$ in Remark 4.3.1, we have $x_{u_N} \in \mathcal{X}^N$ and $u_N \in \mathcal{U}^N$. The stage cost (6.2) is bounded on $\mathcal{X}$ and $\mathcal{U}$. The value of α in (6.19) is bounded on $\mathcal{X}$ by definition. As a result, we conclude that (6.19) is satisfied.

As shown in [69], strict dissipativity with respect to some steady-state x_s is sufficient to prove existence of an approximate turnpike in the solution trajectories at x_s under an additional reachability assumption.

The continuity of $f(\cdot)$ and $l(\cdot,\cdot)$ implies continuity of $V_N(\cdot)$, such that the turnpike property is sufficient to demonstrate convergence of the closed-loop system to a neighbourhood of the optimal equilibrium x_s in the absence of terminal conditions.

We have seen that $V_N(x)$ as in (6.17) may be a Lyapunov function. Instead, we proceed by taking $\lambda(x)$ as a storage function, and defining a *rotated* stage cost [71]

$$L(x_k,u_k) := l(x_k,u_k) + \lambda(x_k) - \lambda(x_{k+1}) - l(x_s,u_s) \tag{6.20}$$

and its associated optimisation programme:

$$\widetilde{V}_N(x_0) = \inf_{u_N\in\mathcal{U}^N} \sum_{k=0}^{N-1} L(x^k_{u_N}, u^k_N). \tag{6.21}$$

Remark 6.13. By inspection, $L(x_s,u_s) = 0$ and $\min\limits_{x,u} L(x,u) = \min\limits_{x,u\in\mathcal{Z}} l(x,u)$. By corollary, the optimisation programme (6.21) admits the same solution as (2.76). As a result, the optimisation programme (6.21) is subject to the same constraints as (2.76).

Using the rotated stage cost, the strict dissipativity property can be phrased as

$$L(x,u) \geq \alpha(\|x - x_s\|). \tag{6.22}$$

We now proceed in a Lyapunov-type direction and examine the behaviour of the rotated value function $\widetilde{V}_N$ from one sampling time to the next. The following lemmata are the consequences of (6.22) and the turnpike property.

Lemma 6.14. *By (6.22) and the existence of a turnpike, then*

$$\widetilde{V}_N(x_1) \leq \widetilde{V}_N(x_0) + \delta(N). \tag{6.23}$$

Proof. Due to the turnpike property, we have (6.15). Moreover, (6.20) and the local Lipschitzity of f and l, and hence λ, around x_s and u_s, there exists a sufficiently small $\delta > 0$ such that the rotated stage cost at the terminal state is bounded by

$$L(x_N, u_N) \leq \delta(N) \leq \delta \tag{6.24}$$

Furthermore, using (6.22), we have

$$\begin{aligned} \widetilde{V}_N(x_1) &= \widetilde{V}_N(x_0) + L(x_N, u_N) - L(x, u) \\ &\leq \widetilde{V}_N(x_0) - \alpha(\|x - x_s\|) + \delta(N) \\ &\leq \widetilde{V}_N(x_0) + \delta(N) \\ &\leq \widetilde{V}_N(x_0) + \delta \end{aligned} \tag{6.25}$$

which completes the proof of the lemma. □

Lemma 6.14 shows that due to turnpike property, $\widetilde{V}_N$ is a practical Lyapunov function for the closed-loop system with the nonlinear eMPC controller.

Lemma 6.15. *$\widetilde{V}_N(x)$ is positive definite with respect to x_s.*

Proof.

$$\begin{aligned} \widetilde{V}_N(x_0) &= \sum_{k=0}^{N-1} L(x_k, u_k) \\ &= \sum_{k=0}^{N-1} \Big[l(x_k, u_k) + \lambda(x_k) - \lambda(x_{k+1}) \Big] \\ &= \sum_{k=0}^{N-1} l(x_k, u_k) + \lambda(x_0) - \lambda(x_N) \\ &\approx V_N(x_0) + \lambda(x_0) \end{aligned} \tag{6.26}$$

□

which ensures that $\widetilde{V}_N(x)$ is positive definite with respect to x_s by ensuring $\lambda(x) > 0$ as in (6.18). We now state the main result of this section.

Theorem 6.16. *The rotated value function $\widetilde{V}_N(\cdot)$ is strictly decreasing along nontrivial system trajectories and is positive definite with respect to x_s, with the offset $\delta(N)$, as proved by Lemmata 6.14 and 6.15 respectively. Therefore, $\widetilde{V}_N(\cdot)$ is a practical Lyapunov function for the economic MPC closed loop. Its existence implies that x_s is a practically asymptotically stable equilibrium of the closed-loop with region of attraction $\mathcal{X}$.*

The level sets of the rotated value function $\widetilde{V}_N(x_0)$ are shown for a particular value of d in Figure 6.9.

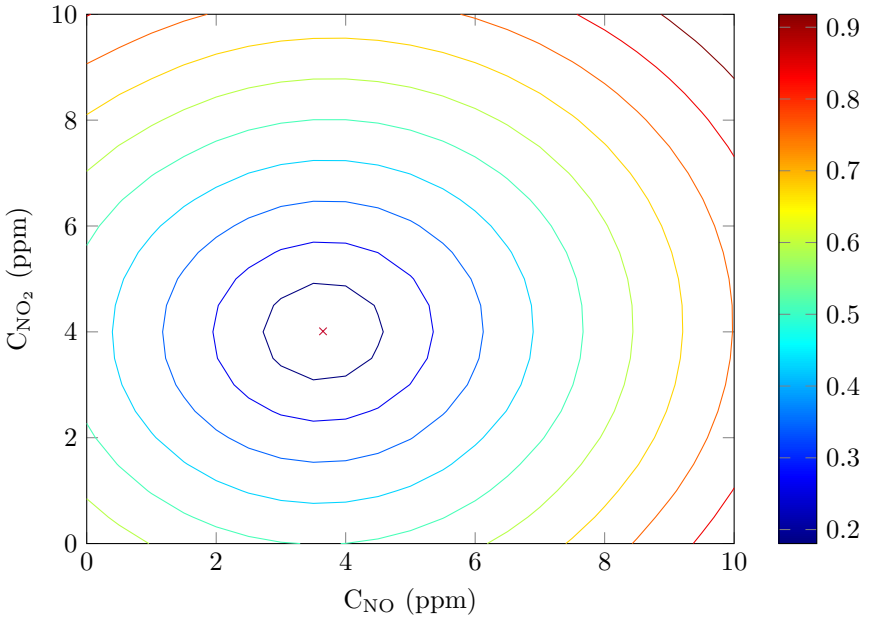


FIGURE 6.9: Level sets of the rotated value function $\widetilde{V}_N(x_0)$ with normalised cost for various x_0 using a particular value of d. The optimal equilibrium (marked with a red cross) has $C_{NO} = 3.65$ppm and $C_{NO_2} = 4.01$ppm.

Numerical Validation

The MPC formulation developed is applied to a calibrated SCR model as used in Section 6.2. We first show that the closed-loop system exhibits the properties described previously, including the turnpike property. We compare the performance of the eMPC controller to a tracking-MPC controller, demonstrating that the economic formulation yields improved closed-loop performance as measured by the closed-loop cost. The test scenario is a stabilisation of the reactor at the optimal steady-state with inlet NO and NO_2 each at 50ppm:

$$d = \left[2.2 \times 10^{-3}, \quad 2.2 \times 10^{-3}\right]^T, \tag{6.27}$$

and the initial state $x_0 = 0$. The optimal steady-state for this case is

$$x_s^\star = \left[1.6 \times 10^{-4}, \quad 1.7 \times 10^{-4}, \quad 1.5 \times 10^{-4}, \quad 0.21\right]^T. \tag{6.28}$$

The stage cost $l(\cdot,\cdot)$ in (2.65) was chosen in an economic least-squares form as in (6.2) in order to minimise the total quantity of tailpipe emissions.

Typically, a prediction horizon of around $t_f = 30$ seconds is required such that the fast storage dynamics (4.11)-(4.12) are captured, and a sufficient period of the slower dynamics (θ) are also captured. A sampling time of $\Delta t = 0.5$ seconds was found to be the maximum tolerable for adequate rejection of fast disturbances. This fixes $N = 60$ in (2.73). It was determined that the weights $Q_w = 1,\ w = \{1,2,3\}$ provided good controller performance. A constraint of 10 ppm was added for outlet ammonia concentration, and an actuator constraint of $0 \leq u \leq 1000$ ppm was chosen for the ammonia dosage.

The economic stage cost (6.2) was compared with a tracking-MPC stage cost

$$l_t(x,u) = \|x - x_s^\star\|_Q^2, \tag{6.29}$$

where x_s is the optimal equilibrium (6.28) for the value of d as defined by (6.27), as demonstrated in Figure 6.9, and Q is as defined in (6.3). It is demonstrated by Figure 6.10b that the economic stage cost (6.2) achieves consistently better closed-loop performance than tracking MPC for all prediction horizons N. Additionally, it does not require the computation of $x_s^\star$ at each time step, which is a nontrivial function of d. Using an economic stage cost, the turnpike property is visualised for various N in Figure 6.10a. Here, it can be seen that the closed-loop trajectories remain close to $x_s^\star$ in the period after the effects of initial conditions have passed, until one or two time steps before the end of the horizon, when the trajectories leave the turnpike towards a locally lower-cost solution.

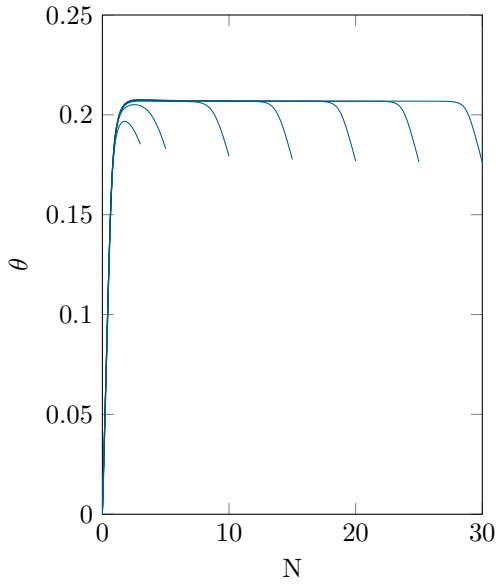
The convergence of NO_x and ammonia to their respective equilibria under MPC closed-loop is shown for both economic and tracking MPC formulations in Figure 6.10c. The results are summarised in Table 6.3. An intuitive indication of the cause of the better performance of economic MPC can be found by examination of the ammonia storage θ during transient operation. As shown in Figure 6.10d, economic MPC allows storage to rise above the steady-state value $x_{s,4}^\star$ in the transient, leading to improved NO_x conversion in this period, which results in lower stage costs and thus a lower value function as shown in Figure 6.10b. Despite achieving better NO_x performance and higher ammonia storage, the eMPC controller uses less urea solution. The outlet ammonia is necessarily higher in the case of eMPC, yet the constraint is respected.

6.2.3 Concluding Remarks

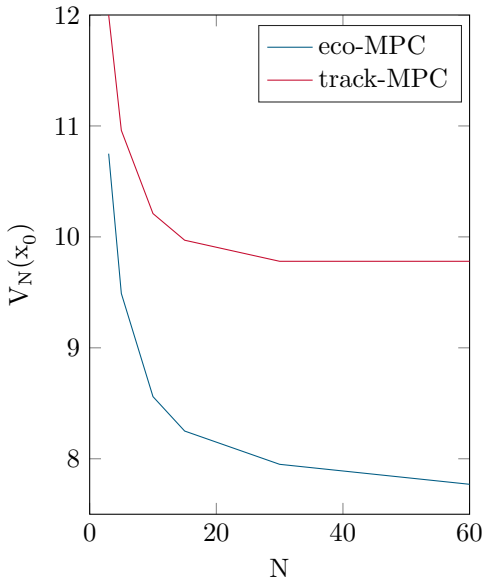
A nonlinear economic MPC controller has been formulated to deal directly with the constraints of operation of an SCR catalyst used in a diesel engine. It has been proven possible, unlike previous formulations of this problem, to avoid the calculation of an

TABLE 6.3: Comparison of eMPC and tracking MPC.

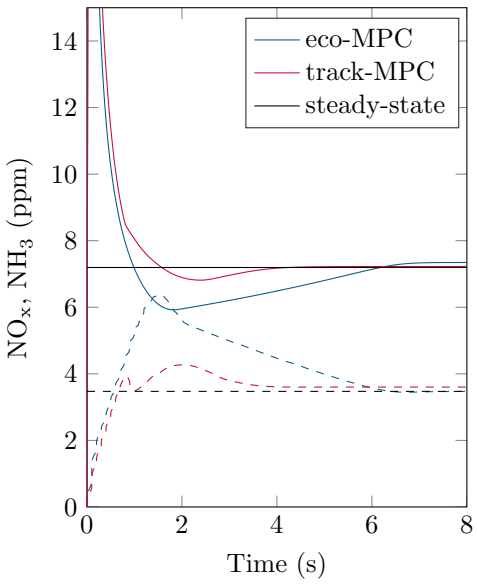
Statistic	Improvement vs. tracking MPC (%)
NO_x emissions	2.5
NH_3 emissions	−13.1
Urea consumption	1.6



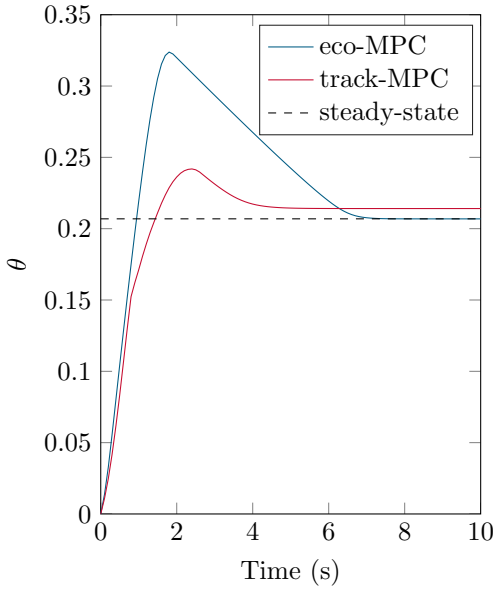
(a) Turnpike properties of θ with an economic stage cost (6.2) for $N = \{3, 5, 10, 15, 20, 25, 30\}$



(b) Dimensionless value function $V_N(x_0)$ for economic and tracking MPC for $x_0 = 0$ against prediction horizon N



(c) Closed loop convergence of NO_x (solid) and ammonia (dashed) concentrations under economic and tracking MPC for $x_0 = 0$



(d) Closed loop convergence of θ for economic and tracking MPC for $x_0 = 0$

FIGURE 6.10: Comparison of economic and tracking MPC

ammonia storage reference directly, but instead model the complete catalyst and thus constrain and weight the tailpipe gas concentrations. Such a formulation is feasible and stable, such that implementations of this controller are suitable for use in an automotive environment. The improvements are particularly profound in cold start operation, which accounts for a significant proportion of NO_x emissions, especially in urban environments. The control objectives for SCR are relatively complex: rather than a regulation problem, the controller must continually make a tradeoff between ammonia slip and NO_x reduction. The NMPC controller can meet these objectives, as well as obeying constraints, during the cold start period as well as during normal operating conditions. This is in contrast to industry standard techniques of disabling the controller or limiting ammonia injection, resulting in suboptimal emissions reduction and ad-hoc switching between control laws. The NMPC controller makes the optimal tradeoff between NO_x reduction and ammonia slip and can do so for all operating conditions.

The possibilities of NMPC implementation for SCR in a vehicle environment are promising. The relatively long sampling period tolerable for SCR control provides adequate opportunity to improve NO_x reduction via such control schemes. The disadvantage is a more computationally expensive controller, however, it has been estimated that the controller is feasible given such sampling periods.

6.3 Thermal Management of SCR

Due to minimal interaction with the engine fuelling control loop, electric catalyst heaters are an excellent candidate for retrofitting to existing diesel vehicles fitted with SCR whose NO_x conversion performance is lower than ideal. A catalyst heater is modelled as a gas heater attached to the front face of the catalyst monolith, as shown in Figure 4.6. It is modelled as being close-coupled with the catalyst in order to maximise the heat transfer between the heater elements and the catalyst monolith. The monolith is not heated directly as this leads to localised hot spots which can damage the catalyst – instead, the exhaust gas is heated before it enters the SCR brick.

6.3.1 Adding Thermal Dynamics to SCR Model

A pair of thermocouples in the SCR catalyst enclosure monitor the gas temperatures before and after the monolith. Monitoring the SCR monolith temperature directly is difficult due to sensor embedding, so this is inferred via the gas temperature sensors. Since reaction rates are functions of the monolith temperature T_{mon}, a simple first order thermal model of the relationship between the upstream gas temperature $T_{exh,in}$ and the monolith temperature is utilised. The heat release from the chemical kinetics is neglected in this model. A parameter k_M models the convectional heat transfer from

the exhaust gases to the monolith, whilst a second parameter k_A models heat loss from the monolith to the environment at temperature T_{amb}, which is

$$\dot{T}_{mon} = k_M(T_{exh,in} + \Delta T_{heater} - T_{mon}) - k_A(T_{mon} - T_{amb}). \tag{6.30}$$

where ΔT_{heater} is given in (4.14). This new state is added to the existing CSTR-based model, and lump NO and NO_2 into a single NO_x parameter. Thus, the reactor model can be written as

$$\begin{bmatrix} \dot{C}_{NO_x} \\ \dot{\phi} \\ \dot{C}_{NH_3} \\ \dot{T}_{mon} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \frac{1}{V} & 0 \\ 0 & \frac{k_M}{c_p \dot{m}} \end{bmatrix} u(t) + \begin{bmatrix} \frac{1}{V} & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & k_M \end{bmatrix} d(t) + \begin{bmatrix} \phi R_{oxi} - C_{NO_x}(\phi R_{red} + F) \\ R_{ads} C_{NH_3} \Phi - \phi(R_{ads} C_{NH_3} + R_{des} + R_{red} C_{NO_x} + R_{oxi}) \\ \phi R_{des} - C_{NH_3}(R_{ads}(\Phi - \phi) + F) \\ -k_M T_{mon} - k_A(T_{mon} - T_{amb}) \end{bmatrix}. \tag{6.31}$$

where u is a vector of control inputs:

$$u = \begin{bmatrix} NH_{3,inj}, & P_{heater} \end{bmatrix}^T, \tag{6.32}$$

and d is a disturbance vector:

$$d = \begin{bmatrix} NO_{x,in}, & T_{exh,in} \end{bmatrix}^T. \tag{6.33}$$

6.3.2 Combined Chemical and Thermal Control

The MPC controller was configured with a 20 second prediction horizon with a timestep of 0.5 seconds. The upper limit for ammonia injection was 1×10^{-5} kmol/s and the upper limit for the catalyst heater was 3.6 kilowatts, which are both representative of real hardware. The lower limit was set to 1×10^{-12} for both inputs for numerical stability reasons. The weighting matrices Q and R in (2.65) were configured to ensure a tailpipe ammonia slip of less than 10 ppm. No weight was put on the ammonia storage ϕ.

The calibrated catalyst model developed in Section 4.3 was used as the plant model, and the predictive controller was benchmarked over the first 250 seconds of a cold-start NRTC test cycle, where the SCR catalyst is soaked at 50°C before commencing.

Enabling the catalyst heater in the MPC controller raises the temperature of the catalyst monolith significantly faster than relying solely on heat in the exhaust gas stream. Since a cold start is occurring, the controller opts to enable the heater at maximum power (3.6 kW) for the first part of the cycle, as shown in Figure 6.11a. The result of this is raising of the SCR monolith temperature significantly faster than when the catalyst heater is not used, as shown in Figure 6.11b.

Higher catalyst temperatures increase rates of reaction in SCR, which limit NO_x conversion in this temperature region. The result is that, as shown in Figure 6.11c, the quantity of NO_x emitted from the tailpipe is significantly reduced. In fact, over this portion of the cold-start test cycle, the cumulative NO_x emitted is reduced by 39%. In addition, the ammonia slip is reduced when the catalyst heater is enabled, due to the higher monolith temperatures allowing more ammonia to take part in SCR reactions. This is demonstrated in Figure 6.11d, wherein it is shown that over this period, the cumulative ammonia slip is reduced by 10.7%.

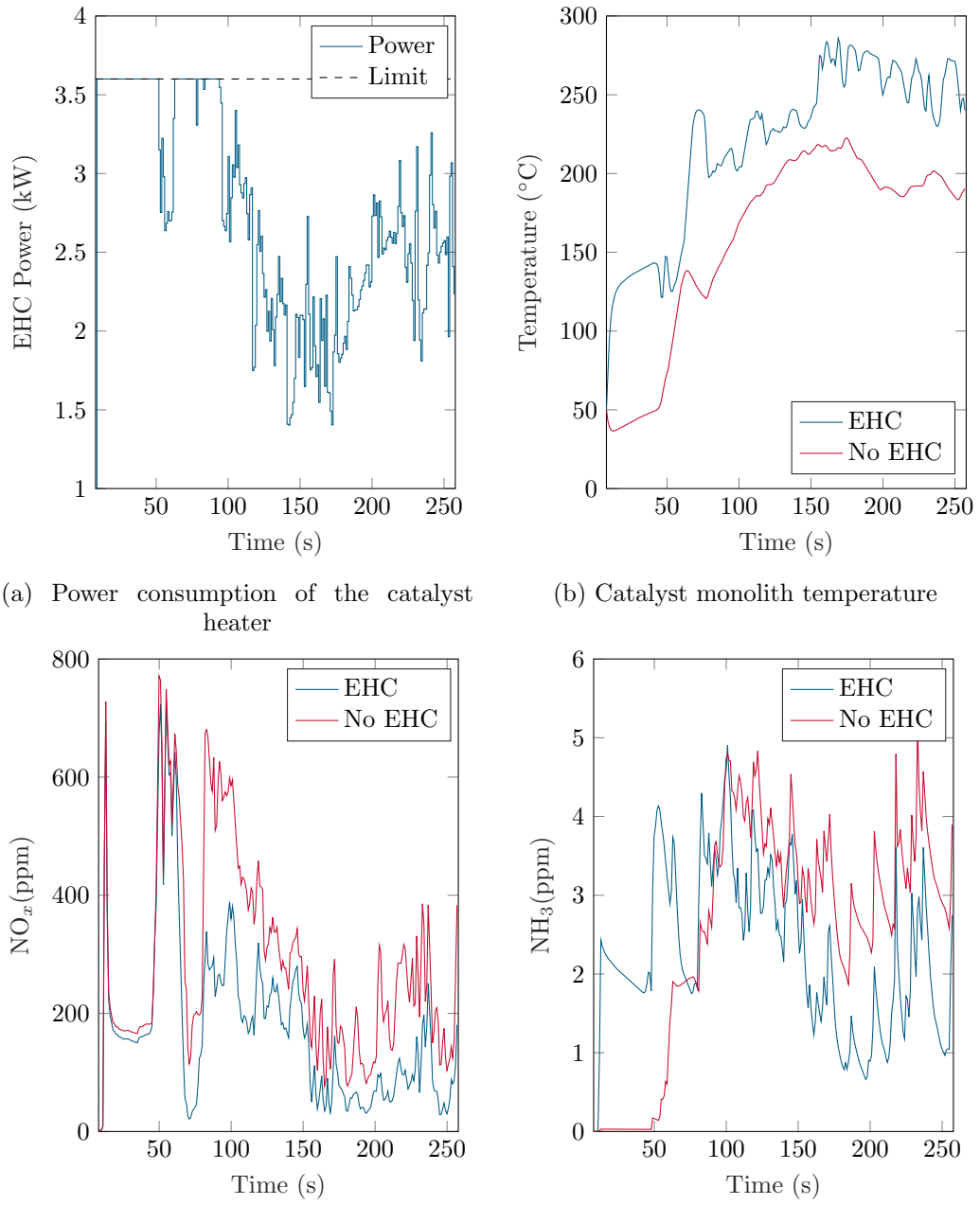
It is clear that increased conversion of NO_x implies increased ammonia consumption: this is true in this case, with the catalyst heater enabled the system consumes 56.8% more ammonia.

6.4 Conclusions

Both linear and nonlinear MPC have proven promising for the control of an SCR catalyst. Linear MPC performs very well when the state is near the linearisation point of the model. However, further away from this, it is shown that the linear controller cannot respect constraints. The main factor that causes this discrepancy was determined to be temperature, since reaction rates increase exponentially with temperature. Furthermore, it is precisely the temperature effects that are causing deviation of SCR from its optimal operation, as discovered via the PEMS equipment in Chapter 3. As such, nonlinear MPC was introduced to overcome this difficultly. A cold-start test cycle was used to demonstrate that nonlinear MPC can maintain constraint satisfaction and optimal SCR operation throughout the temperature range.

Lastly, in this chapter, it was shown that the thermal issues of SCR as found in Chapter 3 can be overcome to some extent by adding an actuator that allows the controller to force heating of the catalyst where required. The additional degree of freedom granted by doing this allows the control designer to fine-tune the fundamental tradeoff between NO_x and CO_2 emissions.

The principal disadvantage of nonlinear MPC is the increased computational cost. It is possible that linear parameter varying (LPV) models or a gain-scheduled linear controller could be satisfactory approximations to achieve the wide operational region of nonlinear



(a) Power consumption of the catalyst heater

(b) Catalyst monolith temperature

(c) Tailpipe NO_x

(d) Ammonia slip

FIGURE 6.11: Comparison of monolith temperature along with tailpipe NO_x and ammonia with and without the catalyst heater enabled, as well as the power drawn by the heater when enabled.

MPC whilst retaining the lower computation cost of the linear variant. Computational costs will be examined further in Chapter 9, including methods to overcome these in order to permit real-time implementation of the nonlinear MPC controller for SCR.

Chapter 7

State Estimation for SCR

In Chapters 5 and 6 it was assumed that full state measurement of the plant was available, whereas in reality the quantity of ammonia stored in the catalyst cannot be measured, but must instead be estimated online. In order to design a controller for a real implementation, this chapter explores various methods of state estimation that can be applied to the SCR system. The problem mentioned in Section 2.3 of cross-sensitivity of NO_x sensors to ammonia in a manner that is both temperature dependent and nonlinear is also addressed.

7.1 Parallelised Particle Filtering

In Chapter 2, the existing literature on state estimation for SCR was examined. This fundamentally comprises the estimation of three key parameters. Firstly, the level of ammonia stored in the catalyst, otherwise known as the *ammonia coverage ratio* is required for any practical SCR controller. For the model based controllers developed in this thesis, the online knowledge of this parameter is essential, because it appears as a state in the control oriented models developed in Chapter 4. Secondly, tailpipe ammonia concentration is typically estimated, since, as discussed in Chapter 2, commercial ammonia sensors of sufficient accuracy and precision for SCR control are not yet widely available. Thirdly, the cross-sensitivity of NO_x sensors to ammonia is to be estimated in order that reliable NO_x measurements can be made. It was noted that this factor varies nonlinearly with temperature and across production batches, so design-time calibration is infeasible in this case. It was argued that the approaches used in the literature mainly fall into the classical observer category. The model based observers were based on the Kalman Filter and Extended Kalman Filter (EKF).

In this chapter, two new methods for state estimation for SCR are presented. These are the Unscented Kalman Filter and the Particle Filter. Both of these should provide

better estimation performance than the KF and EKF since the model is highly nonlinear, and these methods do not rely on local linearisation of the model. Their computational demands are also investigated in order to determine the feasibility of these estimators for online applications.

The particle filter can be viewed as a generalisation of the UKF. The UKF chooses a minimal set of points ($2L$, where L is the state dimension) deterministically around the current state, and passes them through the nonlinear system dynamics before reconstructing the statistics, assuming a Gaussian distribution. Conversely, the particle filter is a sequential Monte Carlo (SMC) technique which does not assume any knowledge of the noise distributions. It chooses many (often 1000 or more) points stochastically, thereby further increasing estimation accuracy by better approximation of the posterior distribution [153], [154]. N_s particles are drawn from the initial state distribution $p(x_0)$, and are equally weighted with weights q_i such that

$$\sum_{i=0}^{N_s-1} q_i = 1. \tag{7.1}$$

Each particle is then propagated through the state equation using the known probability density function (pdf) of the state noise such that the prior particles are generated as

$$x_{k,i}^{-} = f(x_{k-1,i}^{+}, u_{k-1}, w_k) \quad \forall i \in 1 \dots N_s. \tag{7.2}$$

The weight of each particle is then updated based upon the likelihood that such a particle would generate the known measurement, $\tilde{q}_i = p(y_k|x_{k,i}^{-})$ and then $\tilde{q}_i, i \in [1, N_s]$ is scaled such that (7.1) holds:

$$q_i = \frac{\tilde{q}_i}{\sum_{j=1}^{N_s} \tilde{q}_j}. \tag{7.3}$$

The particles are now resampled based on the weights q_i in order to produce the *a posteriori* particle set $x_{k,i}^{+}$ which is an approximation of the desired posterior distribution $p(x_k|y_k)$. The resampling step is typically implemented by drawing a random sample from the uniform distribution $r \sim \mathcal{U}[0,1]$ for each particle, then choosing the M'th particle where M satisfies [154]

$$\sum_{j=0}^{M-1} q_j < r \leq \sum_{j=0}^{M} q_j, \tag{7.4}$$

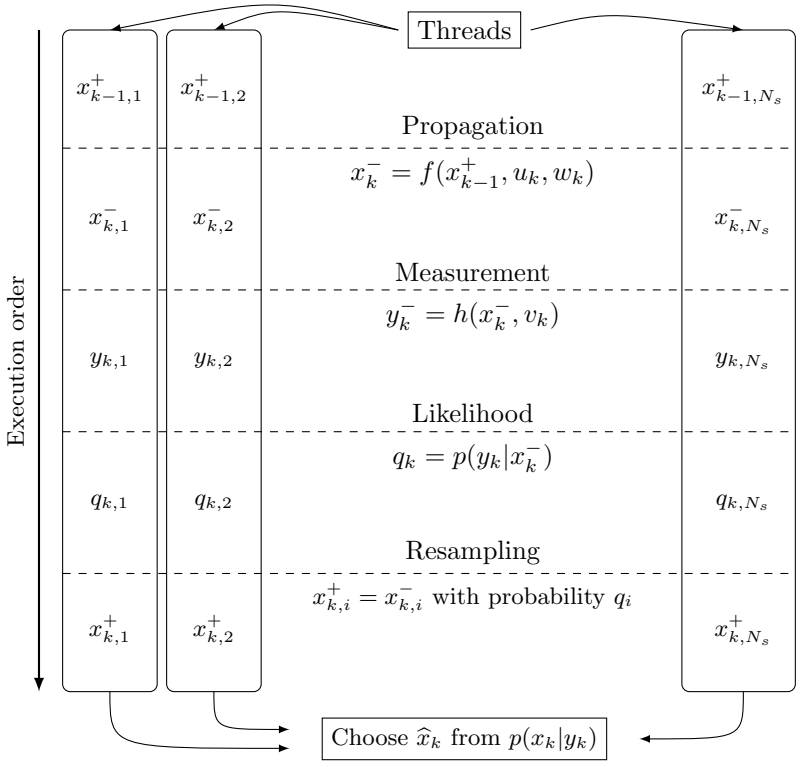


FIGURE 7.1: Parallelisation of the particle filter

where $q_0 = 0$ and $M \leq N_s$. The resampled particles along with their weights describe the posterior distribution, so any metric can be chosen for the state estimate. In the case of a unimodal distribution, it is common to take the mean of the posterior distribution:

$$\widehat{x}^+_k = \frac{1}{N_s} \sum_{l=1}^{N} x^+_{k,l}. \tag{7.5}$$

where $\widehat{x}_k$ is the estimate of x_k.

7.2 Parallelisation

Operation of the filter in real time is desirable for use in a control strategy. For an SCR control system that has a sample time of 10ms, a solution within 10% of this (1ms) is a suitable target, in order that enough time remains within the sampling period for sensor acquisition and control step calculation. With this in mind, one of the key benefits of the particle filter is its amenability to parallelisation. This benefit comes from the fact that each particle $x^{(i)}_k$ is independent, as shown in Figure 7.1.

Of the parallel architectures commonly available in a desktop computing environment, the graphics processing unit (GPU) seems particularly well suited to the particle filtering task, since it is capable of high degrees of parallelisation. GPU-based particle filtering

methods have been examined before with great success [155]. The compute unified device architecture (CUDA) by NVIDIA permits the writing of software for GPU applications. In this model, *kernels* are executed in parallel by threads, which in turn are grouped into *blocks*. The scheduler assigns blocks to the execution engines, which on the GPU are known as *streaming multiprocessors* (SMs). The SMs divide the blocks into *warps* which are executed in parallel according to a single instruction, multiple data (SIMD) model. The SMs do not execute code as quickly as a typical desktop processor; instead the speed increase comes from large-scale parallelism.

This is particularly well suited to the particle filter, since execution can be deeply parallelised, since each particle q_i can be treated individually (with one exception, the resampling stage, as discussed next). Firstly, the existing particles are propagated through the model equation f. Then, the measurement function h is applied to calculate their likelihoods $p(y_k|x_k^-)$, which are subsequently normalised. Finally, the resampling step creates a new set of particles for the next time step. This is shown schematically in Figure 7.1. With $N_s = 1000$ particles, the degree of parallelism that GPUs offer can be exploited to a great extent in this application.

Resampling

The q_j in the sums (7.4) require data from all particles, which is the source of significant speed decrease in a parallel implementation. Two novel options for countering this dependency were introduced by [156], known as *parallel resampling*: effecting the resampling process without computing the cumulative sum of weights. The threads shown in Figure 7.1 then become truly independent with the exception of accesses to single elements q_j. The resulting penalty is not as onerous, since the CUDA architecture provides areas of *shared memory* which can be exploited in this scenario since access is limited to read-only.

The algorithm chosen for this implementation is Metropolis sampling, which is founded upon Metropolis-Hastings algorithms for drawing samples from a distribution $P(x)$ which does not have an analytical form [154]. Metropolis-Hastings algorithms are a Markov chain Monte Carlo (MCMC) method which iterate a Markov chain which has the desired distribution $P(x)$ as its *stationary distribution* $\pi(x)$, meaning that as samples are drawn repeatedly, the distribution of those samples will tend towards the desired $P(x)$. The algorithm commences by assignment of an ancestor, denoted a, to each particle $a_i = i$, such that:

$$a = \begin{bmatrix} a_1, & a_2, & \ldots, & a_s \end{bmatrix} = \begin{bmatrix} 1, & 2, & \ldots, & N_s \end{bmatrix}. \tag{7.6}$$

A vector u of N_s random numbers are drawn from the uniform distribution $\mathcal{U}[0,1]$. A vector $\widetilde{a}$ of N_s random integers in the range $[1, N_s]$ denotes candidate ancestral particles. If

$$u_i \leq \frac{q_{\widetilde{a}_i}}{q_{a_i}}, \tag{7.7}$$

then the ancestor of particle i becomes the jth particle, $a_i = \widetilde{a}_j$. This procedure should be repeated until convergence is reached:

$$\|P^B(i,\cdot) - \pi(\cdot)\| \leq \epsilon, \tag{7.8}$$

where $\pi(\cdot)$ is the true stationary distribution, $P^B(i,\cdot)$ is the B^{th} iterate of the Metropolis resampling algorithm in [156], and $\epsilon > 0$ is chosen to be a small positive number. An upper bound on B is given by

$$B > B^\star \triangleq \frac{\log \epsilon}{\log(1-\beta)}, \tag{7.9}$$

where β is defined as in [156] by

$$\beta \triangleq \min_{i \in \{1,\ldots,N_s\}} \left(\frac{\frac{1}{N}\sum_{j=1}^{N_s} q_j}{q_i} \right) \geq \frac{1}{N} > 0. \tag{7.10}$$

7.3 Estimator Design

To generate the state estimator model, the following key assumptions are made:

Assumption 7.3.1. NO and NO_2 are lumped into a single NO_x state, since this is representative of the sensor characteristics.

Assumption 7.3.2. The only modelled reactions are (2.4) and (2.11).

Assumption 7.3.3. Mass transfer is assumed to be infinitely fast, such that reactions are either temperature or species limited.

The reaction rate applied to the single lumped reduction equation is given as

$$R_{red} = r_{red}\, C_{NO_x} \theta \tag{7.11}$$

where r_{red} is of the form (4.4). Using the CSTR model (4.7), the reactions are phrased as follows

$$\begin{aligned}\dot{C}_{NO_x} &= \frac{NO_{x,in}}{V} + \Phi R_{red} - \frac{F}{V} C_{NO_x} \\ \dot{C}_{NH_3} &= \frac{NH_{3,in}}{V} + \Phi(R_{des} - R_{ads}) - \frac{F}{V} C_{NH_3} \\ \dot{\theta} &= R_{ads} - R_{des} - R_{red}\end{aligned} \tag{7.12}$$

where $NH_{3,in}$ and $NO_{x,in}$ are the inlet molar flow rates in kmol/s of ammonia and NO_x respectively. As the dynamic of the ammonia concentration inside the reactor occurs much faster than the NO_x reduction and the ammonia storage reactions, make a quasi-steady state approximation

$$\dot{C}_{NH_3} = 0 \implies C_{NH_3} = \frac{r_{des}\,\theta\,\Phi + \dfrac{C_{NH_{3,in}}}{V}}{r_{ads}\,\Phi\,(1-\theta) + \dfrac{F}{V}}, \tag{7.13}$$

which allows calculation of the value of C_{NH_3} as required in (7.12). Excluding the dynamic equation for ammonia concentration, and augmenting a zero dynamic state to track k_{sens} (since no information on its dynamics exists), the model becomes

$$\begin{bmatrix}\dot{C}_{NO_x} \\ \dot{\theta} \\ \dot{k}_{sens}\end{bmatrix} = \begin{bmatrix} C_{NO_x}\left(r_{red}\,\theta\,\Phi + \dfrac{F}{V}\right) \\ r_{ads} C_{NH_3} - \theta(r_{ads}\,C_{NH_3} + r_{des} + r_{red}\,C_{NO_x}) \\ 0\end{bmatrix} \tag{7.14}$$

on which the estimation procedure is based. Asymptotic observability of the system is assumed (4.13) as defined in [157].

Forward Euler integration was used to discretise the model with a sample time of 10ms. As shown in Figure 7.2, the values of θ and k_{sens} are required by the ACU for control calculation. However, the sensor measurements are a function of states:

$$y_k = C_{NO_x,k} + k_{sens,k}\,C_{NH_3} + v_k, \tag{7.15}$$

where v_k is an independent zero-mean white Gaussian noise sequence with variance 10^{-18} kmol/m^3, which was determined from benchmarking of a production sensor. Computation of C_{NO_x} follows from the sensor output and $\widehat{k}_{sens}$.

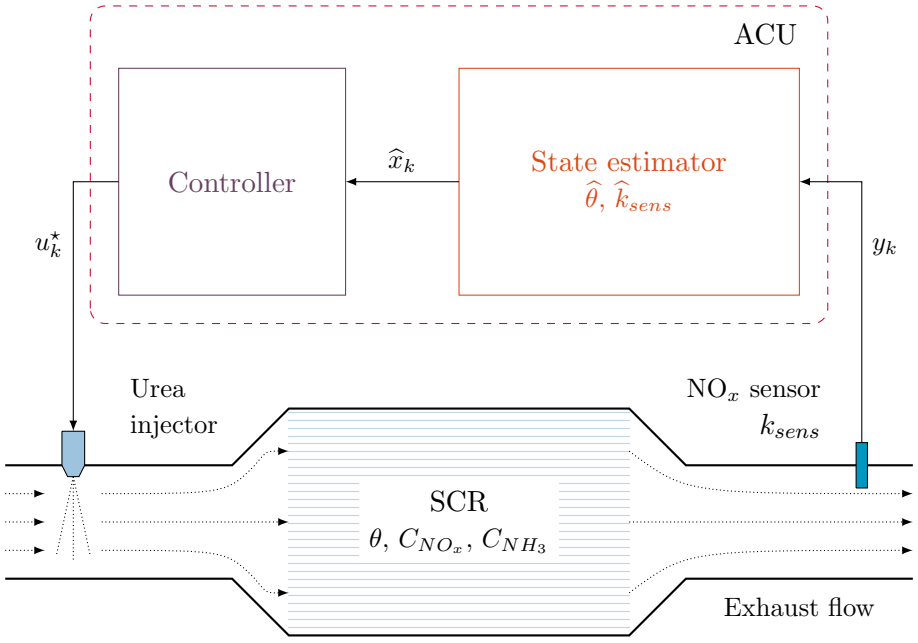


FIGURE 7.2: Closed loop SCR schematic with state estimator for the state estimate $\widehat{x}$ required by the controller.

7.4 Results

The particle filter was simulated and compared to an EKF and UKF using data from a diesel engine on an NEDC cycle. The Q matrix, as given in (2.15), were configured for the EKF and UKF as:

$$Q = \begin{bmatrix} 10^{-4} & 0 & 0 \\ 0 & 10^{-12} & 0 \\ 0 & 0 & 10^{-6} \end{bmatrix} \tag{7.16}$$

and with $R = 10^{-4}$. The inlet NO_x concentration was fixed at 117.6ppm, and the inlet ammonia concentration was varied as shown in Figure 7.3. The thermal mass of the sensor was assumed to be small, thus temperature dynamics of the sensor were neglected. This causes worst-case transients in k_{sens}; the inclusion of temperature dynamics in reality will have the effect of low-pass filtering the true value. The EKF and UKF were initialised with all states being 0. The particle filter was configured with $N_s = 1000$ particles and the initial distribution $p(x_0)$ as uniform over the 3 states:

$$\begin{aligned} x_{0,i}^{(1)} &\sim \mathcal{U}[0, 5 \times 10^{-7}] \\ x_{0,i}^{(2)} &\sim \mathcal{U}[0, 2 \times 10^{-3}] \\ x_{0,i}^{(3)} &\sim \mathcal{U}[0, 1] \end{aligned} \tag{7.17}$$

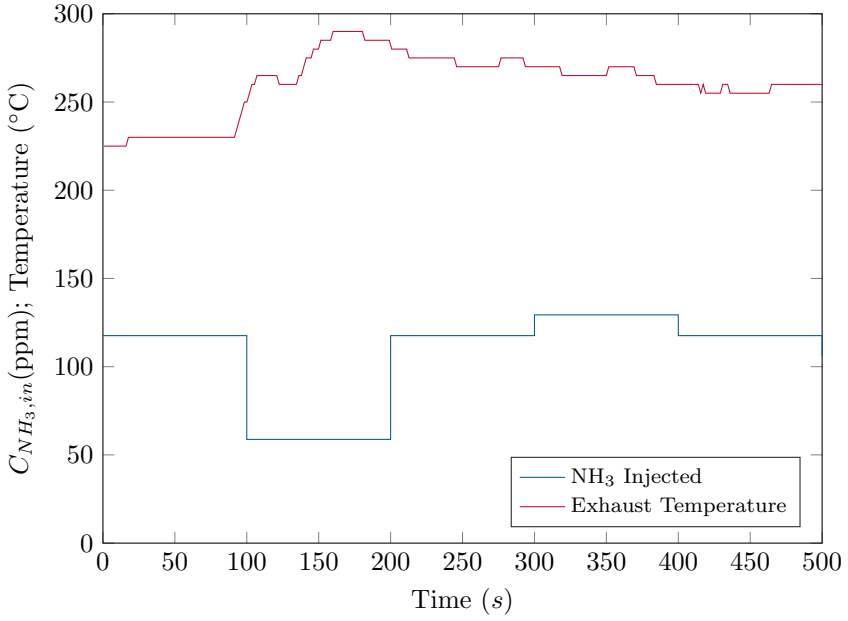


FIGURE 7.3: Ammonia injection concentration and exhaust temperature during the test cycle

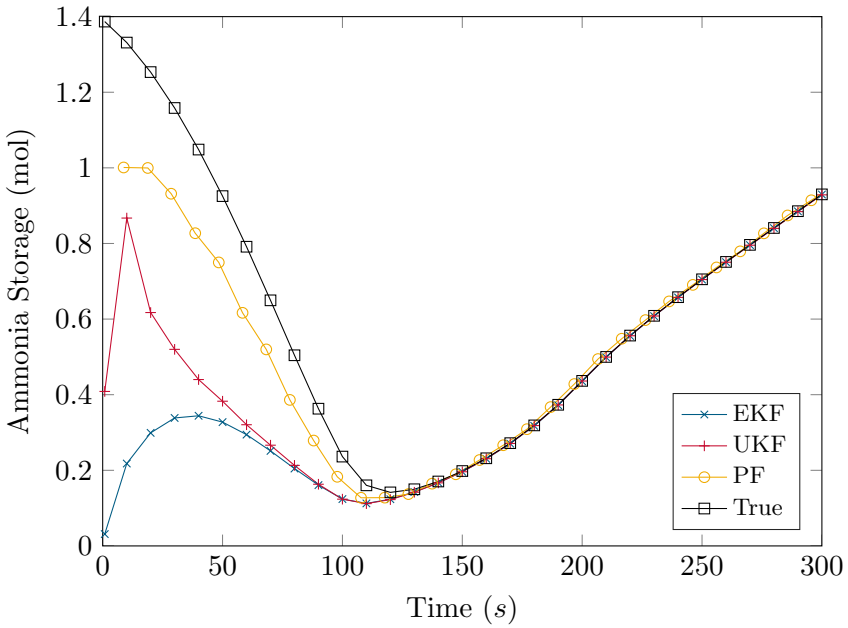


FIGURE 7.4: Convergence of estimated and true ammonia storage on the catalyst surface

in order to represent the lack of information as to the initial state but confine the particles to the feasible ranges of the state variables. It was found that $B = 3$ in (7.9) was the typical upper limit with $\epsilon = 10^{-3}$ in practice, and was thus fixed instead of calculating $B^\star$ at each iteration. The resulting convergence to the true ammonia storage and cross sensitivity factors are shown in Figures 7.4 and 7.5 respectively. The tracking performance of the filters is shown for the post-convergence period in Figure 7.6.

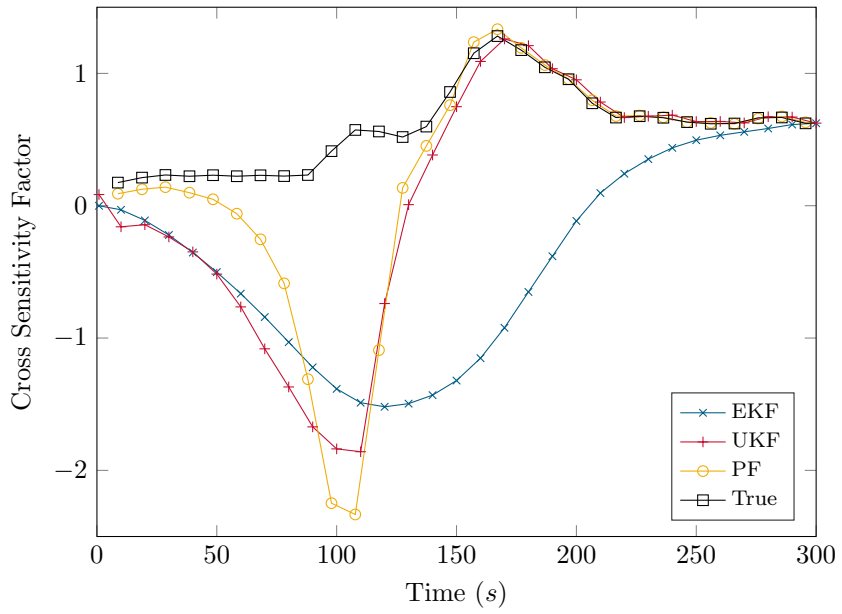


FIGURE 7.5: Estimated and true cross sensitivity factors in the convergence period

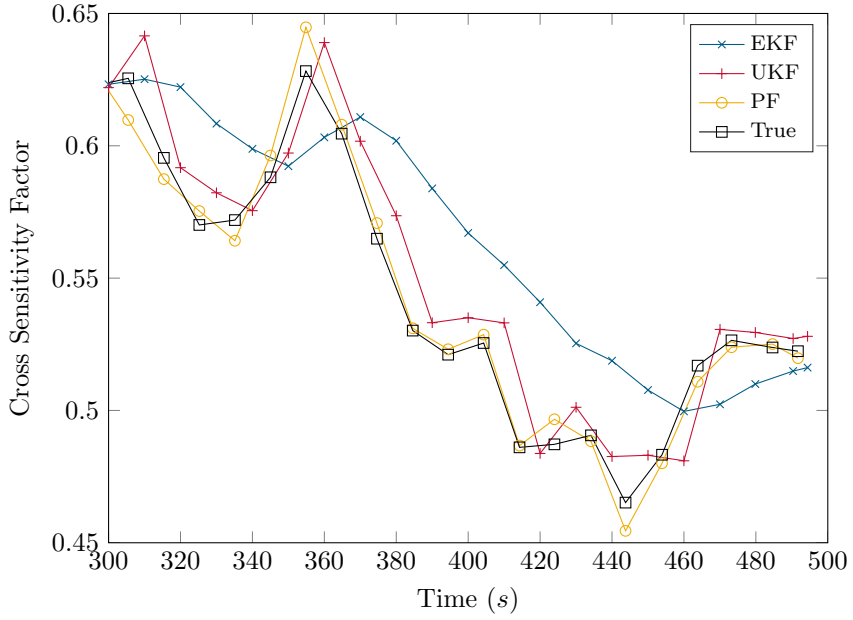


FIGURE 7.6: Tracking of cross sensitivity factor in the period after filter convergence

The particle filter demonstrates two distinct advantages. Firstly, the distribution of the initial particles according to (7.17) results in the best initial estimates during convergence, as shown in Figure 7.5. Secondly, the particle filter demonstrates the best tracking of the cross sensitivity factor during fast temperature transients, as in Figure 7.6. The UKF cannot make use of an initial distribution $p(x_0)$, thus state estimates are relatively poor until convergence. However, during the tracking period, the UKF provides performance significantly better than the EKF and approaching the particle filter. The divergence of k_{sens} in all filters in the region around 100s can be explained by the sensor output equation (7.15), since an overestimate of NO_x concentration can be countered by an underestimate in ammonia concentration, resulting in identical sensor output. Table 7.1 shows the mean solution times for each filter on an Intel Core i7 CPU, confirming that the EKF is computationally the cheapest, whilst the particle filter is the most expensive. The UKF provides an excellent middle ground, with computational time well within the sampling period.

TABLE 7.1: Average CPU solution time for the EKF, UKF and two particle filters

Filter	Solution time (ms)
EKF	0.17
UKF	0.33
PF ($N_s = 100$)	0.76
PF ($N_s = 1000$)	1.09

Comparative results for the particle filter on the CPU and GPU (NVIDIA Quadro 4000) are shown in Table 7.2. The root mean square (RMS) tracking error for k_{sens} for the period after convergence (200 second onwards, see Figure 7.5) is also shown as a function of the number of particles. The CPU implementation scales poorly with particle number, quickly exceeding the available 1ms, whilst GPU implementation exhibits more desirable scaling characteristics.

TABLE 7.2: Average particle filter execution times on the CPU and GPU

N_s	CPU (ms)	GPU (ms)	k_{sens} RMS Error
100	0.76	0.71	0.081
500	0.95	0.73	0.052
1000	1.09	0.77	0.036
10000	4.84	0.90	0.022

7.5 Conclusions

This chapter examined state estimation for SCR, both in order to estimate quantities required for control that cannot be directly measured, and to address the issues seen in commercially available NO_x sensors. Accurate estimation is key to optimal control of SCR to meet strict emissions limits for internal combustion engines. Due to the additive nature of the ammonia cross-sensitivity of NO_x sensors, it is difficult to solve to sufficient accuracy. This is compounded by the nonlinear temperature dependence of the cross-sensitivity factor, which has too high a variance to be feasibly characterised, either online or offline. A particle filtering approach has been shown to converge faster than existing Kalman filter methods, and tracks the cross sensitivity accurately during temperature transients, even without a tailpipe ammonia sensor.

Furthermore, a parallel application of this particle filter using a graphics processing unit has been demonstrated, which has computational ability particularly well suited to the particle filtering problem. Whilst not currently applicable to vehicles, this implementation is well suited to testbed SCR development where high fidelity is required. Furthermore, future embedded platforms may see an increase in parallel execution capability, for example field programmable gate arrays (FPGAs). A parallel resampling technique was exploited to reduce the data dependency of the canonical filter formulation. The parallelisation allowed simulation of an order of magnitude more particles within the allotted timeframe, which improved filter tracking.

Chapter 8

Investigation of Embedded MPC Implementation

So far, this thesis has examined the use of model-based control schemes for SCR. In particular, it was found that nonlinear MPC performs particularly well in maintaining constraint-admissible optimal operation of the catalyst across the entire operating range. This includes the addition of a catalyst heater to improve NO_x removal at low temperature, which are typical in the environments where maximum NO_x conversion is critical. Thus far, the implementation of MPC has not been discussed. This is typically a major hurdle to be overcome for applications of predictive control, since the computational requirements are significant, as discussed in Subsection 2.4.6.

In this chapter, the implementation of nonlinear MPC is examined on low power, resource constrained hardware. Particularly, the idea of explicit MPC, as introduced in Chapter 2, is extended into the nonlinear domain. To demonstrate the controller a benchmark system (an inverted pendulum) was used in place of the SCR, since it has a published dynamic model, and other investigations have been carried out regarding its control. This supplies some benchmarks and comparisons against which the controller developed in this chapter can be tested.

8.1 Approximate Grid Explicit MPC

To demonstrate control of a fast nonlinear system, an interesting variant of the inverted pendulum is employed, namely the reaction wheel pendulum [158]. The inverted pendulum has been a popular demonstration platform for nonlinear control for at least fifty years [159]. The transition of the pendulum from a stable (downwards) equilibrium to an unstable (upright) one, known as *swing up*, has been the subject of much research. The most common strategy is an energy based swing up law [160]. It is typical to use

one control law to swing up the pendulum and then another to balance it around the upright position – for example, [161] uses the energy based swing up method followed by a linear parameter varying (LPV) MPC approach to stabilise. Further, [162] uses energy based approaches for both regimes. One successful attempt at a global control law, (i.e., swing up and stabilisation) is made by [163], which uses NMPC for the control of a linear inverted pendulum. The control problem is challenging due to the system being underactuated, with torque applied to a reaction wheel attached to the end of the pendulum, but not to the pendulum itself. A dual-controller approach to swing up and balance the reaction wheel pendulum is introduced by [158]. Subsequent works, including [164] and [165], demonstrate classical nonlinear global controllers. Here, NMPC for global stabilisation of this pendulum is motivated. The nonlinearities in the dynamics mean the control actions to swing up the pendulum cannot be determined by a linear MPC controller, since there exists no single linearisation for which the plant dynamics are adequately captured over any possible state trajectory.

The concept of explicit MPC has found success in the control of fast systems in which the sampling period has typically been too small to allow for online optimisation in the required time frame [63]. The canonical formulation of explicit MPC uses a linear MPC scheme such that the optimisation problems to be solved are quadratic programmes. In this formulation, polytopes in the state space are uniquely defined by the constraints that are active in that region of the state space (the *active set*). For a given polytope, the optimal control law becomes piecewise affine (PWA) in the state x, giving $u^\star(x) = F_i x + g_i$ for the polytope i [166]. The QPs are solved multiparametrically (mp-QP), so that the solution $U^\star(x)$ can be found. As such, the coefficients of the PWA law can be stored offline, and the online search for the optimal solution $u^\star$ reduces to locating the relevant polytope and evaluating the PWA control law found there. This can be done very quickly using approaches such as the *binary search tree* [167].

Unfortunately, in the case of a nonlinear model, the convexity is lost and the solution is no longer a quadratic programme, resulting in the usual explicit MPC strategy being inapplicable. There has been some work regarding overcoming this problem, including [166], in which approaches such as approximation of the cost function by multiple QPs is discussed. In this chapter, several characteristics of the system are exploited to allow a uniform grid to be placed over the state space, and the full NLP (2.76) is solved at each vertex.

Define a generalised coordinate to index the grid:

$$x' = \begin{bmatrix} x'_1, & x'_2, & \ldots, & x'_n \end{bmatrix}^T \in \mathbb{N}^n \tag{8.1}$$

where n is the number of states of the system (2.63). Let $\mathcal{P} = \{1, 2, \ldots, n\}$. Additionally, define the error at a vertex x' in dimension p as:

$$\delta_p(x') = u_0^\star(x') - u_0^\star(\widetilde{x}'(p)) \tag{8.2}$$

where $\widetilde{x}'$ is the adjacent preceeding vertex to x' in dimension p, i.e.,

$$\widetilde{x}'(p) = \left[x'_1, \ldots, x'_p - 1, \ldots, x_n\right]^T \tag{8.3}$$

Let N_p be the number of grid vertices in dimension p. Define the current error in dimension p as:

$$\epsilon_p = \max_{x'} \delta_p(x') \; \forall x' \in \mathcal{X}' \tag{8.4}$$

where $\mathcal{X}' \subseteq \mathcal{X}$ is the set of vertices x' for which the solution will be validated. The grid is iteratively refined until ϵ_p is below a certain tolerance $\bar{\epsilon}$ in all dimensions.

Algorithm 8.3 Generation of an approximate explicit NMPC solution grid

1: **for** each dimension $p \in \mathcal{P}$ **do**
2: Truncate to known bounds and $N_p \leftarrow 3$.
3: **end for**
4: **for** each vertex x' **do**
5: Solve the NLP (2.76) and choose the first element $u_0^\star$ as the solution for the current vertex, $u_0^\star(x')$. Store this solution against this vertex.
6: **end for**
7: **for** each dimension $p \in \mathcal{P}$ **do**
8: evaluate ϵ_p as in (8.4).
9: **if** $\epsilon_p < \bar{\epsilon}$ **then**
10: terminate this dimension and return to step 3 for the next dimension
11: **end if**
12: **end for**
13: **if** all dimensions are complete **then**
14: Terminate.
15: **else**
16: Let $N_p := 2N_p - 1$ and return to step 3.
17: **end if**

8.2 Algorithms

Once Algorithm 8.3 has completed, the result will be a grid of values of $u_0^\star$ for each vertex x'. For real time implementation, this grid is stored alongside the truncation and N_p for each dimension p. The online controller requires the current state x_0, which in general is either measured or provided by an observer.

Algorithm 8.4 Online algorithm to find approximated optimal solution in the solution grid

1: Find the current state x_0 either by measurement or state observer
2: **if** $x_0 \notin \mathcal{X}$ **then**
3: $\quad \widetilde{x}_0 \leftarrow \arg\min_{\widetilde{x}_0 \in \mathcal{X}} \|x_0 - \widetilde{x}_0\|$
4: $\quad x_0 \leftarrow \widetilde{x}_0$
5: **end if**
6: Perform n-dimensional linear interpolation in the solution grid at x_0, the interpolant is the approximated control input $\hat{u}_0^\star$.
7: Apply the control input $\hat{u}_0^\star$ to the plant for the sampling period, then return to Step 1.

In this work, the n-dimensional linear interpolation is performed using the `interpn` function in MATLAB. Note that Step 2 of Algorithm 8.4 should not be required if the truncations were chosen properly, but is included to prevent failure of the algorithm if the state deviates from the solution grid.

8.3 Application to the Reaction Wheel Pendulum

The explicit NMPC controller will be applied to solve the global stabilisation, i.e., the swing-up and the subsequent balancing problem, of a reaction wheel pendulum using a single controller. The solution technique described in Subsection 2.4.4 will be used to transcribe the continuous-time finite horizon optimal control problem and apply the designed controller to a continous-time dynamic model of the pendulum.

8.3.1 Reaction Wheel Pendulum Model

The model of the Quanser reaction wheel pendulum module will be used as shown, together with the simplified free body diagram of its mechanical part, in Figure 8.1 [158]. This system consists of a pendulum which is free to rotate in the vertical plane, along with a reaction wheel at the free end of the pendulum. Torque is applied to this reaction wheel via a motor attached to the pendulum. As such, the system has two degrees of freedom (DOF), since only the wheel is actuated.

The continuous-time equations of motion of the reaction wheel pendulum are as follows [168]:

$$\begin{bmatrix} I_p + I_w & I_w \\ I_w & I_w \end{bmatrix} \begin{bmatrix} \ddot{\theta}_p \\ \ddot{\theta}_w \end{bmatrix} + \begin{bmatrix} -m_p g L_p \sin(\theta_p) \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \tag{8.5}$$

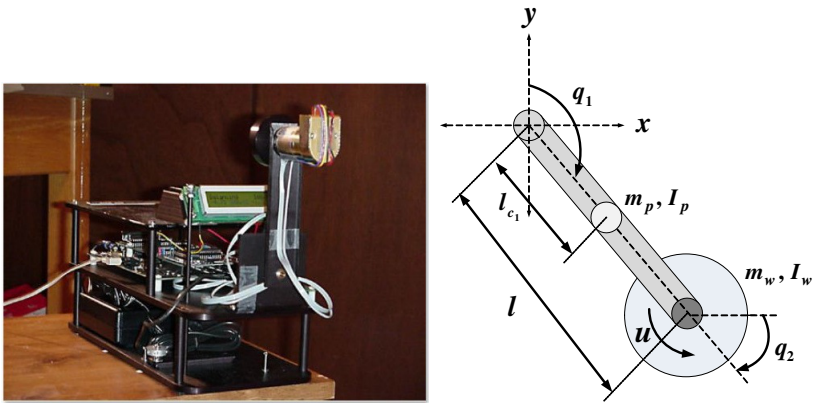


FIGURE 8.1: The Reaction Wheel Pendulum

where m_p and L_p are the mass and length of the pendulum respectively, I_p is the moment of inertia of the pendulum about its pivot and I_w is that of the wheel. g is acceleration due to gravity. The coordinates θ_p and θ_w are the angular positions of the pendulum and wheel respectively. u is the input torque from the DC motor. The physical parameters of the system are listed in Table 8.1. Introducing the following change of coordinates as in [168]:

$$q_1 = \theta_p \tag{8.6}$$

$$q_2 = \theta_p + \theta_w \tag{8.7}$$

leads to the simplified equations of motion:

$$\begin{bmatrix} I_p & 0 \\ 0 & I_w \end{bmatrix} \begin{bmatrix} \ddot{q}_1 \\ \ddot{q}_2 \end{bmatrix} + \begin{bmatrix} -m_p g L_p \sin(q_1) \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} u \tag{8.8}$$

The state vector is chosen to read $x = \begin{bmatrix} q_1, & q_2, & \dot{q}_1, & \dot{q}_2 \end{bmatrix}^T$.

TABLE 8.1: Specifications of the reaction wheel pendulum

Symbol	Description	Value	Unit
m_p	Mass of pendulum	0.2164	kg
L_p	Total length of pendulum	0.2346	m
J_p	Pendulum Moment of inertia	2.233×10^{-4}	kgm^2
m_w	Mass of reaction wheel	0.085	kg
J_w	Wheel Moment of inertia	2.495×10^{-5}	kgm^2
u_{max}	Maximum torque	± 1	Nm

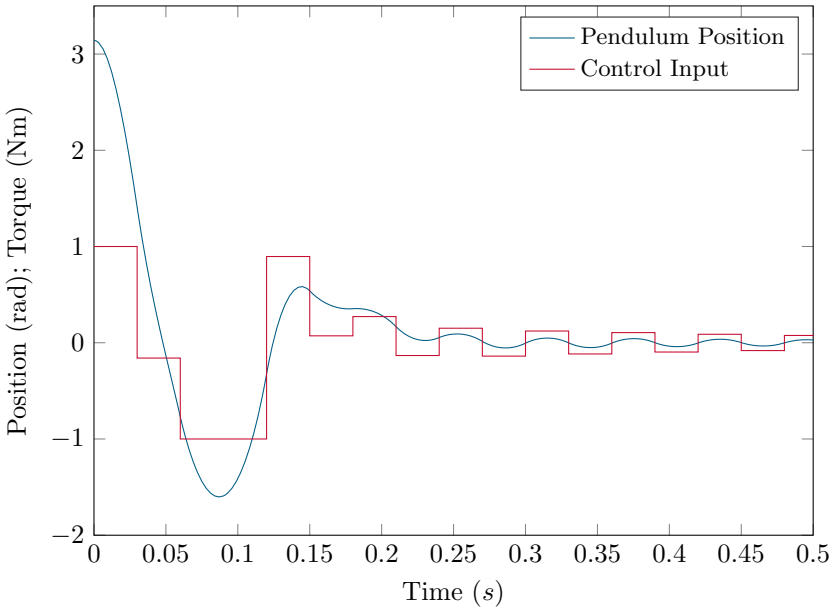


FIGURE 8.2: Swing-up of the reaction wheel pendulum using the NMPC controller

8.3.2 Control Design for the Inertia Wheel Pendulum

The NMPC controller was configured for these simulations with a prediction horizon $H_p = 0.3$ seconds with a shooting mesh grid of $N = 61$ points during which $N = 10$ control moves are equispaced. This results in a sampling time of 30ms for the controller. These values were chosen to provide the smallest NLP, in order that the solution of the NLP (2.76) is as fast as possible such that a fair comparison is possible. Crucially, the balancing performance became significantly worse at larger sampling periods. The stage cost in (2.65) was tuned to be of the following form:

$$l(x, u) = \|x_1\|_{Q_1} + \|x_2\|_{Q_2} \tag{8.9}$$

where $Q_1 = 1000$ and $Q_2 = 5$, to reflect the primary objective of inverting and balancing the pendulum, followed by the secondary objective of holding the reaction wheel static at its origin. The terminal cost $E(\cdot)$ in (2.65) is often required as a guarantee of closed-loop stability. Since no such proof is attempted here, this term was discarded from the cost function. The maximum control torque that can be applied to the reaction wheel is $u_{max} = \pm 1$Nm. The OCP (2.66) was transcribed using `ICLOCS` [103], with `IPOPT` being used to solve the resulting NLP (2.76) [101]. The resulting pendulum arm position and control input during the swing-up are shown in Figure 8.2.

Transcription and solution of the OCP requires a worst case time of 0.465s with an average time of 0.120s on an INTEL Core i7 desktop computer. This is above the sampling period of 30ms and renders the controller infeasible for real time control.

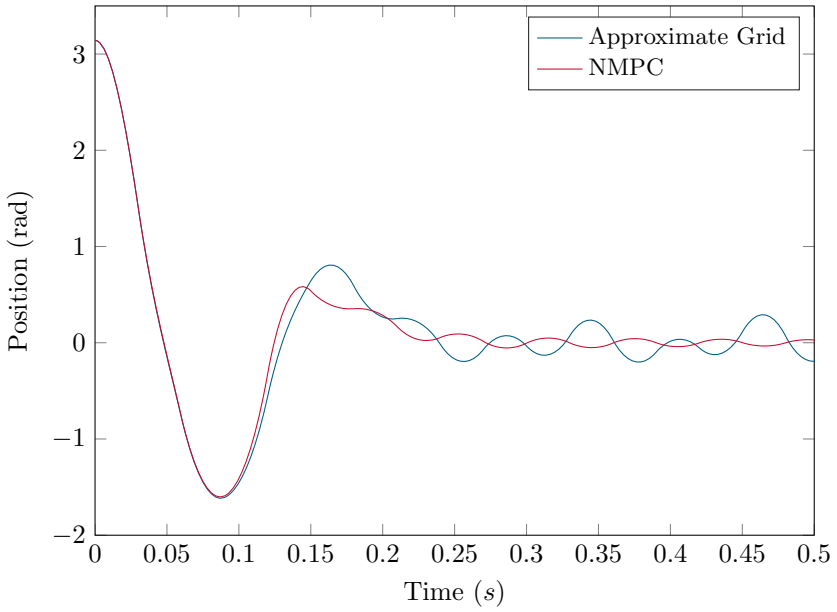


FIGURE 8.3: Comparison of the pendulum arm during swing-up between standard NMPC and the approximate explicit algorithm

8.4 Results and Analysis

8.4.1 Simulations

Applying the approximate grid explicit NMPC Algorithm 8.3 to the reaction wheel pendulum system, the truncations $q_1 \in [-\pi, \pi]$, $q_2 \in [-10, 70]$, $\dot{q_1} \in [-100, 100]$ and $\dot{q_2} \in [-1500, 1500]$ were chosen as a result of swing-up simulation results from the full NMPC controller. The set $\mathcal{X}'$ is selected as the central 50% in each dimension p, such that the balancing control region is guaranteed to be of sufficient fidelity. The swing up region towards the edges of the state space (in q_1 and $\dot{q_1}$) can tolerate less accuracy, as will be demonstrated later. $\bar{\epsilon}$ was chosen to be 0.1. The algorithm ceases with 20 points in the x_1 dimension, 50 points in the x_3 dimension, and only 5 points in each of the x_2 and x_4 dimensions. This correlates with our expectation that the angular velocity of the pendulum has the most significant impact on the control action, followed by the position of the arm. The position and velocity of the reaction wheel are significantly less important to achieve the required control action. The position of the pendulum during swing-up is compared in Figure 8.3 between the NMPC and approximate grid explicit NMPC, and the control input comparison is shown in Figure 8.4.

The control action as a function of the position and speed of the pendulum arm are shown in Figure 8.5 (for one set of values of x_2 and x_4). It is clear that there are large regions of the state space where the torque limit is in effect and therefore governs the control action. Between these, there is a region where the control action changes in a smooth manner between one saturation and the other. The visibility of this gradually changing value justifies our approximation of linear interpolation between grid vertices.

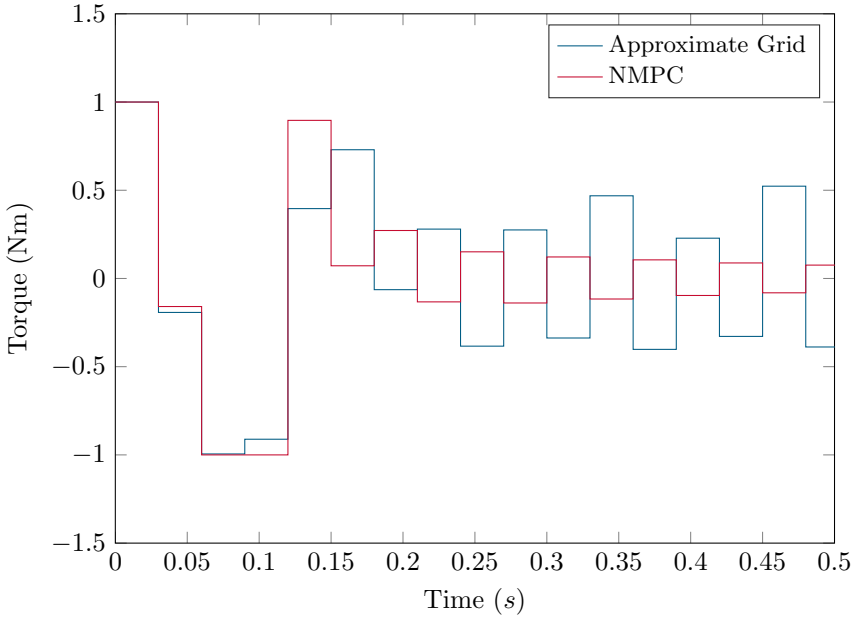


FIGURE 8.4: Comparison of the control input during swing-up between standard NMPC and the approximate explicit algorithm

The bottom-left and top-right corners of the image are regions in which the control law falls sharply back to approximately zero. These mark regions in which the maximum input torque does not permit an immediate return to the upright position. Instead, the controller relaxes the control input to around zero until the pendulum returns to the controllable region. On these boundaries the linear approximation is significantly less accurate; this becomes the main source of error in the algorithm, as seen in Figure 8.3. However, it is tolerable in practice, particularly since the system is returning to the higher fidelity interpolation region, from where more precise control will occur.

8.4.2 Computational Effort

Given that state measurements are available, the online part of the controller amounts to a 4 dimensional linear interpolation in the solution grid. Solution of the NLP online, even on a relatively powerful desktop computer, is around two orders of magnitude too slow for real time implementation. On the same desktop computer, the approximate grid NMPC algorithm requires worst case and average times of 1.8ms and 1.1ms respectively. These are more than one order of magnitude below the sampling time of the reaction wheel pendulum plant and the controller is therefore applicable in real time.

The resultant solution grid has $25,000$ vertices, to each of which is attached a control input in double precision floating point (8 bytes). Thus, storage of the grid requires around 200kB of memory. This memory requirement, and the required online interpolation in the solution grid, are expected to be achievable on a microprocessor platform, making this solution highly applicable to real-time controllers. For example, the STM32F407

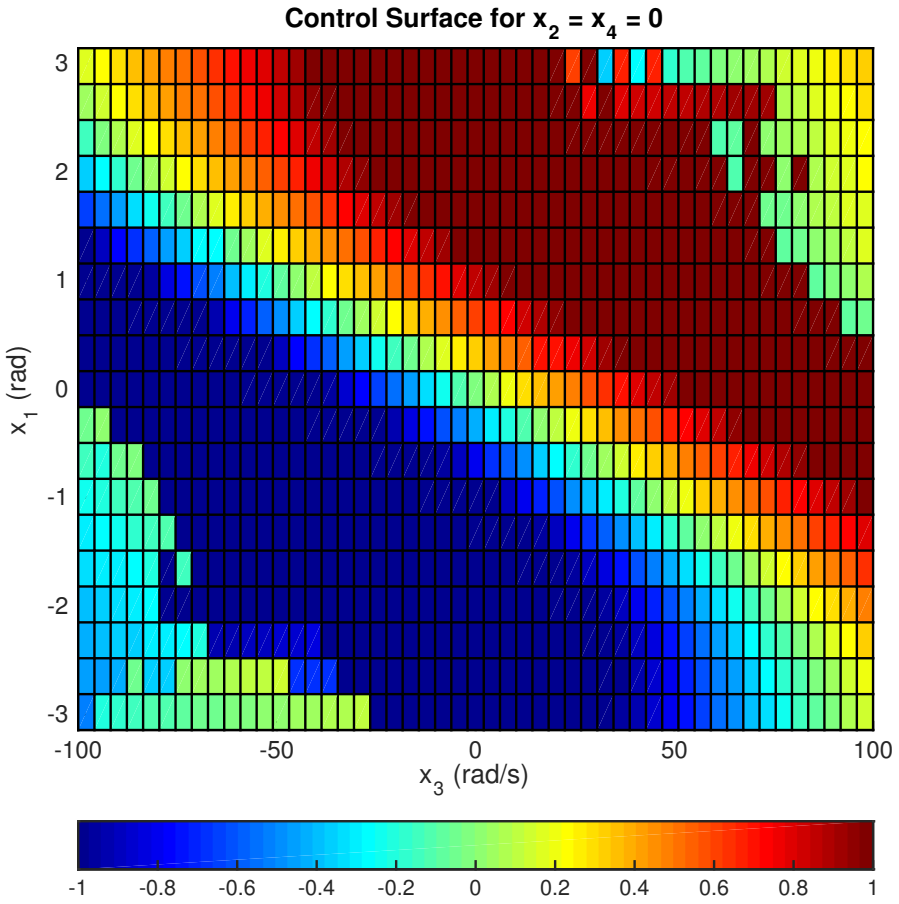


FIGURE 8.5: Control action as a function of pendulum angle and velocity, with reaction wheel angular position and velocity zero ($x_2 = x_4 = 0$).

device from ST Microelectronics has a core clock of 168MHz and 1MB of flash program memory [169], and is available for around GBP 7. In contrast to a PWA solution that requires $n + 1$ numbers to store per polytope (where n is the number of states), this solution requires only 1. To improve the storage situation still further, adjacent vertices whose values of $u_0^\star$ are the same can be merged such that the grid becomes non-uniform (such as the large single colour areas in Figure 8.5). This reduction of the solution grid is likely to decrease the storage requirements considerably for systems like the reaction wheel pendulum wherein inputs are at the constraints $\pm u_{max}$ for large portions of the state space.

8.5 Conclusions

It was demonstrated that despite being able to achieve the desirable optimal performance without the switching logic requirements, NMPC is not applicable online to some systems. The reaction wheel pendulum was employed as an example of a fast nonlinear control problem. The difficultly in online application comes from the time required for solution of the NLP. To overcome this problem, an explicit NMPC algorithm was derived

for the determination of a solution grid. Insensitivity of the control input to a subset of the states was exploited to reduce the size of the solution space. The tradeoff between approximation error, and the requirements for storage and computation time can be simply tuned by adjustment of the error tolerance $\bar{\epsilon}$. This approximate solution grid provided will allow real-time implementation of the controller with performance close to the full NMPC controller, even with a modest error tolerance.

It is possible that such a scheme could be employed for the control of SCR. The error tolerances and the storage required for the offline grid would be determined by deriving an approximate explicit grid controller for the SCR system. It is required that the state space be truncated in this algorithm, so research will be required as to the operating ranges in which the SCR controller would be expected to operate. However, if these issues can be addressed, the offline approximate grid nonlinear MPC algorithm may provide a useful route to real-time application of the MPC controllers for SCR that were developed in Chapter 6.

Chapter 9

Hardware Accelerated MPC

In this chapter, methods of embedding the MPC controllers designed in Chapter 6 are explored. In order to achieve this, the frameworks for numerical solution of optimal control problems as introduced in Chapter 2 are employed, and are implemented on embedded hardware. Whilst real time solution of the whole MPC problem in the current setup is possible using a desktop computer, it is found to be infeasible on embedded hardware that is highly resource constrained. Consequently, each part of the algorithm is profiled to quantify its execution speed and understand which portions consume the most time. To permit real time embedded solution, a system-on-chip (SoC) recently introduced to the market is leveraged: the Xilinx Zynq 7000, which is available in automotive grade.

The real-time iterations (RTI) approach is a particular arrangement of the SQP algorithm as introduced in Subsection 2.4.6. The fundamental concept is that the SQP algorithm becomes restricted to the solution of only one QP at each time step. The RTI algorithm is split into two phases: the preparation phase in which numerical integration of the model is computed, and the feedback phase after the state measurement has been found the associated QP can be constructed and solved. Multiparametric solution of the QPs (with qpOASES, for example) results in an extremely fast feedback phase. The requirement for robust integrators with sensitivity generation in the preparation phase, however, means this is often quite slow in practice. If the system of the ODEs to be solved demonstrates highly stiff properties, this numerical integration can be too slow to be solved within the sampling time of the plant. Implicit Runge-Kutta (IRK) integrators have proven popular in recent nonlinear MPC literature: they cope well with stiff ODEs and it is relatively simple to add the required sensitivity generation [98].

The models developed in Chapter 4 are typically highly stiff, due to the combination of thermal and chemical differential equations. Furthermore, in the MPC controllers developed in Chapter 6, a long prediction horizon is used in order to capture the storage

and thermal dynamics of the catalyst. This combination leads to the numerical integration part of the RTI algorithm becoming exceptionally computationally difficult. To counter this, the Zynq-7000 SoC is employed, which combines ARM Cortex A9 microprocessors with close-coupled FPGA fabric. As discussed in Chapter 2, FPGAs have previously been leveraged for MPC. However, typically the full MPC algorithm is implemented in the FPGA. Conversely, *hardware accelerated* MPC, where a portion of the algorithm runs on the FPGA and the rest on a microprocessor, has traditionally been limited by the bandwidth between the two. The close-coupled nature of the Zynq platform is used to overcome this to a large extent in this chapter, by using the FPGA to accelerate the numerical integration whilst the rest of the RTI algorithm runs on the microprocessor.

9.1 Review of Implicit Integration Methods

Recall, as introduced in Subsection 2.4.6, that we are interested in numerical integration of the general system of nonlinear differential equations describing a continuous time dynamic model, as introduced in Subsection 2.4.4, that we can write as

$$\dot{x}(t) = f(x(t), u(t)), \tag{9.1}$$

where the ODE f depends on time t only through the system states x and inputs u. The problem is in general the finding of the solution to the IVP over the time interval Δt from some initial state x_k to some new state x_{k+1} given some control input u defined over the time interval

$$F(x_k, u_k) = x_{k+1} = \int_{t_0}^{t_0+\Delta t} f(x(t), u(t))\,\mathrm{d}t, \tag{9.2a}$$

$$x(t_0) = x_k. \tag{9.2b}$$

Solution of (9.2) requires numerical integration of the plant. Since stiff systems of ODEs are the focus of this work, implicit integration methods are examined, since they exhibit larger regions of stability compared to explicit methods. An implicit method finds the solution to

$$Y(x_n, x_{n+1}) = 0 \tag{9.3}$$

where x_n is the current state and x_{n+1} is the desired state at some time in the future. In particular, implicit Runge-Kutta (IRK) methods are A-stable, meaning that if the solution of the ODE tends to some finite value, so does the numerical integrator output

[170]. If this property holds regardless of step size, the method is said to be A-stable. Furthermore, IRK methods are amenable to sensitivity generation which is typically required in gradient-based optimisation.

9.1.1 Implicit Runge-Kutta Methods

The s-stage Runge-Kutta (RK) method computes the solution of the differential equation at a time future point in time as

$$x_{n+1} = x_n + h\sum_{i=1}^{s} b_i k_i \tag{9.4}$$

where h is the integration time step and the coefficients k_i are found as

$$k_i = f\left(x_n + h\sum_{j=1}^{s} a_{ij}k_j, u_n\right) \ \forall i \in 1,\ldots,s \tag{9.5}$$

where it assumed that u_n is a known constant for each n. A particular Runge-Kutta method is uniquely defined by its Butcher tableau [171]:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} \tag{9.6}$$

where a_{ij} and b_i appear in (9.4) and (9.5). The c_i quantities are employed when the ODE is a function of time, so they are ignored in this discussion since (9.1) is not a time-varying function.

If the Butcher tableau is lower triangular, the RK method is explicit, whereas IRK methods do not display this property. As a result, in (9.5), the equations for k_i are interdependent. Therefore, solution of the system of equations for k_i is achieved by defining the vector k

$$k = \begin{bmatrix} k_1, & k_2, & \ldots, & k_s \end{bmatrix} \tag{9.7}$$

and defining a new function

$$G(k) = \begin{bmatrix} f\left(t_n + c_1 h, x_n + h\sum_{j=1}^{s} a_{1j}k_j\right) - k_1 \\ f\left(t_n + c_2 h, x_n + h\sum_{j=1}^{s} a_{2j}k_j\right) - k_2 \\ \vdots \\ f\left(t_n + c_s h, x_n + h\sum_{j=1}^{s} a_{3j}k_j\right) - k_s \end{bmatrix} \tag{9.8}$$

and seeking a solution to the equation

$$G(k) = 0. \tag{9.9}$$

This equation has no exact solution in general, so instead a k that minimises the mean-squared error is sought:

$$k^\star = \arg\min_k \|G(k)\|^2. \tag{9.10}$$

This is a nonlinear least squares problem, solution of which can be achieved by applying Newton's method, which for a matrix problem like (9.10) can be written as:

$$\left.\frac{\partial G}{\partial k}\right|_{k=\tilde{k}_i} \Delta k = -G(\tilde{k}_i) \tag{9.11}$$

with the iterative update

$$\tilde{k}_{i+1} = \tilde{k}_i + \Delta k. \tag{9.12}$$

9.1.2 Least Squares Problems

Equation (9.11) is in fact a linear least squares (LLSQ) problem, which can be generically written as

$$Ax = b, \tag{9.13}$$

with $A \in \mathbb{R}^{m\times m}$ and $b \in \mathbb{R}^m$. This is in general an overdetermined system, and the solution is defined as the x which minimises the residual error, defined as

$$\hat{x} = \arg\min_x \|b - Ax\|^2. \tag{9.14}$$

Some of the most stable algorithms when A is real and square employ QR decomposition, in which the matrix A is decomposed as $A = QR$, where Q is an orthogonal matrix and R is upper or lower triangular, depending on the implementation. This is shown in Algorithm 9.5. The number of iterations of (9.12) required depends on the properties of (9.1), principally the degree of nonlinearity and the stiffness of the system. Once k has been found, the IRK update (9.4) is used to compute x_{n+1}.

Algorithm 9.5 Algorithm for solution of a linear least squares problem using QR decomposition.

```
1: procedure LEAST_SQUARES(A, b)
2:     {Q, R} ← QR(A)                    ▷ QR decomposition of A
3:     c ← Q^T b
4:     x̂ ← R^{-1} c                      ▷ Gaussian elimination
5:     return x̂
6: end procedure
```

9.1.3 Sensitivity Generation

In addition to the solution of the IVP (2.77), gradient based optimisation methods require the derivatives of the cost and constraint functions with respect to the decision variable z. The derivative of the defect constraints (2.76d) requires the matrix $\frac{\partial F(s_n, q_n, p)}{\partial z}$.

There are many options for calculating this quantity, see [97] for a discussion. The approach favoured in many real-time MPC applications, such as [98], is derived by applying the implicit function theorem (IFT) to the IRK update equation (9.4):

$$\frac{\mathrm{d}k}{\mathrm{d}z_n} = -\frac{\partial G}{\partial k}^{-1} \frac{\partial G}{\partial z_n} \tag{9.15a}$$

$$\frac{\partial x_{n+1}}{\partial z_n} = \frac{\partial x_n}{\partial z_n} + h \sum_{i=1}^{s} b_i \frac{\mathrm{d}k_i}{\mathrm{d}z_n}. \tag{9.15b}$$

The matrix $\frac{\partial G}{\partial k}$ and its factorisation has already been calculated as part of the IRK iterations (9.11).

9.2 Parallelisation of Multiple Shooting

The direct multiple shooting algorithm was introduced in Subsection 2.4.6, wherein it was discussed that this framework involves the splitting of the prediction horizon into N segments. Numerical integration of each of these intervals is carried out independently, with constraints added to the optimisation programme to ensure the consistency of trajectories across segment boundaries.

Defining the number of integration steps to cross each multiple shooting segment as $\beta = \frac{H}{h}$, Algorithm 9.6 displays the numerical integration part of the direct multiple shooting algorithm. The outermost loop iterates over each of the N steps of the prediction horizon. The i loop iterates over the integration steps required inside each shooting segment. The innermost j loop executes a fixed number L of Newton iterations as given in (9.11). Note that unlike [97], the quantity $\frac{\partial G}{\partial k}$ is recalculated at each Newton step as this becomes one step of the accelerator pipeline and so does not require additional computation time. Conversely, it generates an improved guess at k at each IRK iteration, which thereby reduces the number of Newton iterations (9.12) required for convergence of $k^\star$.

Algorithm 9.6 Algorithm for numerical integration part of direct multiple shooting using the IRK integrator.

```
1: procedure MULTIPLE_SHOOTING(z)
2:     for n = 1 : N do
3:         for i = 1 : β do
4:             for j = 1 : L do
5:                 Δk ← ∂G/∂k|_k^{-1} G(k)
6:                 k ← k + Δk
7:             end for
8:             s_{n+1} ← s_n + h Σ_{i=1}^{s} b_i k_i
9:             compute ∂s_{n+1}/∂z_n as given in (9.15)
10:        end for
11:    end for
12: end procedure
```

In Algorithm 9.6, each Newton step is dependent on the previous iteration, since line 5 uses the value of k_n computed at the previous iteration. Now a trait of direct multiple shooting is exploited: each segment is independent. Thus it is possible to rearrange Algorithm 9.6 as shown in Algorithm 9.7.

Algorithm 9.7 computes a Newton step for each of the N segments in turn and does this L times, after which k_n is a solution for (9.9) in each segment n. The IRK update (9.4) is then applied and the sensitivity (9.15) calculated for each segment $n \in \{0, \ldots, N-1\}$.

9.3 Hardware Accelerator Design

Algorithm 9.7 requires $N \times L \times \beta$ solutions of the LLSQ problem (9.11). This is computationally expensive, particularly where β and L are large, which is commonly the case for stiff systems, and where N is large, which implies a long horizon. However, the arrangement in Algorithm 9.7 permits batches of $N \times L$ LLSQ problems to be solved in a pipelined manner. As such, in this section, a pipelined hardware LLSQ accelerator is designed.

Algorithm 9.7 Restructured direct multiple shooting integration algorithm to remove data dependencies from inner loops.

1: **procedure** MULTIPLE_SHOOTING_PARALLEL(z)
2: **for** $i = 1 : \beta$ **do**
3: **for** $j = 1 : L$ **do**
4: **for** $n = 1 : N$ **do**
5: $\Delta k \leftarrow \frac{\partial G}{\partial k}\big|_{k_n}^{-1} G(k_n)$
6: $k_n \leftarrow k_n + \Delta k$
7: **end for**
8: **end for**
9: **for** $n = 1 : N$ **do**
10: $s_{n+1} \leftarrow s_n + h \sum_{i=1}^{s} b_i k_{n,i}$
11: compute $\frac{\partial s_{n+1}}{\partial z_n}$ as given in (9.15)
12: **end for**
13: **end for**
14: **end procedure**

It should be noted that the QR decomposition (line 2 of Algorithm 9.5) is the most expensive part of the LLSQ solver, yet it does not require the vector b. The accelerator can therefore be optimised by beginning the QR decomposition as soon as A is received from the CPU, before b is transmitted. It should also be noted that line 4 of Algorithm 9.5 does not involve a matrix inversion since the matrix R is triangular, and $\hat{x}$ is therefore found by Gaussian elimination.

The accelerator was constructed with a 5-stage pipeline as shown in Figure 9.1. Each stage of the pipeline is separated by a block RAM (BRAM) in the FPGA fabric configured in ping-pong mode. Note that in stage 2, the accelerator is already computing the QR decomposition of A whilst b is transferred from the CPU. This arrangement permits the accelerator to consume either an A or a b matrix at each stage of pipeline execution, and produces a result $\hat{x}$ at each iteration.

The advanced extensible interconnect standard, version 4 (AXI4) is adopted by Xilinx for intellectual property (IP) blocks which are implemented in the programmable logic. The point-to-point address-free AXI-Stream interconnect was chosen for the data input and output interfaces of the least-squares accelerator, since it provides low latency for bulk data transfer. An AXI-Lite interface is also provided by the accelerator for initialisation, control and diagnostics.

Resource utilisation for the accelerator is shown in Table 9.1. Block RAMs (BRAMs) are 18 kilobyte memories which are used for creating the pipeline stages and storage inside the pipeline stages. The DSP48 slices are configurable multiply-accumulate blocks that can be leveraged to perform arithmetic operations rather than implementing custom logic. Flip-flops (FFs) and look-up tables (LUTs) are used to construct custom logic to implement the hardware accelerator.

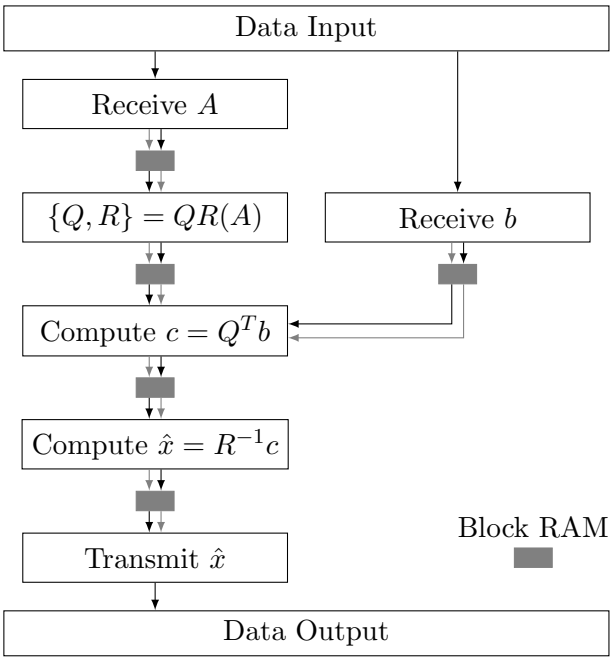


FIGURE 9.1: Pipelining arrangement of the hardware LLSQ solver, showing block RAMs configured as ping-pong memory to create the pipeline.

TABLE 9.1: Resource utilisation of the LLSQ accelerator, showing percentages for the ZC702 device.

Stage\Resource	BRAM	DSP48	FF	LUT
$QR(A)$	2	40	8765	12178
$c = Q^T b$	8	5	734	1243
$\hat{x} = R^{-1}c$	0	5	1412	1940
Pipeline	4	0	3328	192
Other	0	0	26	83
Total	14	50	14503	15820
Utilisation (%)	5	22	13	29

9.4 Accelerator Implementation

The Xilinx Zynq ZC702 SoC contains a pair of ARM Cortex-A9 application microprocessors coupled with an Artix-7 series FPGA. A development board as shown in Figure 9.2 provides test and debug facilities as well as flash storage for the program code. The LLSQ accelerator was implemented into the programmable logic (PL) of the Zynq, along with a DMA controller as shown in Figure 9.3, which is an IP block provided by the vendor. The DMA controller interfaces between the streaming interface of the accelerator and the memory mapped CPU. The DMA was configured to read and write directly from on-chip memory (OCM) via the accelerator coherency port (ACP) of the processing system, since this gives the lowest latency access to the CPU memory system. Whilst the development board provides off-chip DDR memory, it was disabled for this

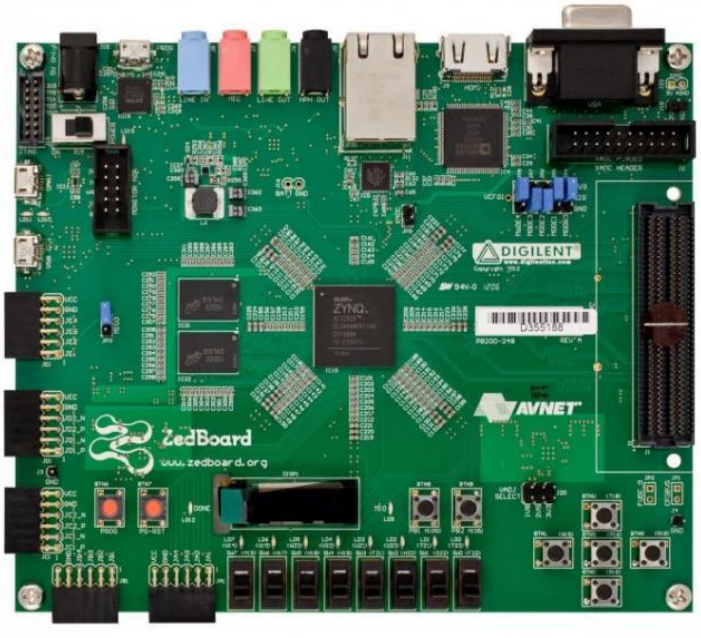


FIGURE 9.2: The Zedboard Xilinx Zynq-7000 development board.

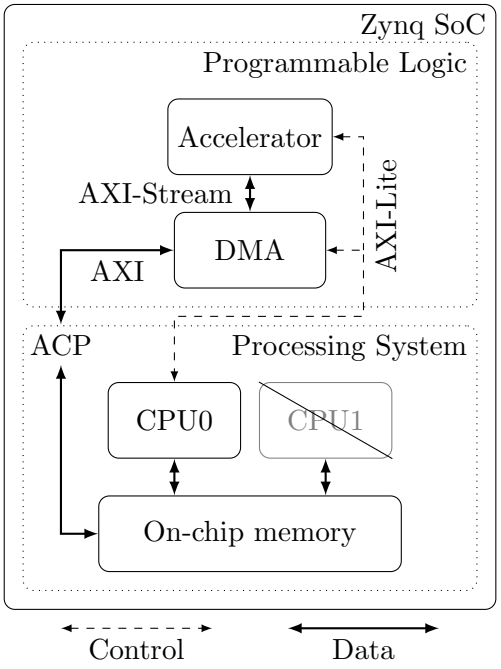


FIGURE 9.3: Schematic of the Zynq SoC architecture, with only CPU0 used in this application whilst CPU1 remains in suspend mode.

application since it is typically not available in small embedded systems. Program code was stored in the flash read-only memory (ROM).

The performance of the implemented accelerator was tested by comparing the computation time for varying numbers of LLSQ problems as solved by the accelerator versus a pure software implementation, with results shown in Table 9.2. The reported times

TABLE 9.2: Reported speedup as function of number of sequential independent LLSQ problems.

No. of problems	Software Time (ms)	Accelerated Time (ms)	Speedup
1	0.271	0.044	6.16
5	1.251	0.113	11.1
10	2.478	0.200	12.4
50	12.12	0.893	13.6
100	24.46	1.761	13.9
1000	242.2	17.38	13.9

include the time taken for transfer of data to and from the accelerator over the AXI-Stream interface, but do not include the time taken to initialise and configure the various hardware, including the accelerator and the DMA controller. It is clear that even with just one LLSQ problem, the accelerator outperforms the software implementation, however it is questionable as to whether the additional cost of the hardware and design time are worth it for few LLSQ problems. For greater numbers of problems, the speedup increases. When there are 50 or more problems to be solved, the accelerator reaches its maximum speedup.

9.5 Application to SCR Controller Implementation

The MPC problem to be applied is that formulated in Section 6.3, where the MPC problem is to simultaneously control the urea injection and electric catalyst heater. As such, the model comprises $n_x = 4$ states and $n_u = 2$ inputs. A Radau-IIA IRK method was chosen, which has $s = 3$ stages. This gives $m = s \times n_x = 12$ in (9.13). With this size of LLSQ problem, the pipelined accelerator was able to execute at a initiation interval (II) of 2064 cycles at a clock speed of 117 MHz, meaning new inputs are accepted and new outputs are generated every 17.6 μs. At this rate, the accelerator is able to solve around $56,600$ LLSQ problems of size 12×12 per second.

The real-time iterations algorithm [172] with condensing was implemented to solve the optimal control problem. The QP which is formed at each time step and whose Hessian has size $Nn_u \times Nn_u$ was solved with the embedded version of qpOASES [83]. The OCP was designated a $H = 1$ second control interval and a $H_p = 20$ second prediction horizon, giving $N = 20$ in (2.73). Due to a highly stiff model, the integration time step can be no larger than 100 milliseconds for successful convergence of the numerical integration, giving $\beta = 10$. The number of Newton iterations was fixed at $L = 20$, since this permitted convergence in the worst-case conditions, such as step changes in disturbances. This value was fixed rather than using a runtime heuristic in order to produce an algorithm with a fixed and deterministic execution time.

TABLE 9.3: Profiler result of real-time iterations on the Zynq-7000 with and without hardware acceleration of the least-squares solver.

Function	Native Time (ms)	Accelerated Time (ms)
`solve_odes`	1365	320
`qp_condense`	174	174
`qp_solve`	131	131
`solve_sensitivities`	88	88
`generate_matrices`	88	-
Total	1950	772

The most time consuming five functions in the real-time iterations algorithm, both with and without the hardware least-squares accelerator, are shown in Table 9.3. Notice that without acceleration, the numerical integration of the ODEs (`solve_odes`) takes around 70% of the time per iteration. However, the accelerated solver carries out the numerical integration in less than 25% of the time required by the native version. The time for generation of the matrices for the least-squares problems is not shown since this generation is part of the accelerator pipeline. The time required for condensing (`qp_condense`) and solving the QP (`qp_solve`) is unchanged, the speedup comes only from the acceleration of the numerical integration.

9.6 Concluding Remarks

Real-time solution of optimal control problems remains a barrier to their wider uptake by industry. Recent work has provided algorithms to calculate the optimal controls (or an approximation thereof) at each time step, but the numerical integration of the plant and the associated sensitivity generation can preclude real-time application, particularly for stiff systems on long horizons. Implicit integrators are known to be efficient for stiff systems due to their large region of stability. However, the requirement for solution of nonlinear least-squares problems can lead to excessive computational complexity.

In this chapter, a hardware accelerator was implemented to solve the Newton steps for calculation of the solution of the nonlinear least squares problem. It was shown that this accelerator could be pipelined and rearrangement of the multiple shooting algorithm could exploit the speedup that this provided. The real-time iterations algorithm was implemented on a Zynq SoC with the hardware accelerator attached to the microprocessor via a low-latency streaming interface. The speedup provided was tabulated versus the number of least-squares problems to be solved, and linked to the stiffness and horizon length of the MPC problem. The MPC controllers for SCR designed in Chapter 6 were implemented using the hardware accelerator and it was demonstrated that this topology permits real-time execution of the MPC controllers. This is despite a stiff system of

differential equations arising from the combination of chemical and thermal dynamics, as well as a long prediction horizon which is required to adequately capture the plant dynamics.

Chapter 10

Conclusions and Further Research

10.1 Conclusions

Recent research into the effects of NO_x on human health have prompted investigation into areas where these pollutants are highly concentrated. One key area of concern is urban environments, which has resulted in legislation to drastically reduce the permitted emissions from road vehicles, since lean-burn internal combustion engines are one of the principal culprits. The problem is largely still unsolved, and major cities across the world remain in breach of safe exposure levels. To meet their targets, the automotive industry has invested heavily in aftertreatment systems, of which SCR is currently favoured across a broad range of applications, from passenger vehicles to heavy duty. This thesis aims to find solutions to solve these problems by improving SCR efficiency through advanced control.

This thesis began by examining the current state of the art in automotive aftertreatment, concentrating particularly on SCR. Control and estimation strategies were assimilated to produce a broad overview of the current methods applied in industrial applications of SCR. The control of the SCR catalyst was shown to be an open problem, with many previous works in the literature under various control frameworks.

To understand why current applications of SCR were not achieving the design levels of NO_x reduction in practice, urban passenger buses in Southampton were instrumented with PEMS and telematics equipment. Extensive data was acquired from these experiments which gave significant insight into the operation of SCR catalysts in real world driving. The results led to the understanding that catalyst temperature is critical to effective operation, and therefore simultaneous chemical and thermal management of SCR was required in order to achieve the desired NO_x removal in low speed driving conditions.

With the aim of model based control and estimation in mind, this thesis then used the modelling techniques in the literature to construct high fidelity catalyst models and control oriented models, which are less accurate but permit faster-than-realtime execution. In order to provide a pathway to application of model based controllers for SCR in industry, the reference governor framework was explored and proven feasible for SCR. This proved that this approach allows integration of systematic constraint handling into existing feedback controllers, as are currently utilised in industry. The applicability of this approach on automotive hardware was also demonstrated. The model based control research then expanded to MPC, wherein linear and nonlinear control approaches were explored. The performance improvement of MPC over industry standard map-based controllers was thoroughly proven. The nonlinear approach was proven to be more capable with regards to constraint satisfaction across the operating temperature range of the catalyst, and the natural extension of MPC to MIMO problems was demonstrated by the inclusion of a catalyst heater to provide thermal management of the catalyst alongside chemical control.

Real applications of SCR controllers require state estimation, as was discussed in the review of the existing literature. The highly nonlinear nature of the models developed in this thesis led to performance issues with existing estimators. As a result, two novel estimators were designed based on the UKF and particle filter frameworks. Their performance was shown to be noticeably improved over current strategies, and the UKF was considered applicable for real-time applications. The computational complexity of the particle filter was mitigated by leveraging the parallel computing capabilities of modern GPUs.

Real-time application of model based control is still a barrier to their wider adoption by industry. This thesis has designed two novel extensions to the literature on fast solution of optimal control problems. Firstly, the explicit MPC framework was extended into the nonlinear MPC domain by an approximate grid strategy, which employed parallel computing for fast generation of the offline data but relied on truncation of the state space which is not always practical. Secondly, the real-time iterations approach was investigated, where it was determined that online solution of the stiff model was the principal barrier to performance. In this vein, a hardware accelerated least squares solver was designed and implemented on the Xilinx Zynq SoC. This approach permits real-time solution of stiff systems on long horizons in MPC, which is typically a difficult combination.

This thesis reconciles the recent academic progress in the design and application of model-based control and estimation with the tightening emissions legislation for lean-burn internal combustion engines. It has covered the design and calibration of models, various model-based control approaches and their pathways to integration with existing software and hardware, and strategies to manage their inherent computational complexity. All of these steps are required for the successful industrial adoption of model

based control. The results presented in this thesis indicate that predictive control shows promise for achieving the stated aims of improving SCR efficiency. It is therefore recommended that further research is carried out with the aim of demonstrating that this technology can be embedded successfully into an automotive electronic control unit and benchmarked on a testbed to prove that the performance gains can be realised in a practical setting.

10.2 Further Research

Model predictive control holds numerous advantages over classical control, many of which have been clearly exhibited in its application to SCR, and similar benefits can be expected in other areas of the automotive industry. This thesis has shown that MPC achieves improved performance, decreased calibration effort and constraint certification, which are all highly desirable in the automotive industry and beyond. Consequently, the main technical barrier to its uptake is now the computational aspects, which are eminently surmountable, even on automotive grade embedded hardware. This situation will improve as microprocessor and embedded computational hardware continues to increase in capability and fall in cost. The automotive industry is highly conservative, so the legal and political barriers are also something which must be overcome. This thesis has presented some technical pathways to enable this adoption. The future for MPC in the automotive industry is thus hugely promising, and therein lie improvements in all arenas of automotive control, including but not limited to performance, emissions and safety.

Research continuing from this thesis could take one of many directions. Some further directions which are important to address could be listed as follows, which is far from being exhaustive:

- Examination of the robustness of predictive control in the face of manufacturing and ageing-related variability. The literature around robust and adaptive MPC is expansive and many of the techniques and results presented therein could be applied to the SCR control problem.
- Demonstration of predictive controllers on a testbed SCR. This would require calibration of an SCR model, such as those in Chapter 4, to match a physical catalyst. It is not necessary to embed the controller on automotive grade hardware in order to verify its performance.
- Embedding real-time predictive control on an automotive ECU with the hardware acceleration components designed as custom application-specific integrated circuits (ASICs). This research would involve demonstrating that the custom hardware

designed in Chapter 9 can be ported to custom silicon and linked to an application processor via a fast interconnect, perhaps Ethernet or CAN.

- Examine further expanding the authority of the predictive controller to include other system components such as high/low pressure EGR, boost pressure and injection timing in order to quantify if any or all of these variables permit further emissions reductions. Whilst this implies modelling of these components, this is often carried out during control design to permit model-in-the-loop verification.
- Significant effort is being expended to improve the chemical and thermal properties of the catalysts which are used for SCR. Combining advanced control with these efforts would be a highly worthwhile direction for future work.

List of References

[1] Department for Transport. *Vehicle Licensing Statistics*. URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/198753/vls-2012.pdf (visited on 14th Aug. 2017).

[2] Working Group on Nitrogen Dioxide. *Position Paper on Air Quality: Nitrogen Dioxide*. 1997. URL: http://ec.europa.eu/environment/air/pdf/pp_no2.pdf (visited on 14th Aug. 2017).

[3] Environmental Protection Agency. *Nitrogen Dioxide*. 2014. URL: http://www.epa.gov/air/ozonepollution/basic.html (visited on 14th Aug. 2017).

[4] J. Hussain, K. Palaniradja, N. Alagumurthi and R. Manimaran. "Effect of Exhaust Gas Recirculation (EGR) on Performance and Emission characteristics of a Three Cylinder Direct Injection Compression Ignition Engine". In: *Alexandria Engineering Journal* 51.4 (2012), pp. 241–247.

[5] "Four of world's biggest cities to ban diesel cars from their centres". In: *The Guardian* (2016). URL: https://www.theguardian.com/environment/2016/dec/02/four-of-worlds-biggest-cities-to-ban-diesel-cars-from-their-centres (visited on 14th Aug. 2017).

[6] "Paris bans old polluting diesel cars". In: *Engineering and Technology* (2017). URL: https://eandt.theiet.org/content/articles/2017/01/paris-bans-old-polluting-diesel-cars/ (visited on 14th Aug. 2017).

[7] J. Reijnders, M. Boot and P. de Goey. "Impact of aromaticity and cetane number on the soot-NO_x trade-off in conventional and low temperature combustion". In: *Fuel* 186 (2016), pp. 24–34.

[8] M. Zheng, G. T. Reader and J. Hawley. "Diesel engine exhaust gas recirculation – a review on advanced and novel concepts". In: *Energy Conversion and Management* 45.6 (2004), pp. 883–900.

[9] A. Maiboom, X. Tauzia and J.-F. Hétet. "Experimental study of various effects of exhaust gas recirculation (EGR) on combustion and emissions of an automotive direct injection diesel engine". In: *Energy* 33.1 (2008), pp. 22–34.

[10] I. P. Kandylas and G. C. Koltsakis. "NO_2-Assisted Regeneration of Diesel Particulate Filters: A Modeling Study". In: *Industrial & Engineering Chemistry Research* 41.9 (2002), pp. 2115–2123.

[11] B. Guan, R. Zhan, H. Lin and Z. Huang. "Review of state of the art technologies of selective catalytic reduction of NO_x from diesel engine exhaust". In: *Applied Thermal Engineering* 66.1-2 (2014), pp. 395–414.

[12] C. Schär. "Control of a selective catalytic reduction process". Doctoral. Swiss Federal Institute of Technology, 2003.

[13] M Koebel, M Elsener and G Madia. "Reaction pathways in the selective catalytic reduction process with NO and NO_2 at low temperatures". In: *Industrial & Engineering Chemistry Research* 40 (2001), pp. 52–59.

[14] M. Koebel, G. Madia and M. Elsener. "Selective catalytic reduction of NO and NO_2 at low temperatures". In: *Catalysis Today* 73.3-4 (2002), pp. 239–247.

[15] P. R. Ettireddy and A. Kotrba. "Low Temperature SCR Catalysts Optimized for Cold-Start and Low-Load Engine Exhaust Conditions". In: *SAE Technical Paper 2015-01-1026* (2015).

[16] M. Colombo, I. Nova, E. Tronconi, V. Schmeißer and M. Weibel. "Mathematical modelling of cold start effects over zeolite SCR catalysts for exhaust gas aftertreatment". In: *Catalysis Today* 231.1 (2014), pp. 99–104.

[17] Y. Miao, L. D. Chen, Y. He and T. W. Kuo. "Study of SCR cold-start by energy method". In: *Chemical Engineering Journal* 155.1-2 (2009), pp. 260–265.

[18] D. Culbertson, M. Khair, S. Zhang, J. Tan and J. Spooler. "The Study of Exhaust Heating to Improve SCR Cold Start Performance". In: *SAE International Journal of Engines* 8.3 (2015).

[19] K. Chilumukuru, A. Gupta, M. Ruth, M. Cunningham, G. Kothandaraman, L. Cumaranatunge and H. Hess. "Aftertreatment Architecture and Control Methodologies for Future Light Duty Diesel Emission Regulations". In: *SAE International Journal of Engines* 10.4 (2017).

[20] O. P. Bhardwaj, K. Krishnamurthy, D. Blanco-Rodriguez, B. Holderbaum and T. Krfer. "Comparative Study to Assess the Potential of Different Exhaust Gas Aftertreatment Concepts for Diesel Powered Ultra-Light Commercial Vehicle Applications in View of Meeting BS VI Legislation". In: *Symposium on International Automotive Technology 2017*. SAE International, 2017.

[21] A. Auld, A. Ward, K. Mustafa and B. Hansen. "Assessment of Light Duty Diesel After-Treatment Technology Targeting Beyond Euro 6d Emissions Levels". In: *SAE International Journal of Engines* 10.4 (2017).

[22] F. Willems, R. Cloudt, E. van den Eijnden, M. van Genderen, R. Verbeek, B. de Jager, W. Boomsma and I. van den Heuvel. "Is Closed-Loop SCR Control Required to Meet Future Emission Targets?" In: *SAE World Congress and Exhibition*. SAE International, 2007.

[23] Q. Song and G. Zhu. "Model-based Closed-loop Control of Urea SCR Exhaust Aftertreatment System for Diesel Engine". In: *SAE 2002 World Congress and Exhibition*. SAE International, 2002.

[24] C. Schar, C Onder and H Geering. "Control of an SCR catalytic converter system for a mobile heavy-duty application". In: *IEEE Transactions on Control Systems Technology* 14.4 (2006), pp. 641–653.

[25] C. Ong, A. Annaswamy, I. V. Kolmanovsky, P. Laing and D. Reed. "An Adaptive Proportional Integral Control of a Urea Selective Catalytic Reduction System based on System Identification Models". In: *SAE International Journal of Fuels and Lubricants* 3.1 (2010), pp. 625–642.

[26] D. Upadhyay and M. Van Nieuwstadt. "Model based analysis and control design of a urea-SCR deNO$_x$ aftertreatment system". In: *Journal of Dynamic Systems, Measurement, and Control* 128.3 (2006), p. 737.

[27] M. Devarakonda, G. Parker, J. H. Johnson, V. Strots and S. Santhanam. "Model-Based Estimation and Control System Development in a Urea-SCR Aftertreatment System". In: *SAE International Journal of Fuels and Lubricants* 1.1 (2008), pp. 646–661.

[28] T. McKinley and A. Alleyne. "Adaptive model predictive control of an SCR catalytic converter system for automotive applications". In: *IEEE Transactions on Control Systems Technology* 20.6 (2012), pp. 1533–1547.

[29] C. Ciardelli, I. Nova, E. Tronconi, B. Konrad, D. Chatterjee, K. Ecke and M. Weibel. "SCR-DeNO$_x$ for diesel engine exhaust aftertreatment: unsteady-state kinetic study and monolith reactor modelling". In: *Chemical Engineering Science* 59.22-23 (2004), pp. 5301–5309.

[30] E. Tronconi, I. Nova, C. Ciardelli, D. Chatterjee, B. Bandl-Konrad and T. Burkhardt. "Modelling of an SCR catalytic converter for diesel exhaust after treatment: Dynamic effects at low temperature". In: *Catalysis Today* 105.3-4 (2005), pp. 529–536.

[31] L. Sharifian. "Simulation of NOx reduction over a Fe-Zeolite catalyst in an NH3-SCR system and calibration of the related parameters". Doctoral Dissertation. Swiss Federal Institute of Technology Zurich, 2011.

[32] C. Ericson. "Model based optimization of a complete diesel engine/SCR system". Doctoral Dissertation. Lund University, 2009. ISBN: 9789162877248.

[33] M.-F. Hsieh and J. Wang. "Development and experimental studies of a control-oriented SCR model for a two-catalyst urea-SCR system". In: *Control Engineering Practice* 19.4 (2011), pp. 409–422.

[34] M.-F. Hsieh and J. Wang. "Adaptive and efficient ammonia storage distribution control for a two-catalyst selective catalytic reduction system". In: *Journal of Dynamic Systems, Measurement, and Control* 134.1 (2012).

[35] S. Borchers, P. Rumschinski, S. Bosio, R. Weismantel and R. Findeisen. "A set-based framework for coherent model invalidation and parameter estimation of discrete time nonlinear systems". In: *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. 2009, pp. 6786–6792.

[36] H.-C. Cho, S. Takase, J.-H. Song and Y. Shimizu. "Impedancemetric NO_x Sensor Using YSZ-Based Solid Electrolyte Attached with Oxide Receptor". In: *14th International Meeting on Chemical Sensors*. 2012.

[37] J. Zou, Y. Zheng, J. Li, Z. Zhan and J. Jian. "Potentiometric NO_2 Sensors Based on Thin Stabilized Zirconia Electrolytes and Asymmetric $(La_{0.8}Sr_{0.2})_{0.95}MnO_3$ Electrodes". In: *Sensors* 15.7 (2015), pp. 17558–17571.

[38] M.-F. Hsieh and J. Wang. "An Extended Kalman Filter for NO_x Sensor Ammonia Cross- Sensitivity Elimination in Selective Catalytic Reduction Applications". In: *American Control Conference*. 2010, pp. 3033–3038.

[39] M Penza, M. A. Tagliente, L Mirenghi, C Gerardi, C Martucci and G Cassano. "Tungsten trioxide (WO_3) sputtered thin films for a NO_x gas sensor". In: *Sensors and Actuators* 50.x (1998), pp. 9–18.

[40] Q. Diao, X. Liang, J. Li, S. Yang, C. Yin and G. Lu. "Mixed-potential type NO_x sensor using stabilized zirconia and Cr_2O_3-WO_3 nanocomposites". In: *2011 16th International Solid-State Sensors, Actuators and Microsystems Conference*. 2011, pp. 146–149.

[41] B. T. Marquis and J. F. Vetelino. "A semiconducting metal oxide sensor array for the detection of NOx and NH3". In: *Sensors and Actuators B: Chemical* 77.1 (2001). Proceeding of the Eighth International Meeting on Chemical Sensors IMCS-8 - Part 2, pp. 100 –110.

[42] D. Y. Wang, S. Yao, M. Shost, J.-H. Yoo, D. Cabush, D. Racine, R. Cloudt and F. Willems. "Ammonia Sensor for Closed-Loop SCR Control". In: *SAE International Journal of Passenger Cars - Electronic and Electrical Systems* 1.1 (2008), pp. 323–333.

[43] Delphi. *Delphi Ammonia Sensor*. URL: `https://www.delphi.com/manufacturers/auto/sensors/engine-and-transmission/diesel/ammonia` (visited on 4th Aug. 2017).

[44] A. Frobert, S. Raux, Y. Creff and E. Jeudy. "About Cross-Sensitivities of NO_x Sensors in SCR Operation". In: *SAE 2013 World Congress and Exhibition*. SAE International, 2013.

[45] M.-F. Hsieh and J. Wang. "Observer-Based Estimation of Diesel Engine Aftertreatment System NO and NO_2 Concentrations". In: *Proceedings of the ASME 2010 Dynamic Systems and Control Conference* 2 (2010), pp. 1–8.

[46] H. S. Surenahalli, G. Parker and J. H. Johnson. "Extended Kalman Filter Estimator for NH_3 Storage, NO, NO_2 and NH_3 Estimation in a SCR". In: *SAE 2013 World Congress and Exhibition*. SAE International, 2013.

[47] R. Kalman. "A new approach to linear filtering and prediction problems". In: *Journal of Basic Engineering* 82.1 (1960), pp. 35–45.

[48] M. Hsieh and J Wang. "An extended Kalman filter for ammonia coverage ratio and capacity estimations in the application of Diesel engine SCR control and onboard diagnosis". In: *American Control Conference (ACC), 2010* (2010), pp. 5874–5879.

[49] J. Figura, J. Pekar, P. Krejza, D. Mracek, D. von Wissel and T. Zhang. "NO_2/NO_x Ratio and NH_3 Storage Estimation of Automotive SCR Multi-Brick Systems". In: *WCX 17: SAE World Congress Experience*. SAE International, 2017.

[50] M.-F. Hsieh and J Wang. "Observer-based estimation of selective catalytic reduction catalyst ammonia storage". In: *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 224.9 (2010), pp. 1199–1211.

[51] H. Zhang, J. Wang and Y.-Y. Wang. "Robust Filtering for Ammonia Coverage Estimation in Diesel Engine Selective Catalytic Reduction Systems". In: *Journal of Dynamic Systems, Measurement, and Control* 135.6 (2013), p. 064504.

[52] G. Hommen, F. Kupper and X. Seykens. "Robust , Model-Based Urea Dosing Control for SCR Aftertreatment Systems using a Cross-Sensitive Tailpipe NO_x Sensor". In: (2017). DOI: 10.4271/2017-01-0938.Copyright.

[53] Y. Y. Wang, H. Zhang and J. Wang. "NO_x Sensor Reading Correction in Diesel Engine Selective Catalytic Reduction System Applications". In: *IEEE/ASME Transactions on Mechatronics* 21.1 (2016), pp. 460–471.

[54] H. Zhang, J. Wang and Y.-Y. Wang. "Removal of NO_x sensor ammonia cross sensitivity from contaminated measurements in diesel-engine selective catalytic reduction systems". In: *Fuel* 150 (2015), pp. 448–456.

[55] A. Bonfils, Y. Creff, O. Lepreux and N. Petit. "Closed-loop control of a SCR system using a NO_x sensor cross-sensitive to NH_3". In: *Journal of Process Control* 24.2 (2014), pp. 368–378.

[56] J. Richalet, A. Rault, J. Testud and J. Papon. "Model predictive heuristic control". In: *Automatica* 14.5 (1978), pp. 413–428.

[57] C. Cutler and B. Ramaker. “Dynamic matrix control – A computer control algorithm”. In: *1980 Joint Automatic Control Conference, San Francisco, CA, USA*. 1980.

[58] E. N. Hartley and J. M. Maciejowski. “Predictive control for spacecraft rendezvous in an elliptical orbit using an FPGA”. In: *2013 European Control Conference (ECC)*. 2013, pp. 1359–1364.

[59] C. Göhrle, A. Wagner, A. Schindler and O. Sawodny. “Active suspension controller using MPC based on a full-car model with preview information”. In: *2012 American Control Conference (ACC)*. 2012, pp. 497–502.

[60] S. J. Qin and T. A. Badgwell. “A survey of industrial model predictive control technology”. In: *Control Engineering Practice* 11 (2003), pp. 733–764.

[61] J. M. Maciejowski. *Predictive Control with Constraints*. Prentice Hall/Pearson Education, 2002.

[62] D. Q. Mayne. “Model predictive control: Recent developments and future promise”. In: *Automatica* 50.12 (2014), pp. 2967–2986.

[63] A. Bemporad, M. Morari, V. Dua and E. Pistikopoulos. “The explicit linear quadratic regulator for constrained systems”. In: *Automatica* 38 (2002).

[64] R. Gondhalekar and J. ichi Imura. “Recursive feasibility guarantees in move-blocking MPC”. In: *2007 46th IEEE Conference on Decision and Control*. 2007, pp. 1374–1379.

[65] M. Cannon. *C21 Model Predictive Control*. 2015. URL: `http://www.eng.ox.ac.uk/~conmrc` (visited on 14th Aug. 2017).

[66] J. Löfberg. “Oops! I cannot do it again: Testing for recursive feasibility in MPC”. In: *Automatica* 48.3 (2012), pp. 550 –555. ISSN: 0005-1098.

[67] J. B. Rawlings, D. Angeli and C. N. Bates. “Fundamentals of economic model predictive control”. In: *Proceedings of 51st IEEE Conference on Decision and Control (CDC)* (2012), pp. 3851–3861.

[68] J. B. Rawlings, D. Bonne, J. B. Jørgensen, A. N. Venkat and S. B. Jørgensen. “Unreachable setpoints in model predictive control”. In: *IEEE Transactions on Automatic Control* 53.9 (2008), pp. 2209–2215.

[69] T. Faulwasser, M. Korda, C. N. Jones and D. Bonvin. “Turnpike and dissipativity properties in dynamic real-time optimization and economic MPC”. In: *53rd IEEE Conference on Decision and Control*. 2014, pp. 2734–2739.

[70] M. A. Müller, D. Angeli and F. Allgöwer. “On Necessity and Robustness of Dissipativity in Economic Model Predictive Control”. In: *IEEE Transactions on Automatic Control* 60.6 (2015), pp. 1671–1676.

[71] “A Lyapunov function for economic optimizing model predictive control”. In: *IEEE Transactions on Automatic Control* 56.3 (2011), pp. 703–707.

[72] D. Angeli, R. Amrit and J. B. Rawlings. "On Average Performance and Stability of Economic Model Predictive Control". In: *IEEE Transactions on Automatic Control* 57.7 (2012), pp. 1615–1626.

[73] L. Grüne. "NMPC without terminal constraints". In: *IFAC Proceedings Volumes* 45.17 (2012), pp. 1 –13.

[74] L. Grüne and M. Stieler. "A Lyapunov function for economic MPC without terminal conditions". In: *53rd IEEE Conference on Decision and Control*. 2014, pp. 2740–2745.

[75] T. Damm and L. Grüne. "An Exponential Turnpike Theorem for Dissipative Discrete Time Optimal Control Problems". In: *SIAM Journal on Control and Optimization* 52.3 (2014), pp. 1935–1957.

[76] S. Boyd and L. Vandenberghe. *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[77] M. S. K. Lau, S. P. Yue, K. V. Ling and J. M. Maciejowski. "A comparison of interior point and active set methods for FPGA implementation of model predictive control". In: *2009 European Control Conference (ECC)*. 2009, pp. 156–161.

[78] S. Koehler, C. Danielson and F. Borrelli. "A Primal-Dual Active-Set Method for Distributed Model Predictive Control". In: *American Control Conference (ACC), 2015* May (2015), pp. 4759–4764.

[79] L. A. Rodriguez and A. Sideris. "An active set method for constrained linear quadratic optimal control". In: *Proceedings of the 2010 American Control Conference*. 2010, pp. 5197–5202.

[80] M. Cannon, B. Kouvaritakis and J. A. Rossiter. "Efficient active set optimization in triple mode MPC". In: *IEEE Transactions on Automatic Control* 46.8 (2001), pp. 1307–1312.

[81] V. V. Naik, D. N. Sonawane, D. D. Ingole and D. Ginoya. "Model Predictive Control of DC Servomotor using Active Set Method". In: *2013 IEEE International Conference on Control Applications (CCA)* (2013), pp. 820–825.

[82] L. Zhong and F. Mora-Camino. "Aircraft Fault Tolerant Control Based on Active Set Method". In: *25th Chinese Control and Decision Conference* (2013), pp. 3516–3521.

[83] H. J. Ferreau, H. G. Bock and M Diehl. "An online active set strategy to overcome the limitations of explicit MPC". In: *International Journal of Robust and Nonlinear Control* 18.8 (2008), pp. 816–830.

[84] K. Edlund, L. E. Sokoler and J. B. Jørgensen. "A primal-dual interior-point linear programming algorithm for MPC". In: *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. 2009, pp. 351–356.

[85] L. E. Sokoler, A. Skajaa, G. Frison, R. Halvgaard and J. B. Jørgensen. "A warm-started homogeneous and self-dual interior-point method for linear economic model predictive control". In: *52nd IEEE Conference on Decision and Control*. 2013, pp. 3677–3683.

[86] E. Klintberg and S. Gros. "A Parallelizable Interior Point Method for Two-Stage Robust MPC". In: *IEEE Transactions on Control Systems Technology* PP.99 (2017), pp. 1–11.

[87] J Liu, H Peyrl, A Burg and G. A. Constantinides. "FPGA implementation of an interior point method for high-speed model predictive control". In: *2014 24th International Conference on Field Programmable Logic and Applications (FPL)* (2014), pp. 1–8.

[88] R. S. Vadamalu and C. Beidl. "Online MPC based PHEV Energy Management using conic interior-point methods". In: *2016 IEEE Intelligent Vehicles Symposium (IV)*. 2016, pp. 466–471.

[89] A. Tatulea-Codrean, D. Haßkerl, M. Urselmann and S. Engell. "Steady-state optimization and nonlinear model-predictive control of a reactive distillation process using the software platform do-mpc". In: *2016 IEEE Conference on Control Applications (CCA)*. 2016, pp. 1513–1518.

[90] J. L. Jerez, P. J. Goulart, S. Richter, G. A. Constantinides, E. C. Kerrigan and M. Morari. "Embedded Predictive Control on an FPGA using the Fast Gradient Method". In: *2013 European Control Conference (ECC)* (2013), pp. 3614–3620.

[91] M. Pereira, D. Limon, D. M. de la Pea and T. Alamo. "MPC implementation in a PLC based on Nesterov's fast gradient method". In: *2015 23rd Mediterranean Conference on Control and Automation (MED)*. 2015, pp. 371–376.

[92] L. Ferranti and T. Keviczky. "A Parallel Dual Fast Gradient Method for MPC Applications". In: *2015 54th IEEE Conference on Decision and Control (CDC)* (2015), pp. 2406–2413.

[93] J. Betts. *Practical Methods for Optimal Control and Estimation Using Nonlinear Programming*. Second. Society for Industrial and Applied Mathematics, 2010.

[94] M. Diehl, H. J. Ferreau and N. Haverbeke. "Efficient Numerical Methods for Nonlinear MPC and Moving Horizon Estimation". In: *Nonlinear Model Predictive Control: Towards New Challenging Applications*. Ed. by L. Magni, D. M. Raimondo and F. Allgöwer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 391–417.

[95] M. R. Kristensen, J. B. Jørgensen, P. G. Thomsen and S. B. Jørgensen. "An ESDIRK method with sensitivity analysis capabilities". In: *Computers and Chemical Engineering* 28.12 (2004), pp. 2695–2707.

[96] M. Diehl, H. Bock, J. P. Schlöder, R. Findeisen, Z. Nagy and F. Allgöwer. "Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations". In: *Journal of Process Control* 12.4 (2002), pp. 577–585.

[97] R. Quirynen, M. Vukov and M. Diehl. "Auto Generation of Implicit Integrators for Embedded NMPC with Microsecond Sampling Times". In: *IFAC Proceedings Volumes* 45.17 (2012). 4th IFAC Conference on Nonlinear Model Predictive Control, pp. 175–180.

[98] R. Quirynen, S. Gros and M. Diehl. "Fast auto generated ACADO integrators and application to MHE with multi-rate measurements". In: *2013 European Control Conference (ECC)*. 2013, pp. 3077–3082.

[99] J. V. Frasch, S. Sager and M. Diehl. "A parallel quadratic programming method for dynamic optimization problems". In: *Mathematical Programming Computation* 7.3 (2015), pp. 289–329.

[100] A. Domahidi. *FORCES*. 2012. URL: `http://forces.ethz.ch/` (visited on 14th Aug. 2017).

[101] A. Wächter and L. T. Biegler. "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming". English. In: *Mathematical Programming* 106.1 (2006), pp. 25–57.

[102] B. Houska, H. J. Ferreau and M. Diehl. "An auto-generated real-time iteration algorithm for nonlinear MPC in the microsecond range". In: *Automatica* 47.10 (2011), pp. 2279–2285.

[103] P. Falugi, E. Kerrigan and E. van Wyk. *ICLOCS: Imperial College London optimal control software (user guide)*. 2010.

[104] A. Smeti, W. Chagra and M. Ksouri. "Implementation of a Predictive controller on the STM32 board". In: *2013 International Conference on Control, Decision and Information Technologies (CoDIT)*. 2013, pp. 250–253.

[105] T. A. Johansen, W. Jackson, R. Schreiber and P. Tondel. "Hardware Synthesis of Explicit Model Predictive Controllers". In: *IEEE Transactions on Control Systems Technology* 15.1 (2007), pp. 191–197.

[106] P. Dua, V. Sakizlis, L. S. Kershenbaum and E. N. Pistikopoulos. "Model based parametric controller for the operation of an experimental reactor". In: *Computer Aided Chemical Engineering* 18 (2004). European Symposium on Computer-Aided Process Engineering-14, pp. 637 –642.

[107] E. Kerrigan, J. Jerez, G. Constantinides and K.-V. Ling. "Model predictive control for deeply pipelined field-programmable gate array implementation: algorithms and circuitry". In: *IET Control Theory & Applications* 6.8 (2012), pp. 1029–1041.

[108] K. V. Ling, J. Maciejowski, A. Richards and B. F. Wu. “Multiplexed model predictive control”. In: *Automatica* 48.2 (2012), pp. 396–401.

[109] K. Ling, B. Wu and J. Maciejowski. “Embedded Model Predictive Control (MPC) using a FPGA”. In: *IFAC Proceedings Volumes* 41.2 (2008). 17th IFAC World Congress, pp. 15250–15255.

[110] B. Stellato and P. J. Goulart. “Real-time FPGA implementation of direct MPC for power electronics”. In: *2016 IEEE 55th Conference on Decision and Control* (2016), pp. 1471–1476.

[111] A. Suardi, E. C. Kerrigan and G. A. Constantinides. “Fast FPGA prototyping toolbox for embedded optimization”. In: *2015 European Control Conference* (2015), pp. 2589–2594.

[112] F. Xu, H. Chen, W. Jin and Y. Xu. “FPGA implementation of nonlinear model predictive control”. In: *The 26th Chinese Control and Decision Conference (2014 CCDC)*. 2014, pp. 108–113.

[113] Y. Shoukry, M. W. El-Kharashi and S. Hammad. “An embedded implementation of the Generalized Predictive Control algorithm applied to automotive active suspension systems”. In: *Computers & Electrical Engineering* 39.2 (2013), pp. 512–529.

[114] N. Rajaei, X. Han, X. Chen and M. Zheng. “Model Predictive Control of Exhaust Gas Recirculation Valve”. In: *SAE Technical Paper*. SAE International, Apr. 2010.

[115] M. Iwadare, M. Ueno and S. Adachi. “Multi-Variable Air-Path Management for a Clean Diesel Engine Using Model Predictive Control”. In: *SAE Int. J. Engines* 2 (Apr. 2009), pp. 764–773.

[116] D. Gagliardi, T. Othsuka and L. del Re. “Direct C/GMRES Control of The Air Path of a Diesel Engine”. In: *IFAC Proceedings Volumes* 47.3 (2014). 19th IFAC World Congress, pp. 3000–3005.

[117] H. Borhan, G. Kothandaraman and B. Pattel. “Air handling control of a diesel engine with a complex dual-loop EGR and VGT air system using MPC”. In: *2015 American Control Conference (ACC)*. 2015, pp. 4509–4516.

[118] N. Khaled, M. Cunningham, J. Pekar, A. Fuxman and O. Santin. “Multivariable Control of Dual Loop EGR Diesel Engine with a Variable Geometry Turbo”. In: *SAE Technical Paper*. SAE International, Apr. 2014.

[119] Q. Zhu, R. Prucka, M. Prucka and H. Dourra. “A Nonlinear Model Predictive Control Strategy with a Disturbance Observer for Spark Ignition Engines with External EGR”. In: *SAE International Journal of Commercial Vehicles* 10 (Mar. 2017), pp. 360–372.

[120] T. Albin, D. Ritter, D. Abel, N. Liberda, R. Quirynen and M. Diehl. "Nonlinear MPC for a two-stage turbocharged gasoline engine airpath". In: *2015 54th IEEE Conference on Decision and Control (CDC)*. 2015, pp. 849–856.

[121] R. Cloudt and F. Willems. "Integrated Emission Management strategy for cost-optimal engine-aftertreatment operation". In: *SAE International Journal of Engines* 4.1 (2011), pp. 1784–1797.

[122] P. Mentink, F. Willems, F. Kupper and E. V. den Eijnden. "Experimental Demonstration of a Model-Based Control Design and Calibration Method for Cost Optimal Euro-VI Engine-Aftertreatment Operation". In: *SAE Technical Paper*. SAE International, Apr. 2013.

[123] M. Cheng, L. Feng and B. Chen. "Nonlinear Model Predictive Control of a Power-Split Hybrid Electric Vehicle with Electrochemical Battery Model". In: *SAE Technical Paper*. SAE International, Mar. 2017.

[124] H. Wang, Y. Huang, A. Khajepour, H. He and C. Lv. "MPC-based power management strategy for a series hybrid electric tracked bulldozer". In: *2017 IEEE International Conference on Industrial Technology (ICIT)* (2017), pp. 1403–1408.

[125] K. Olia, M. Shahverdi, M. Mazzola and A. Sherif. "Developing a Model Predictive Control-Based Algorithm for Energy Management System of the Catenary-Based Electric Truck". In: *SAE Technical Paper*. SAE International, Oct. 2016.

[126] M. Shahverdi, M. Mazzola, S. Abdelwahed, M. Doude and D. Zhu. "MPC-based power management system for a plug-in hybrid electric vehicle for relaxing battery cycling". In: *2016 IEEE Transportation Electrification Conference and Expo, ITEC 2016* (2016), pp. 1–6.

[127] U. Sartori, F. Biral, E. Bertolazzi and S. Onori. "On-Line power management optimization of a hybrid electric vehicle with Non linear MPC and battery recharge equivalent cost". In: *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*. 2016, pp. 5094–5100.

[128] V. Puig, R. Costa-Castell and J. L. Sampietro. "Economic MPC for the energy management of hybrid vehicles including fuel cells and supercapacitors". In: *2016 UKACC 11th International Conference on Control (CONTROL)*. 2016, pp. 1–6.

[129] J. Buerger and M. Cannon. "Nonlinear MPC for supervisory control of hybrid electric vehicles". In: *2016 European Control Conference (ECC)*. 2016, pp. 135–140.

[130] A. Chandrasekar, S. Sengupta, S. Hingane, C. Gururaja and S. Pandit. "Comparative Analysis of Model Predictive Control (MPC) and Conventional Control in Supervisory Controller of a Retrofit HEV". In: *SAE Technical Paper*. SAE International, Jan. 2017.

[131] T. L. McKinley and A. G. Alleyne. "Model Predictive Control: A Unified Approach for Urea-Based SCR Systems". In: *SAE International Journal of Fuels and Lubricants* 3.1 (2010), pp. 673–689.

[132] C.-j. Chiang, C.-l. Kuo, C.-c. Huang and J.-y. Lee. "Model predictive control of SCR aftertreatment system". In: *2010 5th IEEE Conference on Industrial Electronics and Applications* (2010), pp. 2058–2063.

[133] P. Chen and J. Wang. "Integrated diesel engine and selective catalytic reduction system active NO_x control for fuel economy improvement". In: *2013 American Control Conference*. 2013, pp. 2196–2201.

[134] P. Chen and J. Wang. "Estimation and adaptive nonlinear model predictive control of selective catalytic reduction systems in automotive applications". In: *Journal of Process Control* 40 (2016), pp. 78–92.

[135] S. Stadlbauer, H. Waschl and L. del Re. "Adaptive SCR Model for MPC Control Including Aging Effects". In: *SAE Technical Paper*. SAE International, Apr. 2015.

[136] P. Chen and J. Wang. "Coordinated Active Thermal Management and Selective Catalytic Reduction Control for Simultaneous Fuel Economy Improvement and Emissions Reduction During Low-Temperature Operations". In: *Journal of Dynamic Systems, Measurement, and Control* 137.12 (2015), p. 121001.

[137] Y. Liu, J. Zhou, L. Fiorentini, M. Canova and Y. Y. Wang. "Model Predictive Control of a two-stage turbocharged Diesel engine air-path system for rapid catalyst warm-up". In: *2015 European Control Conference, ECC 2015* (2015), pp. 123–128.

[138] J. Zhao and J. Wang. "Integrated Model Predictive Control of Hybrid Electric Vehicles Coupled With Aftertreatment Systems". In: *IEEE Transactions on Vehicular Technology* 65.3 (2016), pp. 1199–1211.

[139] Transport for London. *Emissions from the TfL Bus Fleet*. Tech. rep. July. 2015.

[140] Transport for London. *In-service emissions performance of Euro 6 / VI vehicles*. Tech. rep. 2015. URL: `http://content.tfl.gov.uk/in-service-emissions-performance-of-euro-6vi-vehicles.pdf`.

[141] Transport for London. *London Bus Fleet Audit*. Tech. rep. March. 2016, pp. 1–6.

[142] Department for Environment Food & Rural Affairs. *Improving air quality in the UK Tackling nitrogen dioxide in our towns and cities*. Tech. rep. December. 2015.

[143] OpenStreetMap Contributors. *Planet dump retrieved from http://planet.osm.org*. 2017. URL: `http://www.openstreetmap.org`.

[144] S. H. Oh and J. C. Cavendish. "Transients of monolithic catalytic converters: Response to step changes in feedstream temperature as related to controlling automobile emissions". In: *Industrial & Engineering Chemistry Product Research and Development* 21.1 (1982), pp. 29–37.

[145] A. Lubeski. “Real-Time Catalytic Converter Temperature Estimator in the Powertrain Controller”. In: *SAE Technical Paper*. SAE International, Mar. 2000.

[146] A. G. Konstandopoulos, M. Kostoglou, E. Skaperdas, E. Papaioannou, D. Zarvalis and E. Kladopoulou. “Fundamental Studies of Diesel Particulate Filters: Transient Loading, Regeneration and Aging”. In: *SAE Technical Paper*. SAE International, Mar. 2000.

[147] I. Kolmanovsky, E. Garone and S Di Cairano. “Reference and command governors: A tutorial on their theory and automotive applications”. In: *2014 American Control Conference (ACC)*. 2014, pp. 226–241.

[148] H. Nakada, P. Martin, G. Milton, A. Iemura and A. Ohata. “An application study of online reference governor to boost pressure control for automotive diesel engines”. In: *2014 American Control Conference*. 2014, pp. 3135–3140.

[149] H. Nakada, G. Milton, P. Martin, A. Iemura and A. Ohata. “Application of Reference Governor Using Soft Constraints and Steepest Descent Method to Diesel Engine Aftertreatment Temperature Control”. In: *SAE Int. J. Engines* 6 (Apr. 2013), pp. 257–266.

[150] A. K. Kaw, E. K. Kalu and D. Nguyen. *Numerical Methods with Applications*. 2nd ed. University of South Florida, 2011.

[151] J. Sowman, D. S. Laila, A. J. Cruden, P. Fussey and A. Truscott. “A predictive control approach to diesel Selective Catalytic Reduction”. In: *2015 European Control Conference (ECC)*. 2015, pp. 3073–3078.

[152] L. Grüne, C. M. Kellett and S. R. Weller. “On a discounted notion of strict dissipativity”. In: *IFAC-PapersOnLine* 49.18 (2016). 10th IFAC Symposium on Nonlinear Control Systems NOLCOS 2016, pp. 247–252.

[153] D. Simon. *Optimal State Estimation - Kalman, H-Infinity and Nonlinear Approaches*. Wiley-Interscience, 2006.

[154] N. J. Gordon, D. J. Salmond and A. F. M. Smith. “Novel approach to nonlinear/non-Gaussian Bayesian state estimation”. In: *IEE Proceedings F - Radar and Signal Processing* 140.2 (1993), pp. 107–1–13.

[155] J. A. Brown and D. W. Capson. “A framework for 3D model-based visual tracking using a gpu-accelerated particle filter”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.1 (2012), pp. 68–80.

[156] L. M. Murray, A. Lee and P. E. Jacob. “Parallel Resampling in the Particle Filter”. In: *Journal of Computational and Graphical Statistics* 25.3 (2016), pp. 789–805.

[157] D. S. Laila and E. Grünbacher. “Nonlinear output feedback and periodic disturbance attenuation for setpoint tracking of a combustion engine test bench”. In: *Automatica* 64 (2016), pp. 29–36.

[158] M. W. Spong, P. Corke and R. Lozano. “Nonlinear control of the Reaction Wheel Pendulum”. In: *Automatica* 37 (2001), pp. 1845–1851.

[159] O. Boubaker. “The inverted pendulum benchmark in nonlinear control theory: A survey”. In: *International Journal of Advanced Robotic Systems* 10 (2013).

[160] K. J. Åström and K. Furuta. “Swinging up a pendulum by energy control”. In: *Automatica* 36 (2000), pp. 287–295.

[161] K. Ling, P. Falugi, J. Maciejowski and L. Chisci. “Robust predictive control of the Furuta pendulum”. In: *IFAC Proceedings Volumes* 35.1 (2002). 15th IFAC World Congress, pp. 37 –42.

[162] S. Durand, F. G. Castellanos, N. Marchand and W Ferm. “Event-Based Control of the Inverted Pendulum: Swing up and Stabilization”. In: *Journal of Control Engineering and Applied Informatics* 15.3 (2013), pp. 96–104.

[163] A. Mills, A. Wills and B. Ninness. “Nonlinear model predictive control of an inverted pendulum”. In: *2009 American Control Conference* (2009), pp. 2335–2340.

[164] J. She, A. Zhang, X. Lai and M. Wu. “Global stabilization of 2-DOF underactuated mechanical systems-an equivalent-input-disturbance approach”. In: *Nonlinear Dynamics* 69 (2012), pp. 495–509.

[165] R. Olfati-Saber. “Global stabilization of a flat underactuated system: the inertia wheel pendulum”. In: *Proceedings of the 40th IEEE Conference on Decision and Control*. Vol. 4. 2001, 3764–3765 vol.4.

[166] T. A. Johansen. “On multi-parametric nonlinear programming and explicit nonlinear model predictive control”. In: *Proceedings of the 41st IEEE Conference on Decision and Control* 3 (2002), pp. 2768–2773.

[167] P. Tøndel, T. Johansen and A. Bemporad. “Evaluation of piecewise affine control via binary search tree”. In: *Automatica* 39 (2003), pp. 945–950.

[168] R. Ortega, M. Spong, F. Gomez-Estern and G. Blankenstein. “Stabilization of a class of underactuated mechanical systems via interconnection and damping assignment”. In: *IEEE Transactions on Automatic Control* 47.8 (2002), pp. 1218–1233.

[169] ST Microelectronics. *Discovery kit for STM32F407/417 line*. 2014. URL: `http://www.st.com/st-web-ui/static/active/en/resource/technical/document/data_brief/DM00037955.pdf` (visited on 14th Aug. 2017).

[170] G. G. Dahlquist. “A special stability problem for linear multistep methods”. In: *BIT Numerical Mathematics* 3.1 (1963), pp. 27–43.

[171] A. Stuart and A. Humphries. *Dynamical Systems and Numerical Analysis*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 1998.

[172] M. Diehl, H. G. Bock and J. P. Schloder. "Real-Time Iterations for Nonlinear Optimal Feedback Control". In: *Proceedings of the 44th IEEE Conference on Decision and Control*. 2005, pp. 5871–5876.