ABSTRACT

Title of dissertation: POSITIVE FILTERED P_N METHOD FOR LINEAR TRANSPORT EQUATIONS AND THE ASSOCIATED OPTIMIZATION ALGORITHM

Ming Tse Paul Laiu, Doctor of Philosophy, 2016

Dissertation directed by: Professor André Tits
Department of Electrical and Computer Engineering

We propose a positive, accurate moment closure for linear kinetic transport equations based on a filtered spherical harmonic (FP_N) expansion in the angular variable. The FP_N moment equations are accurate approximations to linear kinetic equations, but they are known to suffer from the occurrence of unphysical, negative particle concentrations. The new positive filtered P_N (FP_N^+) closure is developed to address this issue.

The FP_N^+ closure approximates the kinetic distribution by a spherical harmonic expansion that is non-negative on a finite, predetermined set of quadrature points. With an appropriate numerical PDE solver, the FP_N^+ closure generates particle concentrations that are guaranteed to be non-negative. Under an additional, mild regularity assumption, we prove that as the moment order tends to infinity, the FP_N^+ approximation converges, in the L^2 sense, at the same rate as the FP_N approximation; numerical tests suggest that this assumption may not be necessary. By numerical experiments on the challenging line source benchmark problem, we

confirm that the FP_N^+ method indeed produces accurate and non-negative solutions.

To apply the FP_N^+ closure on problems at large temporal-spatial scales, we develop a positive asymptotic preserving (AP) numerical PDE solver. We prove that the propose AP scheme maintains stability and accuracy with standard mesh sizes at large temporal-spatial scales, while, for generic numerical schemes, excessive refinements on temporal-spatial meshes are required. We also show that the proposed scheme preserves positivity of the particle concentration, under some time step restriction. Numerical results confirm that the proposed AP scheme is capable for solving linear transport equations at large temporal-spatial scales, for which a generic scheme could fail.

Constrained optimization problems are involved in the formulation of the FP_N^+ closure to enforce non-negativity of the FP_N^+ approximation on the set of quadrature points. These optimization problems can be written as strictly convex quadratic programs (CQPs) with a large number of inequality constraints. To efficiently solve the CQPs, we propose a constraint-reduced variant of a Mehrotra-predictor-corrector algorithm, with a novel constraint selection rule. We prove that, under appropriate assumptions, the proposed optimization algorithm converges globally to the solution at a locally q-quadratic rate. We test the algorithm on randomly generated problems, and the numerical results indicate that the combination of the proposed algorithm and the constraint selection rule outperforms other compared constraint-reduced algorithms, especially for problems with many more inequality constraints than variables.

POSITIVE FILTERED P_N METHOD FOR LINEAR TRANSPORT EQUATIONS AND THE ASSOCIATED OPTIMIZATION ALGORITHM

by

Ming Tse Paul Laiu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:
Professor André Tits, Chair/Advisor
Dr. Cory Hauck, Co-Advisor
Professor Dianne O'Leary
Professor Isaak Mayergoyz
Professor Konstantina Trivisa
Professor Howard Elman

Dedication

To Yun,

and to our daughter, Phoebe.

Acknowledgments

I thank my advisor, André Tits, for his advice and support throughout my graduate studies. I learned a lot from his unlimited attentiveness and patience, for which I truly appreciate.

I thank Cory Hauck for serving as my co-advisor and introducing me to the fields of numerical PDE and kinetic theory. His trust and guidance helped me get through the transition from engineering to mathematics.

I thank Dianne O'Leary for many efficient, concise, and constructive discussions on my work. This work would not be completed without her numerous invaluable suggestions.

I thank Professors Howard Elman, Issak Mayergoyz, and Konstantina Trivisa for serving on my committee. Their inspiring questions and comments made my defense an enjoyable experience.

I thank Graham Alldredge for his assistance in understanding the moment methods during my early graduate career.

I thank Kristopher Garrett for his help in the implementation of the kinetic scheme used in Chapter 2.

I thank Marcos Vasconcelos for many meaningful conversations in our office.

I thank Richard Barnard, Zheng Chen, Eirik Endeve, Vincent Heningburg, Qiwei Sheng, Hoang Tran, and all my other colleagues in ORNL for their support in the final stages of my graduate studies.

Finally, I thank my parents for everything they have done for me.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

In this dissertation, we propose a positive, accurate moment closure for linear kinetic transport equations, develop an asymptotic preserving (AP) numerical PDE solver for the proposed closure, and introduce an efficient optimization algorithm for solving optimization problems that arise in the application of the proposed closure.

Kinetic transport equations are used to model particle-based systems in various areas including rarefied gases [1, 2], radiative transport [3–5], and semiconductors [6]. These equations govern the evolution of a positive scalar function, the kinetic distribution, that depends on position, momentum, and time. While the kinetic distribution gives microscopic information of the particles, solving kinetic equations is computationally expensive in general. In typical settings, the position-momentum phase space is six-dimensional. This makes the numerical simulation of these equations difficult.

Moment methods are commonly used to approximate the solution of kinetic equations. These methods track a finite number of moments (or weighted averages) of the kinetic distribution with respect to the momentum variable. Equations to describe the evolution of these moments are derived directly from the kinetic equation. However, for any finite number of moments, the exact moment equations are not

closed, i.e., they require additional information about the kinetic distribution that is lost when retaining only a finite number of moments. Hence a moment closure is needed to fill in the missing kinetic information and close the system of equations. Since moment closures are expected to provide accurate approximation to the kinetic distribution, optimization is involved in the formulation of moment closures in many previous works [7–11].

In this dissertation, we consider linear kinetic equations with a momentum variable that specifies the traveling direction of unit speed particles by an angle on the unit sphere. In this setting, the most common moment closure method is the spherical harmonic approximation, or P_N method [3,12]. This method is equivalent to a standard spectral discretization of the kinetic equation with respect to the momentum variable. The finite expansion of the kinetic distribution in spherical harmonics provides the necessary closure, and the coefficients of the expansion are related to the tracked moments via an explicit linear mapping.

Although computationally fast, the P_N method suffers from several well-known drawbacks. Like most spectral methods, it may produce highly oscillatory solutions that can lead to local negative values in the particle concentration.[1] Several moment closures have been proposed to enforce non-negativity on the particle concentration. The M_N [7,8,13,14] and PP_N [9,10] closures were proposed to maintain the positivity of solutions by using a positive ansatz for the closure. This is in contrast to the

[1] In this dissertation, the term "concentration" is used when referring to the integral of the kinetic distribution with respect to angle. The concentration is a function of position and time only.

spherical harmonic expansion for the P_N method, which may take on negative values. However, both the M_N and PP_N solutions are still quite oscillatory [9, 10] and furthermore are much more expensive than P_N [15–17]. The recently proposed FP_N closure [11,18] still uses a spherical harmonics expansion, but damps the oscillations via a low pass filter on the moments. While the filter mitigates the occurrence of negative particle concentrations, they are not fully removed. Small negative values in the particle concentration may not affect stability or accuracy when solving linear kinetic models, but for nonlinear models, negative concentrations may make the system unstable.[2] Hence, it is of interest to develop a positive-preserving[3] modification of the FP_N method.

In the work presented in this dissertation, we develop a positive, accurate moment closure based on the FP_N closure. We refer it as the positive filtered P_N (FP_N^+) closure. The FP_N^+ closure approximates the kinetic distribution by a spherical harmonic expansion that is non-negative on a finite, predetermined set of quadrature points. With an appropriate numerical PDE solver, the FP_N^+ closure generates particle concentrations that are guaranteed to be non-negative. Numerical experiments on the challenging line source benchmark problem confirm that the FP_N^+ method indeed produces accurate and non-negative solutions. To apply the FP_N^+ closure on problems at large temporal-spatial scales, we propose a positive asymptotic pre-

[2]For example, when solving radiative transfer equations coupled with a material equation, the negative radiative energy-density can cause a negative material temperature [19, 20].

[3]In this dissertation, the term "positive-preserving" refers to methods that maintain the non-negativity of particle concentration.

serving numerical PDE solver. We prove that the proposed AP solver is efficient and positive-preserving at large temporal-spatial scales, and verify such properties with numerical results. Constrained optimization problems are involved in the formulation of the FP_N^+ closure to enforce non-negativity of the FP_N^+ approximation on the set of quadrature points. To efficiently solve these problems, we propose a constraint-reduced variant of a Mehrotra-predictor-corrector algorithm, with a novel, powerful constraint selection rule. We prove the convergence properties of the proposed optimization algorithm, including global convergence and local convergence rate. The algorithm is tested on randomly generated problems and on the FP_N^+ method. The efficiency of the proposed algorithm is confirmed by the numerical results.

In Chapter 2, we propose the FP_N^+ closure, a modification of the FP_N closure that preserves non-negativity on a set of quadrature points. This set is part of a quadrature rule that is used to evaluate exactly (up to roundoff errors) the moments of the spherical harmonic expansion, up to a given order. As shown in [15], this condition is sufficient to maintain a non-negative particle concentration.

Implementation of the FP_N^+ method requires a PDE solver to update the moment system in time. In Chapter 2, we first test the FP_N^+ method on the second-order finite volume kinetic scheme developed in [15]; see also [10]. A more advanced PDE solver with the "asymptotic preserving" property is proposed in Chapter 3.

The FP_N^+ method requires the solution of a constrained optimization problem to define the closure. The optimization problem can be written as a strictly convex quadratic program (CQP) with a large number of inequality constraints that enforce

positivity on the prescribed quadrature. The optimization provides an accurate ansatz for the FP_N^+ closure, which plays an important role in our analysis on the consistency properties. Under an additional, mild regularity assumption, we prove that as the moment order tends to infinity, the FP_N^+ approximation converges to the underlying target function, in the L^2 sense, as fast as the FP_N approximation. Furthermore, our numerical results show that this property holds even without the additional assumption. For comparison, we also analyze and test the consistency properties of another positive-preserving closure, which we refer to as the uniform damping (UD_N) closure. This closure was originally proposed in [21] to generate spatial reconstructions in the numerical simulation of hyperbolic conservation laws. More recently, it was applied to finite volume, weighted essentially non-oscillatory (WENO) and discontinuous Galerkin schemes in [22]. Because of its simplicity and fast implementation, the method has been applied in a variety of applications; see [23] for review and further references. We prove convergence results for the UD_N closure that are suboptimal when compared to the FP_N closure; numerical tests suggest that our estimates are likely sharp. For smooth problems, the difference in the accuracy of the closures is negligible. However, for problems with less regularity, the difference is substantial.

To conclude Chapter 2, we compute the numerical solution with the FP_N^+ method on the line source benchmark problem [24] and compare it to solutions from the P_N, FP_N, PP_N, and UD_N methods. For the same number of moments, the FP_N^+ method performs much better than the UD_N method. However, enforcing positivity does create some local trade-offs in accuracy when compared to the FP_N method. In

terms of accuracy, the P_N and PP_N methods are not competitive. We also compare the efficiency of the more accurate FP_N^+ closure with the less expensive UD_N closure. In particular, we consider the solution time needed to reach a given level of accuracy in the particle concentration. For the line source problem, we conclude that the FP_N^+ solutions are generally two to ten times faster than the UD_N solutions to reach the same accuracy.

In Chapter 3, an "asymptotic preserving" scheme (see, for example, [25]) is proposed for solving the moment systems of linear transport equations. While the second-order finite volume kinetic scheme in [10, 15] performs well in the transport regime of the kinetic equation, one major drawback is that it is not efficient at large temporal-spatial scales due to the multi-scale behavior of kinetic equations, as discussed in [10]. Specifically, at large temporal-spatial scales, the particle concentration is governed by a diffusion model (referred as "diffusion limit") [26, 27], while the numerical solution of kinetic equations requires the resolution of the much smaller hyperbolic and collisional time scales. Resolving the small time scales dictates a very fine temporal-spatial mesh to maintain stability and accuracy of generic numerical solvers, including the one proposed in [10], hence making the computation cost prohibitive.

Several numerical PDE solvers, e.g. [28–30], have been introduced to address this issue, and they are referred as asymptotic preserving schemes in [31]. The AP schemes preserve stability and accuracy when applied to equations at large temporal-spatial scales, without resolving the size of the temporal-spatial mesh. For transport equations, the AP schemes often require decompositions on the distribution func-

tion, such as even-odd decomposition [32–34], or macro-micro decomposition [35]. Unfortunately, most of the existing AP schemes are either only of first-order accuracy, or do not preserve positivity and hence produce solutions with unphysical negative distribution.

Our goal is to develop a positive preserving AP scheme for the moment equations. The scheme is expected to maintain stability and accuracy with standard mesh sizes near the diffusion limit, while still preserving positivity of the particle concentration. We develop the proposed AP scheme based on the even-odd decomposition approach. We prove that the proposed AP scheme indeed computes the correct diffusion limit at large temporal-spatial scales. We also show that the proposed scheme preserves positivity of the particle concentration, under some hyperbolic time step size restriction. In practice, we only impose the hyperbolic time step restriction when the particle concentration becomes negative, and we prove that a more relaxed parabolic time step size restriction is sufficient to maintain positivity of the particle concentration, under a mild assumption. Results of numerical experiments on the comparison of the proposed AP scheme and the second-order kinetic scheme used in [10,15] shows that the proposed AP scheme is indeed capable for solving the linear transport equation at large temporal-spatial scales, while the second-order kinetic solver in [10] fails.

In Chapter 4, we propose an efficient optimization algorithm for solving the CQPs that arise from the FP_N^+ method, where the CQPs are used to generate positive approximations to the kinetic distribution on a prescribed quadrature set. Primal-dual interior-point methods (PDIPMs) are commonly used to solve CQPs.

(See, e.g., [36].) In standard interior-point methods, the main computing work involves forming a "normal" matrix (or Schur complement matrix) for solving the search direction, which is at a cost proportional to the number of constraints. In the FP_N^+ method, the CQPs contain a large number of inequality constraints, which makes the computational cost of standard interior-point methods prohibitive. Hence, we apply a constraint reduction (CR) technique to reduce the computational burden.

In the context of PDIPM, constraint reduction proved to be an effective technique for solving various types of optimization problems with a large number of inequality constraints. The constraint reduction schemes use an approximate normal matrix (or Schur complement matrix) when solving linear [37–39], convex quadratic [40, 41], or semidefinite programs [42, 43]. With such schemes, only a small portion of constraints are active at the solution, and the other constraints are, in some sense, redundant. The constraint reduction mechanism uses only a wisely selected small subset of all constraints, referred as the "working set," to compute the approximate search direction, which significantly reduces the computation time for problems with many inequality constraints. The guidelines or rules used to choose the working set are referred as "constraint selection rules."

The proposed algorithm is a constraint-reduced variant of a Mehrotra-predictor-corrector (MPC) algorithm. (See [44] for Mehrotra's original MPC algorithm.) The main difference between MPC algorithms and other interior-point algorithms is that, at each iteration, MPC algorithms compute two different directions, predictor (also referred as "affine-scaling" direction) and centering-corrector, while the generic

interior-point algorithms only use one direction. The additional centering-corrector direction points towards the central path, and provides a second order correction on the search direction. Thus, for most problems, the MPC algorithm usually requires fewer iteration count than generic interior-point methods. Since the two directions are computed by using the same normal matrix, the computational cost per iteration of MPC algorithm is only slightly more than that of generic interior-point algorithms for problems with many inequality constraints. In general, the additional cost per iteration is largely compensated by the reduction of iteration counts, which leads to a faster convergence to the optimal solution.

A constraint-reduced variant of a Mehrotra-predictor-corrector algorithm is proposed in [37] for solving linear programs, and is extended in [41] specifically for solving CQPs from training support vector machines. We generalize the algorithm proposed in [41] so that it can be applied to general CQPs, and we refer the proposed algorithm as Algorithm CR-MPC. We also prove that, with proper assumptions on the CQP and some suitable condition on constraint selection rule, Algorithm CR-MPC converges globally to the solution, at a locally q-quadratic rate. The condition on the constraint selection rules is less restrictive than the assumptions made on the constraint selection rules for earlier constraint-reduced algorithms, (See, for example, [37, 38, 40].) while preserving the desired convergence properties. We then propose a simple constraint selection rule and prove that the new rule satisfies the condition. We compare the computational performance of Algorithm CR-MPC to the "affine-scaling" algorithm proposed in [40], with both the proposed rule and the adaptive constraint selection rule in [40]. Numerical re-

sults on randomly generated problems with a large number of inequality constraints show that, in most cases, Algorithm CR-MPC with the proposed rule significantly outperforms other tested combinations of algorithms and rules.

Finally, the contribution of our current work is summarized and the possible future work is discussed in Chapter 5. Appendix A gives implementation details of the finite volume kinetic scheme [10, 15] used in the tests for the FP_N^+ method in Chapter 2. Appendices B and C include the details of proofs for the global convergence property and local q-quadratic convergence rate for Algorithm CR-MPC.

Chapter 2: The FP_N^+ Method for Linear Kinetic Transport Equations

In this chapter, we propose the FP_N^+ method for solving linear kinetic transport equations. In Section 2.1, we review the kinetic equation, moment equations, and several moment closures including P_N, FP_N, PP_N, and UD_N closures. Section 2.2 introduces the proposed FP_N^+ closure and illustrates the implementation details in the FP_N^+ method. In Section 2.3, we present the consistency analysis of the FP_N^+ and UD_N closures and the associated numerical convergence results. The numerical results for the line source problem are provided in Section 2.4, and the efficiency and accuracy comparisons between the proposed FP_N^+ and other closures are also included.

2.1 Preliminaries and Notations

2.1.1 Kinetic Equations and Moment Models

As in [10], we consider a linear kinetic model of particles traveling with unit speed[1] which scatter isotropically off of a background material medium. Emission, absorption, and external sources are neglected for simplifying the presentation; they can be included easily. The kinetic description is given by a non-negative

[1]The unit speed assumption reduces the problem from six dimensions to five.

distribution function $f = f(r, \Omega, t)$ where $r = (x, y, z) \in \mathbb{R}^3$ is the spatial position, $\Omega = (\Omega_x, \Omega_y, \Omega_z) \in \mathbb{S}^2$ is the direction of particle travel, and $t \geq 0$ is the time. In terms of the polar angle θ and the azimuthal angle ϕ, $(\Omega_x, \Omega_y, \Omega_z) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$. In what follows, it is often convenient to express functions on $\mathbb{S}^2$ in terms $\mu := \cos\theta$ and ϕ.

The governing linear kinetic equation is of the form

$$\partial_t f + \Omega \cdot \nabla_r f = \frac{\sigma}{4\pi}\langle f \rangle - \sigma f \,, \tag{2.1}$$

where σ is the scattering cross-section, and the angle brackets denote integration with respect to Ω over the angular space $\mathbb{S}^2$, i.e., $\langle f \rangle(r,t) = \int_{\mathbb{S}^2} f(r, \Omega, t)\, d\Omega$.To obtain a unique solution, one must equip (2.1) with appropriate initial and boundary conditions.

Moments $\mathbf{u}^f$ associated to f are defined as

$$\mathbf{u}^f = \mathbf{u}^f(r,t) := \langle \mathbf{m} f(r, \cdot, t) \rangle \,, \tag{2.2}$$

where $\mathbf{m}$ is a vector of basis functions over $\mathbb{S}^2$. Following standard practice, we use spherical harmonic basis functions.[2] For moments up to order N, the spherical harmonics basis $\mathbf{m} : \mathbb{S}^2 \to \mathbb{R}^n$, $n = (N+1)^2$, is given by $\mathbf{m} = [m_0; \mathbf{m}_1; \ldots; \mathbf{m}_N]$, where $\mathbf{m}_\ell$ is the collection of the $2\ell + 1$ harmonics of degree ℓ, which are defined explicitly in [10]. The components of $\mathbf{m}$ form an orthogonal basis for $\mathbb{P}_N(\mathbb{S}^2)$, the space of polynomials in Ω on $\mathbb{S}^2$ with degree at most N. We assume the components of $\mathbf{m}$ are normalized so that $\langle \mathbf{m}\mathbf{m}^T \rangle = I_{n\times n}$.

[2]Spherical harmonics are eigenfunctions of general scattering operators. See, for example, [3, Section 1-4].

Equations for $\mathbf{u}^f$ are derived by multiplying the kinetic equation (2.1) by $\mathbf{m}$ and integrating over $\mathbb{S}^2$, which gives

$$\partial_t \mathbf{u}^f + \nabla_r \cdot \langle \mathbf{m}\Omega f \rangle = -\sigma R \mathbf{u}^f, \tag{2.3}$$

where the $n \times n$ matrix $R = \text{diag}(0, 1, \ldots, 1)$. Equation (2.3) is exact, but it is not closed due to the flux term $\langle \mathbf{m}\Omega f \rangle$. In particular, the spherical harmonic expansion of $\mathbf{m}_N \Omega$ involves harmonics of degree $N+1$ so that $\langle \mathbf{m}\Omega f \rangle$ cannot be expressed as a function of $\mathbf{u}^f$.

In order to close (2.3), we define an operator $\mathcal{E} : \mathbb{R}^n \to L^2(\mathbb{S}^2)$ that maps a given set of moments to a distribution on $\mathbb{S}^2$ that approximates f. Then (2.3) can be closed by substituting the *ansatz* $\mathcal{E}[\mathbf{u}]$ for f, which yields the closed moment system

$$\partial_t \mathbf{u} + \nabla_r \cdot \langle \mathbf{m}\Omega \mathcal{E}[\mathbf{u}] \rangle = -\sigma R \mathbf{u}\,. \tag{2.4}$$

The solution $\mathbf{u} = [u_0; \mathbf{u}_1; \ldots; \mathbf{u}_N]$ of system (2.4) is an approximation of the exact moments $\mathbf{u}^f$. Equation (2.4) can be solved numerically in a variety of ways. In this chapter, we use the kinetic scheme proposed in [10, 15]; see Appendix A.

In slab geometry, the distribution f in (2.1) is independent of x and y, i.e., $\partial_x f = \partial_y f = 0$. Thus one can express the angular dependence of f in terms of $\mu = \Omega_z$ only, thereby reducing the angular domain from $\mathbb{S}^2$ to $[-1, 1]$.[3] Thus, we also consider convergence of the FP_N^+ closure on the interval $[-1, 1]$. In this case, the angle brackets denote integration with respect to $\mu \in [-1, 1]$, and the moment

[3] In spherically symmetric geometries, the effective angular space also reduces to $[-1, 1]$, (See, for example, details in [4, Chapter 5].)

basis $\mathbf{m} : [-1, 1] \to \mathbb{R}^n$, $n = N + 1$, is given by $\mathbf{m} = [m_0; m_1; \ldots; m_N]$, where m_ℓ is the ℓ-th order Legendre polynomial on μ. The components of $\mathbf{m}$ in this case form an orthogonal basis for $\mathbb{P}_N([-1, 1])$, the vector space of polynomials on $[-1, 1]$ of degree at most N. We assume the standard normalization $\langle m_\ell^2 \rangle = \frac{2}{2\ell+1}$. Note that (2.3) and (2.4) still hold true for slab geometry, with the modified angular space and moment basis.

In the remaining parts of Section 2.1 and Section 2.2, we present several moment closures in full geometry. These closures can be formulated analogously in the case of slab geometry with minor modifications, as described in the preceding paragraph.

2.1.2 P_N Closures

The P_N equations approximate the linear kinetic equation (2.1) via a standard spectral method. For $\mathbf{u} \in \mathbb{R}^n$, the P_N operator $\mathcal{E}_{\mathrm{P}_N} : \mathbb{R}^n \to \mathbb{P}_N(\mathbb{S}^2)$ maps moments $\mathbf{u}$ to $\mathbb{P}_N(\mathbb{S}^2)$, with

$$\mathcal{E}_{\mathrm{P}_N}[\mathbf{u}] := \hat{\boldsymbol{\alpha}}_{\mathrm{P}_N}(\mathbf{u})^T \mathbf{m} \,, \tag{2.5}$$

where the P_N ansatz $\mathcal{E}_{\mathrm{P}_N}[\mathbf{u}]$ solves the L^2 entropy minimization problem

$$\underset{g \in L^2}{\text{minimize}} \quad \frac{1}{2} \langle g^2 \rangle \quad \text{subject to} \quad \langle \mathbf{m} g \rangle = \mathbf{u} \,, \tag{2.6}$$

and the expansion coefficients $\hat{\boldsymbol{\alpha}}_{\mathrm{P}_N}(\mathbf{u})$ solve the dual problem of (2.6), and are given by

$$\hat{\boldsymbol{\alpha}}_{\mathrm{P}_N}(\mathbf{u}) := \underset{\boldsymbol{\alpha} \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\langle |\boldsymbol{\alpha}^T \mathbf{m}|^2 \right\rangle - \mathbf{u}^T \boldsymbol{\alpha} \right\} = \langle \mathbf{m}\mathbf{m}^T \rangle^{-1} \mathbf{u} = \mathbf{u} \,. \tag{2.7}$$

Setting $\mathcal{E}[\mathbf{u}] = \mathcal{E}_{\mathrm{P}_N}[\mathbf{u}]$ in (2.4) gives the P_N equations:

$$\partial_t \mathbf{u} + \nabla_r \cdot \left\langle \Omega \mathbf{m}\mathbf{m}^T \right\rangle \mathbf{u} = -\sigma R \mathbf{u} \,. \tag{2.8}$$

2.1.3 Filtered P_N Closures (FP_N)

Filtering is commonly used to mitigate Gibbs phenomena in spectral methods for the spatial discretization of hyperbolic problems [45, 46]. Filtered spherical harmonics expansions for angular moment closures were first proposed in [11] in order to suppress oscillations and mitigate the occurrence of negative concentrations in the P_N solution.

The filter can be embedded directly into the numerical PDE solver for the P_N equations (2.8): before each time step, the moment $\mathbf{u}$ is replaced by $F\mathbf{u}$ where $F = \mathrm{blockdiag}(F_\ell I_{(2\ell+1)\times(2\ell+1)})$ is an $n \times n$ matrix and each $F_\ell \in [0,1]$ is a filtering coefficient, with $F_0 = 1$. Associated to $F\mathbf{u}$ is the ansatz

$$\mathcal{E}_{\mathrm{FP}_N}[\mathbf{u}] := \mathcal{E}_{\mathrm{P}_N}[F\mathbf{u}] = \hat{\boldsymbol{\alpha}}_{\mathrm{FP}_N}(\mathbf{u})^T \mathbf{m} \,, \tag{2.9}$$

where $\hat{\boldsymbol{\alpha}}_{\mathrm{FP}_N}(\mathbf{u}) := \hat{\boldsymbol{\alpha}}_{\mathrm{P}_N}(F\mathbf{u})$ solves the filtered version of dual problem (2.7)

$$\hat{\boldsymbol{\alpha}}_{\mathrm{FP}_N}(\mathbf{u}) = \underset{\boldsymbol{\alpha}\in\mathbb{R}^n}{\mathrm{argmin}} \left\{ \frac{1}{2} \left\langle |\boldsymbol{\alpha}^T \mathbf{m}|^2 \right\rangle - (F\mathbf{u})^T \boldsymbol{\alpha} \right\} = F \hat{\boldsymbol{\alpha}}_{\mathrm{P}_N}(\mathbf{u}) \,. \tag{2.10}$$

We call this the *discrete embedding* of the filter.

The original choice of F_ℓ in [11] was based on an optimization problem that penalizes angular derivatives of the ansatz. In [18], a more general formulation leads to a modified system of equations. There F_ℓ is given by

$$F_\ell = \left[\kappa \left(\frac{\ell}{N+1} \right) \right]^\nu \,, \tag{2.11}$$

where $\kappa : \mathbb{R}^+ \to [0,1]$ is a filter function,

$$\nu = -\frac{\sigma_{\mathrm{F}} \Delta t}{\log[\kappa(N/(N+1))]} \tag{2.12}$$

depends on the time step, and σ_{F} is a tuning parameter. We say κ has order $p > 0$ if $\kappa \in C^p(\mathbb{R}^+)$ and

$$\kappa(0) = 1 \quad \text{and} \quad \kappa^{(k)}(0) = 0 \ \text{ for } \ k = 1, \ldots, p-1 \,. \tag{2.13}$$

The choice of ν in (2.12) ensures the discrete embedding is formally consistent in the limit $\Delta t \to 0$ with a modified version of (2.8), the FP_N equations:

$$\partial_t \mathbf{u}^* + \nabla_r \cdot \left\langle \Omega \mathbf{m}\mathbf{m}^T \right\rangle \mathbf{u}^* = -\sigma R \mathbf{u}^* - \sigma_{\mathrm{F}} L \mathbf{u}^* \,, \tag{2.14}$$

where $L = \mathrm{blockdiag}(L_\ell I_{(2\ell+1)\times(2\ell+1)})$, and $L_\ell = \frac{\kappa\left(\frac{\ell}{N+1}\right)}{\kappa\left(\frac{N}{N+1}\right)}$. We refer to (2.14) as a *continuous embedding* of the filter.

In the following sections, we consider both types of embeddings: discrete and continuous. The discrete approach is more conducive to the consistency analysis in Section 2.3, while the continuous approach is better for assessing the space-time convergence of the PDE solver in Section 2.2.2.1. In Section 2.3.2, the convergence results of the FP_N closures are presented for the 2nd-order Lanczos filter [18], 4th-order spherical spline filter [18], and the 6th-order exponential filter [47]. The filter functions κ are given by

$$\kappa_{\mathrm{Lanczos}}(\eta) = \frac{\sin(\eta)}{\eta}, \quad \kappa_{\mathrm{SSpline}}(\eta) = \frac{1}{1+\eta^4}, \quad \kappa_{\mathrm{Exp}}(\eta) = \exp(c\eta^6), \tag{2.15}$$

where, in the definition of κ_{Exp}, $c = \log(\epsilon_M)$, ϵ_M being the machine precision. In the numerical tests presented in Section 2.4.2, the 4th-order spherical spline filter is used.

While the FP_N closure effectively damps oscillations in the numerical solution, it still suffers from some challenges. These include (i) the occurrence of negative particle concentrations that can affect the stability of nonlinear systems (see [19,20]) and (ii) the lack of a systematic way to choose the tuning parameter σ_{F}. In the remainder of this chapter, we address the former.

2.1.4 Positive P_N Closures (PP_N)

In [9], a positive particle concentration is ensured by imposing point-wise positivity constraints on a discretized version of (2.6). Let $\mathcal{Q}$ and $\mathcal{W}$ be the points and (strictly positive) weights of a quadrature rule on $\mathbb{S}^2$ with degree of precision $2N+1$—that is, the quadrature rule integrates polynomials in $\mathbb{P}_{2N+1}(\mathbb{S}^2)$ exactly (in exact arithmetic). Then the discrete PP_N ansatz $\mathcal{E}_{\mathrm{PP}_N} : \mathbb{R}^n \to \mathbb{R}^{|\mathcal{Q}|}$ maps $\mathbf{u}$ to the unique minimizer for

$$\begin{aligned} \underset{g\in\mathbb{R}^{|\mathcal{Q}|}}{\text{minimize}} \quad & \frac{1}{2}\sum_{k=1}^{|\mathcal{Q}|} w_k |g_k|^2 \\ \text{subject to} \quad & \sum_{k=1}^{|\mathcal{Q}|} w_k \mathbf{m}(\Omega_k) g_k = \mathbf{u}, \\ & g_k \geq 0, \quad \forall k \in \{1,\dots,|\mathcal{Q}|\}\,. \end{aligned} \tag{2.16}$$

where $(\Omega_k, w_k) \in (\mathcal{Q}, \mathcal{W})$ for all $k \in \{1,\dots,|\mathcal{Q}|\}$. If $\mathcal{E}_{\mathrm{P}_N}[\mathbf{u}] \geq 0$ on $\mathcal{Q}$, then $\mathcal{E}_{\mathrm{PP}_N}[\mathbf{u}]$ is just the restriction of $\mathcal{E}_{\mathrm{P}_N}[\mathbf{u}]$ to $\mathcal{Q}$.

In [10], a continuum variant of the PP_N closure was proposed to enforce positivity by adding a log penalty term to (2.6). In this case, the PP_N operator

$\mathcal{E}_{\mathrm{PP}_N} : \mathbb{R}^n \to L^2(\mathbb{S}^2)$ maps $\mathbf{u}$ to the unique minimizer for

$$\underset{g \in L^2(\mathbb{S}^2)}{\text{minimize}} \left\langle \frac{1}{2} g^2 - \delta \log g \right\rangle \qquad \text{subject to } \langle \mathbf{m} g \rangle = \mathbf{u} \,, \tag{2.17}$$

where $\delta > 0$ is a penalty parameter. Although (2.17) is formulated as a continuous problem, a quadrature rule is still required to approximate the integrals in the objective.

While both variants (2.16) and (2.17) of the PP_N closures generate a positive ansatz, numerical solutions of the modified optimization problems (2.16) and (2.17) are significantly more expensive to obtain. Moreover, neither ansatz is a polynomial. A consequence of this is that solutions of the PP_N equations suffer from artifacts, known as *ray effects* [3, Section 4-6], due to the fact that the quadrature rule is not exact.

2.1.5 Uniform Damping Closures (UD_N)

Uniform damping (UD) is a simple method for generating a non-negative polynomial reconstruction from given moments. It was first proposed in [21] as a limiter for finite volume discretizations of hyperbolic PDE, and has recently been used to generate discontinuous Galerkin and finite volume WENO methods [22, 23] that satisfy maximum principles while maintaining high-order.

The UD_N closure is a simple application of the UD method. It works by damping moments $\mathbf{u}_\ell$ uniformly for all $\ell > 0$, while preserving u_0. Given quadrature points and weights $(\mathcal{Q}, \mathcal{W})$, the UD_N operator $\mathcal{E}_{\mathrm{UD}_N} : \mathbb{R}^n \to \mathbb{P}_N(\mathbb{S}^2)$ maps $\mathbf{u}$ to the

ansatz

$$\mathcal{E}_{\mathrm{UD}_N}[\mathbf{u}] := \frac{u_0}{u_0 + \langle m_0 c_N \rangle}(\mathcal{E}_{\mathrm{FP}_N}[\mathbf{u}] + c_N), \quad c_N = -\min\left\{\min_{\Omega_k \in \mathcal{Q}} \mathcal{E}_{\mathrm{FP}_N}[\mathbf{u}](\Omega_k), 0\right\}. \tag{2.18}$$

This ansatz is still a spherical harmonics expansion; hence UD_N solutions do not suffer from ray effects as PP_N solutions do. In addition, it is inexpensive to implement. However, as proved in Theorem 2 in Section 2.3.1 and shown in Section 2.4.2, the UD_N closure may lose accuracy for problems with non-smooth solutions.

2.2 Positive Filtered P_N Closures (FP_N^+)

To overcome the drawbacks of the FP_N, PP_N, and UD_N closures, we design *positive filtered* P_N (or FP_N^+) closures. This closure prevents the occurrence of negative particle concentrations using a polynomial ansatz that is non-negative at a pre-selected set of quadrature points. The FP_N^+ ansatz is defined via the solution of an optimization problem. The FP_N^+ ansatz is more expensive to compute than the UD_N ansatz; however, it is more accurate. The benefits of this additional accuracy are analyzed and explored in Sections 2.3 and 2.4.

2.2.1 Formulation

The FP_N^+ operator $\mathcal{E}_{\mathrm{FP}_N^+} : \mathbb{R}^n \to \mathbb{P}_N(\mathbb{S}^2)$ maps moments $\mathbf{u}$ to the ansatz

$$\mathcal{E}_{\mathrm{FP}_N^+}[\mathbf{u}] := \hat{\boldsymbol{\alpha}}_{\mathrm{FP}_N^+}(\mathbf{u})^T \mathbf{m}\,, \tag{2.19}$$

where $\hat{\boldsymbol{\alpha}}_{\mathrm{FP}_N^+}(\mathbf{u})$ solves

$$\begin{aligned} \underset{\boldsymbol{\alpha}\in\mathbb{R}^n}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{\alpha}^T\mathbf{m} - \mathcal{E}_{\mathrm{FP}_N}[\mathbf{u}]\|^2_{L^2(\mathbb{S}^2)} \\ \text{subject to} \quad & \boldsymbol{\alpha}^T\mathbf{m}(\Omega_k) \geq 0\,, \qquad \forall \Omega_k \in \mathcal{Q}\,, \\ & \left\langle m_0 \boldsymbol{\alpha}^T\mathbf{m}\right\rangle = u_0\,, \end{aligned} \tag{2.20}$$

and $\mathcal{Q}$ is a quadrature set. The FP_N^+ ansatz is the best L^2 approximation to the FP_N ansatz in $\mathbb{P}_N(\mathbb{S}^2)$ that is non-negative on $\mathcal{Q}$ and preserves particle concentration.[4] The set $\mathcal{Q}$ is chosen so that the associated quadrature rule has degree of precision $2N+1$. This implies that the flux term $\langle \Omega \mathbf{m}\mathcal{E}[\mathbf{u}]\rangle$ in (2.4) is evaluated exactly whenever $\mathcal{E}[\mathbf{u}] \in \mathbb{P}_N(\mathbb{S}^2)$. It also ensures that u_0 is non-negative in the next update of the PDE solver (see Theorem 5 in the appendix).

Like the standard filter, the positive-preserving filter (2.20) can be discretely embedded into the numerical PDE solver for the P_N equations (2.8)[5] before each time step; the moment $\mathbf{u}$ is replaced by $\langle \mathbf{m}\mathcal{E}_{\mathrm{FP}_N^+}[\mathbf{u}]\rangle$. If the inequality constraints in (2.20) are not active at the solution, then $\langle \mathbf{m}\mathcal{E}_{\mathrm{FP}_N^+}[\mathbf{u}]\rangle = F\mathbf{u}$. Indeed, in this case, (2.20) is equivalent to the dual problem in (2.10). When the inequality constraints are active, $\langle \mathbf{m}\mathcal{E}_{\mathrm{FP}_N^+}[\mathbf{u}]\rangle$ depends on $\mathbf{u}$ in a nonlinear way that cannot be expressed in closed form. Rather it must be determined from the numerical solution of (2.20). With the continuous embedding, the filter is built in to the equations, but positivity is still embedded in the numerics: at each time step of the numerical PDE solver for the FP_N equations (2.14), the moment $\mathbf{u}^*$ is replaced by $\langle \mathbf{m}\mathcal{E}_{\mathrm{P}_N^+}[\mathbf{u}^*]\rangle$ where $\mathcal{E}_{\mathrm{P}_N^+}$ is given by (2.19) when there is no filter—that is, when $F = I$.

[4] The scalar u_0 is a positive constant multiple of the particle concentration.

[5] See the discussion on discrete and continuous embeddings in Section 2.1.3.

2.2.2 Implementation

In this subsection, we briefly describe the implementation of the FP_N^+ closures, which includes an algorithm for solving the optimization problem (2.20) and a numerical PDE solver for (2.4). Further details on the optimization algorithm is given in Chapter 4. The numerical PDE solver used in this chapter is described in Appendix A, and a more advanced numerical PDE solver with the asymptotic preserving property is proposed in Chapter 3.

2.2.2.1 Numerical PDE Solver

In this chapter, we generate a numerical solution of the FP_N^+ equations using a second-order kinetic scheme that was developed in [15]. (See references therein for early developments of this type of method.) The scheme is based on the following discrete ordinate approximation of (2.1):

$$\partial_t f^{\mathcal{Q}} + \nabla_r \cdot \Omega f^{\mathcal{Q}} = \frac{\sigma}{4\pi} \langle f^{\mathcal{Q}} \rangle_{\mathcal{Q}} - \sigma f^{\mathcal{Q}} \,, \tag{2.21}$$

where $f^{\mathcal{Q}}(x, \Omega, t) \approx f(x, \Omega, t)$ for each ordinate Ω in a quadrature set $\mathcal{Q}$ and $\langle \cdot \rangle_{\mathcal{Q}}$ denotes the quadrature rule associated to $\mathcal{Q}$. With an appropriate choice of quadrature, the P_N equations (2.8) can be derived directly from (2.21). Indeed, by taking quadrature-based moments of (2.21) and using the ansatz $\mathcal{E}_{\mathrm{P}_N}[\mathbf{u}]$ to approximate $f^{\mathcal{Q}}$, we arrive at the following system for the unknowns $\mathbf{u}$:

$$\partial_t \langle \mathbf{m} \mathcal{E}_{\mathrm{P}_N}[\mathbf{u}] \rangle_{\mathcal{Q}} + \nabla_r \cdot \langle \Omega \mathbf{m} \mathcal{E}_{\mathrm{P}_N}[\mathbf{u}] \rangle_{\mathcal{Q}} = \frac{\sigma}{4\pi} \langle \mathbf{m} \rangle_{\mathcal{Q}} \langle \mathcal{E}_{\mathrm{P}_N}[\mathbf{u}] \rangle_{\mathcal{Q}} - \sigma \langle \mathbf{m} \mathcal{E}_{\mathrm{P}_N}[\mathbf{u}] \rangle_{\mathcal{Q}} \,. \tag{2.22}$$

If, as in Section 2.2.1, the quadrature set $\mathcal{Q}$ is chosen so that $\langle\cdot\rangle_{\mathcal{Q}}$ has degree of precision $2N+1$, then (2.22) is equivalent to (2.8). This is our motivation for the choice of quadrature. A similar procedure can also be used to update the FP_N equations in (2.14).

It is known [15] that with an appropriate CFL condition, a finite volume discretization of (2.21) preserves the positivity of $f^{\mathcal{Q}}$. The corresponding kinetic scheme for (2.22) is derived by taking quadrature moments of this discretization and thus preserves positivity of the particle concentration. Details of this scheme and a precise statement of the positivity result (Theorem 5) are given in Appendix A.

2.2.2.2 Solving the FP_N^+ Optimization Problem

If $\hat{\boldsymbol{\alpha}}_{\mathrm{FP}_N}(\mathbf{u})$ satisfies the non-negativity constraints in (2.20), then $\hat{\boldsymbol{\alpha}}_{\mathrm{FP}_N}(\mathbf{u})$ solves (2.20)—that is, $\hat{\boldsymbol{\alpha}}_{\mathrm{FP}_N^+}(\mathbf{u}) = \hat{\boldsymbol{\alpha}}_{\mathrm{FP}_N}(\mathbf{u})$. Otherwise, a numerical optimization algorithm is needed. We discuss such an algorithm here.

Due to the orthonormality of spherical harmonics, the equality constraint $\langle m_0 \boldsymbol{\alpha}^T \mathbf{m} \rangle = u_0$ in (2.20) is equivalent to $\alpha_0 = u_0$. Hence the variable α_0 can be removed from the minimization problem, and (2.20) can be rewritten as

$$\begin{aligned} \underset{\tilde{\boldsymbol{\alpha}} \in \mathbb{R}^{n-1}}{\text{minimize}} \quad & \frac{1}{2} \left\langle |\tilde{\boldsymbol{\alpha}}^T \tilde{\mathbf{m}}|^2 \right\rangle - (\tilde{F}\tilde{\mathbf{u}})^T \tilde{\boldsymbol{\alpha}} \\ \text{subject to} \quad & \tilde{\boldsymbol{\alpha}}^T \tilde{\mathbf{m}}(\Omega_k) \geq -m_0 u_0 \,, \qquad \forall \Omega_k \in \mathcal{Q} \,, \end{aligned} \tag{2.23}$$

where $\tilde{\boldsymbol{\alpha}} = [\alpha_1, \ldots, \alpha_{n-1}]^T$, and similarly for $\tilde{\mathbf{u}}$, $\tilde{\mathbf{m}}$, and $\tilde{F}$. This is a convex quadratic program (CQP), which can be solved using primal-dual interior-point methods. We developed a constraint-reduced (CR) variant of Mehrotra's predictor-

corrector (MPC) algorithm to solve CQPs that arise from the FP_N^+ closure, where the problems have a large number of inequality constraints. We refer the algorithm as Algorithm CR-MPC. Details of Algorithm CR-MPC are provided in Chapter 4.

2.2.2.3 Quadrature

We use two types of quadrature to define the FP_N^+ and UD_N closures and evaluate the numerical flux in the PDE solver. One of them is a product quadrature on the unit sphere [48, 49]. This quadrature rule integrates any integrable function g on $\mathbb{S}^2$ by

$$\int_{\mathbb{S}^2} g(\Omega)d\Omega = \int_{-1}^{1}\int_{0}^{2\pi} g(\mu,\phi)d\phi d\mu \simeq \frac{\pi}{M_{\mathcal{Q}}}\sum_{k=1}^{M_{\mathcal{Q}}}\sum_{j=1}^{2M_{\mathcal{Q}}} w_k g(\mu_k,\phi_j)\,. \tag{2.24}$$

Here $\{\mu_k\}_{k=1}^{M_{\mathcal{Q}}}$ and $\{w_k\}_{k=1}^{M_{\mathcal{Q}}}$ are the Gauss-Legendre abscissas and weights, and $\{\phi_j\}_{j=1}^{2M_{\mathcal{Q}}}$ are equally spaced points from 0 to 2π. For closures with moment order N, we require the quadrature to have degree of precision $2N+1$, so we need a grid of at least $N+1$ (or $(N+1)/2$, for even functions on μ) Gauss-Legendre points in the μ direction and $2(N+1)$ equally spaced points in the ϕ direction.

Another quadrature we use is the Lebedev quadrature [50–54], which requires fewer quadrature points than the product quadrature does to achieve the same degree of precision. This property significantly reduces the computation time of the FP_N^+ method, where the quadrature points not only are used in numerical integration, but also are involved in the formulation of the optimization problem (2.23). Some comparisons of these two types of quadrature are given in Table 2.2, and discussed in Remark 4.

2.3 Consistency Results

In this section, we analyze consistency properties of the FP_N^+ and UD_N approximations and report numerical convergence results, for both full and slab geometries. We consider functions $\Psi = \Psi(\mu, \phi)$ where $\mu = \Omega_z \in [-1, 1]$ and $\phi \in [0, 2\pi]$ is the azimuthal angle on the sphere, and functions $\psi = \psi(\mu)$ which correspond to the slab geometry case discussed in Section 2.1.1.

For $q \in \mathbb{R}$, the fractional Sobolev spaces $H^q([-1, 1])$ is the set of functions ψ such that the norm

$$\|\psi\|_{H^q([-1,1])} := \left(\sum_{\ell=0}^{\infty} \ell^q (1+\ell)^q \left(\frac{2\ell+1}{2} \right) |\alpha_\ell|^2 \right)^{1/2}, \quad \alpha_\ell = \int_{-1}^{1} \psi(\mu) m_\ell(\mu) d\mu \tag{2.25}$$

is finite [55]. In this definition, m_ℓ is the ℓ^{th} Legendre polynomial. The space $H^q(\mathbb{S}^2)$ is the set of functions ψ such that the norm

$$\|\psi\|_{H^q(\mathbb{S}^2)} := \left(\sum_{\ell=0}^{\infty} \sum_{|j|\leq\ell} \ell^q (1+\ell)^q |\alpha_\ell^j|^2 \right)^{1/2}, \quad \alpha_\ell^j = \int_{\mathbb{S}^2} \psi(\Omega) m_\ell^j(\Omega) d\Omega \tag{2.26}$$

is finite [46]. In this definition, m_ℓ^j is the degree ℓ, order j spherical harmonic. In the remainder of this section, we use $\mathcal{S}$ to denote either $[-1, 1]$ or $\mathbb{S}^2$. Recall that $H^0(\mathcal{S}) = L^2(\mathcal{S})$.

For $q > 0$, let $q = v + w$, v a positive integer and $w \in [0, 1)$. Then the space $C^q([-1, 1])$ is defined as the set of functions ψ such that the norm

$$\|\psi\|_{C^q([-1,1])} := \|\psi\|_{L^\infty([-1,1])} + \sup_{\substack{\mu_1,\mu_2\in[-1,1] \\ \mu_1\neq\mu_2}} \frac{|\psi^{(v)}(\mu_1) - \psi^{(v)}(\mu_2)|}{|\mu_1 - \mu_2|^w} \tag{2.27}$$

is finite [55]. Here $\psi^{(v)}$ is the v-th derivative of ψ on $[-1, 1]$. Similarly, the space $C^q(\mathbb{S}^2)$ is defined as the set of functions ψ such that the norm

$$\|\psi\|_{C^q(\mathbb{S}^2)} := \|\psi\|_{L^\infty(\mathbb{S}^2)} + \max_{1\le i<j\le 3} \sup_{0<|\vartheta|\le 1} \frac{\|(I - R_{i,j,\vartheta})D^v_{i,j}\psi\|_{L^\infty(\mathbb{S}^2)}}{|\vartheta|^w}, \tag{2.28}$$

is finite [56]. Here the operator $D_{1,2} := x\partial_x - y\partial_y$, $D_{2,3} := y\partial_y - z\partial_z$, $D_{1,3} := x\partial_x - z\partial_z$, x, y, z are the Cartesian coordinates on the sphere, I denotes the identity operator, and $R_{i,j,\vartheta}$ denotes the rotation operator. For example, $R_{1,2,\vartheta}g(\Omega) = g(\Omega')$, where Ω' is obtained by rotating Ω with angle ϑ in the x-y plane, i.e., for $\Omega = (\Omega_x, \Omega_y, \Omega_z)$, Ω' is given by $(\Omega_x \cos\vartheta - \Omega_y \sin\vartheta, \Omega_x \sin\vartheta + \Omega_y \cos\vartheta, \Omega_z)$. $R_{2,3,\vartheta}$ and $R_{1,3,\vartheta}$ are analogously defined for rotations in y-z and x-z planes, respectively. Note that, for $q \in \mathbb{N}$, the space $C^q(\mathcal{S})$ is the space of functions with a continuous q-th derivative on $\mathcal{S}$. Finally, recall that $C^q(\mathcal{S}) \subset H^q(\mathcal{S})$.

2.3.1 Error Estimates of Approximations

The P_N approximation (2.5) is based on the degree N spherical harmonic expansion of $\psi \in L^2(\mathbb{S}^2)$ with moments $\mathbf{u}^N := \mathbf{u}$.[6] For $\psi \in C^\infty(\mathbb{S}^2)$, this expansion converges to ψ (in the L^2 sense) faster than any negative power of N. For $\psi \in H^q(\mathbb{S}^2)$, it converges to ψ (in the L^2 sense) at rate q [57]. The filtered expansion (2.9) shares the convergence rate q with the P_N approximation if the filter order p satisfies $p \ge q$, but has a slower convergence rate p otherwise; see [47]. Based on these results, we establish the following convergence properties for the FP^+_N approximation.

[6] In this section, we use a superscript to emphasize the dependence of the moment vector on N.

Theorem 1. *For $M > 0$, let $\mathcal{D}_M = \{g \in L^\infty(\mathcal{S}) : \|g\|_{L^\infty(\mathcal{S})} \leq M\|g\|_{L^1(\mathcal{S})}\}$. Then, given a non-negative function $\psi \in C^q(\mathcal{S}) \cap \mathcal{D}_M$, $q \geq 0$, there exists a constant $A(q, M)$ such that*

$$\|\psi - \mathcal{E}_{FP_N^+}[\mathbf{u}^N]\|_{L^2(\mathcal{S})} \leq A(q, M)N^{-s}\|\psi\|_{C^q(\mathcal{S})}, \quad \forall N \in \mathbb{N}\,, \tag{2.29}$$

where $\mathbf{u}^N \in \mathbb{R}^n$ consists of the moments of ψ up to order N, and $s = \min\{q, p\}$, with p the order of filter F in (2.10).

Before proving Theorem 1, we give two lemmas which are used in the proof. The first lemma gives the convergence rate of the FP_N approximation, and the second lemma provides an L^∞ error estimate of the best polynomial approximation for continuous functions.

Lemma 1. *For every $q \in \mathbb{R}$, there exists a constant $A_1(q)$ such that, for all $\psi \in H^q(\mathcal{S})$,*

$$\|\psi - \mathcal{E}_{FP_N}[\mathbf{u}^N]\|_{L^2(\mathcal{S})} \leq A_1(q)N^{-s}\|\psi\|_{H^q(\mathcal{S})}, \quad \forall N \in \mathbb{N}\,, \tag{2.30}$$

where $\mathbf{u}^N \in \mathbb{R}^n$ consists of the moments of ψ up to order N, and $s = \min\{q, p\}$, with p the filter order in (2.10).

Proof. See [47]. □

Lemma 2. *For every $q \geq 0$, there exists a constant $A_2(q)$ such that, for all $\psi \in C^q(\mathcal{S})$,*

$$\min_{\varphi \in \mathbb{P}_N(\mathcal{S})} \|\psi - \varphi\|_{L^\infty(\mathcal{S})} \leq A_2(q)N^{-q}\|\psi\|_{C^q(\mathcal{S})}, \quad \forall N \in \mathbb{N}\,, \tag{2.31}$$

where the minimum is attained.

Proof. From [58, Theorem 2] (for $\mathcal{S} = [-1, 1]$) and [56, Theorem 4.8.1] (for $\mathcal{S} = \mathbb{S}^2$)

$$\inf_{\varphi \in \mathbb{P}_N(\mathcal{S})} \|\psi - \varphi\|_{L^\infty(\mathcal{S})} \leq A_2(q) N^{-q} \|\psi\|_{C^q(\mathcal{S})} \,. \tag{2.32}$$

Since $\mathbb{P}_N(\mathcal{S})$ is a finite dimensional subspace of the Banach space $C^q(\mathcal{S})$, it follows from Theorem 1.1 in [59] that the infimum in (2.32) is attained. □

We now prove Theorem 1 for the case $\mathcal{S} = \mathbb{S}^2$; when $\mathcal{S} = [-1, 1]$, the result can be proved analogously. To simplify notation, we write

$$\|\cdot\|_{C^q} = \|\cdot\|_{C^q(\mathbb{S}^2)}\,; \quad \|\cdot\|_{L^p} = \|\cdot\|_{L^p(\mathbb{S}^2)}\,; \quad \mathcal{E}_{\mathrm{FP}_N} = \mathcal{E}_{\mathrm{FP}_N}[\mathbf{u}^N]\,; \quad \mathcal{E}_{\mathrm{FP}_N^+} = \mathcal{E}_{\mathrm{FP}_N^+}[\mathbf{u}^N]. \tag{2.33}$$

Proof of Theorem 1. If $\psi = 0$, then $\mathbf{u}^N = 0$ and $\mathcal{E}_{\mathrm{FP}_N^+} = 0$, and the claim holds trivially. Hence consider the case for $\psi \neq 0$, i.e., $\langle\psi\rangle > 0$. Using Lemma 2, let $\hat{\varphi}_N$ be the minimizer on the left-hand side of (2.31), and let $\varphi_N = \hat{\varphi}_N + \frac{1}{4\pi}\langle\psi - \hat{\varphi}_N\rangle$. Then $\langle\varphi_N\rangle = \langle\psi\rangle > 0$, and

$$\|\psi - \varphi_N\|_{L^\infty} \leq \|\psi - \hat{\varphi}_N\|_{L^\infty} + \frac{1}{4\pi}\langle|\psi - \hat{\varphi}_N|\rangle \leq 2\|\psi - \hat{\varphi}_N\|_{L^\infty} \leq 2A_2(q)N^{-q}\|\psi\|_{C^q} \,. \tag{2.34}$$

We now modify φ_N to generate a non-negative polynomial that still approximates ψ well. Let $\bar{c}_N = -\min\{\min_{\Omega\in\mathbb{S}^2} \varphi_N(\Omega), 0\} \geq 0$. Then by definition, $\varphi_N + \bar{c}_N$ is non-negative, and $\langle\varphi_N + \bar{c}_N\rangle$ is positive. Hence the function

$$\varphi_N^+ := \frac{\langle\varphi_N\rangle}{\langle\varphi_N + \bar{c}_N\rangle}(\varphi_N + \bar{c}_N) = \frac{\langle\psi\rangle}{\langle\psi + \bar{c}_N\rangle}(\varphi_N + \bar{c}_N) \tag{2.35}$$

is a well-defined, non-negative polynomial on $\mathbb{S}^2$, and $\langle\varphi_N^+\rangle = \langle\varphi_N\rangle = \langle\psi\rangle$. Moreover,

$$\|\varphi_N - \varphi_N^+\|_{L^2} = \frac{\|\langle\bar{c}_N\rangle\varphi_N - \langle\psi\rangle\bar{c}_N\|_{L^2}}{\langle\psi + \bar{c}_N\rangle} = \frac{4\pi\bar{c}_N\sqrt{\langle\varphi_N^2\rangle - \frac{\langle\psi\rangle^2}{4\pi}}}{\langle\psi\rangle + 4\pi\bar{c}_N} \leq 4\pi\bar{c}_N\frac{\|\varphi_N\|_{L^2}}{\langle\psi\rangle} \,. \tag{2.36}$$

By Hölder's inequality, $\|\varphi_N\|_{L^2} \leq \sqrt{4\pi}\|\varphi_N\|_{L^\infty}$. Using the triangle inequality, (2.34), and the fact that $\hat{\varphi}_N$ is the minimizer, we have

$$\|\varphi_N\|_{L^\infty} \leq \|\psi\|_{L^\infty} + \|\psi - \varphi_N\|_{L^\infty} \leq \|\psi\|_{L^\infty} + 2\|\psi - \hat{\varphi}_N\|_{L^\infty} \leq 3\|\psi\|_{L^\infty} \,. \tag{2.37}$$

Applying Hölder's inequality and substituting the bound for $\|\varphi_N\|_{L^\infty}$ in (2.37) into (2.36) yield

$$\|\varphi_N - \varphi_N^+\|_{L^2} \leq \left(24\pi^{3/2}\frac{\|\psi\|_{L^\infty}}{\|\psi\|_{L^1}}\right)\bar{c}_N \leq 24\pi^{3/2}M\bar{c}_N \,, \tag{2.38}$$

where the second inequality comes from the assumption that $\psi \in \mathcal{D}_M$. This bound will be used below in (2.42).

By construction, the vector of expansion coefficients for φ_N^+ is a feasible point of (2.20). Because the corresponding vector of expansion coefficients for $\mathcal{E}_{\mathrm{FP}_N^+}$ solves (2.20), we have

$$\|\mathcal{E}_{\mathrm{FP}_N} - \mathcal{E}_{\mathrm{FP}_N^+}\|_{L^2} \leq \|\mathcal{E}_{\mathrm{FP}_N} - \varphi_N^+\|_{L^2}. \tag{2.39}$$

Hence,

$$\begin{aligned}
\|\psi - \mathcal{E}_{\mathrm{FP}_N^+}\|_{L^2} &\leq \|\psi - \mathcal{E}_{\mathrm{FP}_N}\|_{L^2} + \|\mathcal{E}_{\mathrm{FP}_N} - \mathcal{E}_{\mathrm{FP}_N^+}\|_{L^2} \\
&\leq \|\psi - \mathcal{E}_{\mathrm{FP}_N}\|_{L^2} + \|\mathcal{E}_{\mathrm{FP}_N} - \varphi_N^+\|_{L^2} \\
&\leq \|\psi - \mathcal{E}_{\mathrm{FP}_N}\|_{L^2} + \|\mathcal{E}_{\mathrm{FP}_N} - \psi\|_{L^2} + \|\psi - \varphi_N^+\|_{L^2} \\
&\leq 2\|\psi - \mathcal{E}_{\mathrm{FP}_N}\|_{L^2} + \|\psi - \varphi_N^+\|_{L^2}
\end{aligned} \tag{2.40}$$

We bound each of these terms separately. Lemma 1 and the fact that $\|\psi\|_{H^q} \leq A_3\|\psi\|_{C^q}$ for some constant A_3, gives a bound on the first term:

$$\|\psi - \mathcal{E}_{\mathrm{FP}_N}\|_{L^2} \leq A_1(q)N^{-s}\|\psi\|_{H^q} \leq A_1(q)A_3N^{-s}\|\psi\|_{C^q} \,. \tag{2.41}$$

For the second term, we apply the triangle inequality, Hölder's inequality, and (2.38). This gives

$$\|\psi-\varphi_N^+\|_{L^2} \le \|\psi-\varphi_N\|_{L^2}+\|\varphi_N-\varphi_N^+\|_{L^2} \le \sqrt{4\pi}\|\psi-\varphi_N\|_{L^\infty}+\left(24\pi^{3/2}M\right)\bar{c}_N. \quad (2.42)$$

Since $\psi \ge 0$, $\bar{c}_N \le \|\psi - \varphi_N\|_{L^\infty}$. We substitute this bound into (2.42), combine terms in $\|\psi - \varphi_N\|_{L^\infty}$, and apply the bound in (2.34). This gives

$$\|\psi - \varphi_N^+\|_{L^2} \le \left(\sqrt{4\pi} + 24\pi^{3/2}M\right) \|\psi - \varphi_N\|_{L^\infty} \le A_4(q, M)N^{-q}\|\psi\|_{C^q} \quad (2.43)$$

where $A_4(q, M) = 2A_2(q)\left(\sqrt{4\pi} + 24\pi^{3/2}M\right)$. Finally, by substituting the bounds in (2.41) and (2.43) into (2.40), the claim (2.29) is proved, with $A(q, M) = 2A_1(q)A_3 + A_4(q, M)$. □

For comparison, the next theorem provides error estimates for the uniform damping (UD_N) approximation.

Theorem 2. *For $M > 0$, let $\mathcal{D}_M = \{g \in L^2(\mathcal{S}) : \|g\|_{L^2(\mathcal{S})} \le M\|g\|_{L^1(\mathcal{S})}\}$. Then, given a non-negative $\psi \in H^q(\mathcal{S}) \cap \mathcal{D}_M$, $q \ge 0$, $\epsilon > 0$, there exists a constant $B(q, M, \epsilon)$ such that,*

$$\|\psi - \mathcal{E}_{UD_N}[\mathbf{u}^N]\|_{L^2(\mathcal{S})} \le B(q, M, \epsilon)N^{-(s-a-\epsilon)}\|\psi\|_{H^q(\mathcal{S})}, \quad \forall N \in \mathbb{N}\,, \quad (2.44)$$

where $\mathbf{u}^N \in \mathbb{R}^n$ consists of the moments of ψ up to order N, and $s = \min\{q, p\}$, with p the order of filter F in (2.10). The constant a depends on $\mathcal{S}$: when $\mathcal{S} = [-1, 1]$, $a = 3/4$; when $\mathcal{S} = \mathbb{S}^2$, $a = 1$.

The following lemma is used in the proof of Theorem 2.

Lemma 3. *For every $q \geq 0$ and $\delta > 0$, there exist constants $B_1(q,\delta)$ and $B_2(q,\delta)$ such that, for all $\psi \in H^q([-1,1])$ and $N \in \mathbb{N}$,*

$$\|\psi - \mathcal{E}_{FP_N}[\mathbf{u}^N]\|_{L^\infty([-1,1])} \leq \|\psi - \mathcal{E}_{FP_N}[\mathbf{u}^N]\|_{H^{\frac{1}{2}+\delta}([-1,1])} \leq B_1(q,\delta) N^{-(s-\frac{3}{4}-\frac{3\delta}{2})} \|\psi\|_{H^q([-1,1])}, \tag{2.45}$$

and for all $\psi \in H^q(\mathbb{S}^2)$ and $N \in \mathbb{N}$,

$$\|\psi - \mathcal{E}_{FP_N}[\mathbf{u}^N]\|_{L^\infty(\mathbb{S}^2)} \leq \|\psi - \mathcal{E}_{FP_N}[\mathbf{u}^N]\|_{H^{1+\delta}(\mathbb{S}^2)} \leq B_2(q,\delta) N^{-(s-1-\delta)} \|\psi\|_{H^q(\mathbb{S}^2)}\,, \tag{2.46}$$

where $\mathbf{u}^N \in \mathbb{R}^n$ consists of the moments of ψ up to order N, and $s = \min\{q,p\}$, with p the filter order in (2.10).

The first inequalities in (2.45) and (2.46) are Sobolev embedding theorems that can be found in [55] and [60], respectively. The second inequalities can be found in [61, Theorem 2.2] and [46, Theorem 8.2], respectively.

Proof of Theorem 2. For convenience, we denote $\mathcal{E}_{\mathrm{FP}_N}[\mathbf{u}^N]$ and $\mathcal{E}_{\mathrm{UD}_N}[\mathbf{u}^N]$ as $\mathcal{E}_{\mathrm{FP}_N}$ and $\mathcal{E}_{\mathrm{UD}_N}$, respectively. By the triangle inequality,

$$\|\psi - \mathcal{E}_{\mathrm{UD}_N}\|_{L^2(\mathcal{S})} \leq \|\psi - \mathcal{E}_{\mathrm{FP}_N}\|_{L^2(\mathcal{S})} + \|\mathcal{E}_{\mathrm{FP}_N} - \mathcal{E}_{\mathrm{UD}_N}\|_{L^2(\mathcal{S})}\,. \tag{2.47}$$

The bound for the first term in (2.47) is given by (2.30) in Lemma 1. For the second term, we use the definition of $\mathcal{E}_{\mathrm{UD}_N}$ in (2.18) to compute (recalling that m_0 and c_N are constant over $\mathcal{S}$)

$$\|\mathcal{E}_{\mathrm{FP}_N} - \mathcal{E}_{\mathrm{UD}_N}\|_{L^2(\mathcal{S})} = \frac{\|\langle m_0 c_N\rangle \mathcal{E}_{\mathrm{FP}_N} - \langle m_0\psi\rangle c_N\|_{L^2(\mathcal{S})}}{\langle m_0\psi\rangle + \langle m_0 c_N\rangle} = \frac{B_3 c_N \sqrt{\langle \mathcal{E}^2_{\mathrm{FP}_N}\rangle - \frac{\langle\psi\rangle}{B_3}}}{\langle\psi\rangle + \langle c_N\rangle}\,, \tag{2.48}$$

where $B_3 = \langle 1 \rangle$. Because $\|\mathcal{E}_{\mathrm{FP}_N}\|_{L^2(\mathcal{S})} \leq \|\mathcal{E}_{\mathrm{P}_N}\|_{L^2(\mathcal{S})} \leq \|\psi\|_{L^2(\mathcal{S})}$ and $c_N \leq \|\psi - \mathcal{E}_{\mathrm{FP}_N}\|_{L^\infty(\mathcal{S})}$, it follows from (2.48) and $\psi \in \mathcal{D}_M$ that

$$\|\mathcal{E}_{\mathrm{FP}_N} - \mathcal{E}_{\mathrm{UD}_N}\|_{L^2(\mathcal{S})} \leq \frac{B_3 c_N \|\mathcal{E}_{\mathrm{FP}_N}\|_{L^2(\mathcal{S})}}{\langle \psi \rangle + \langle c_N \rangle} \leq B_3 \frac{\|\psi\|_{L^2(\mathcal{S})}}{\|\psi\|_{L^1(\mathcal{S})}} c_N \leq B_3 M \|\psi - \mathcal{E}_{\mathrm{FP}_N}\|_{L^\infty(\mathcal{S})} \,. \tag{2.49}$$

The bound for the second term in (2.47) is then obtained by applying either (2.45) or (2.46) in Lemma 3 on the right-hand side of (2.49). Finally, by bounding for both terms in (2.47), the claim (2.44) is proved, with

$$B(q, M, \epsilon) = \begin{cases} A_1(q) + B_1(q, 2\epsilon/3) B_3 M \,, & \text{when } \mathcal{S} = [-1, 1] \\ A_1(q) + B_2(q, \epsilon) B_3 M \,, & \text{when } \mathcal{S} = \mathbb{S}^2 \end{cases} \tag{2.50}$$

chosen to be the constant.

Remark 1. *The error estimate in* (2.44) *appears to be sharp for both choices of* $\mathcal{S}$*. This is illustrated in Figures 2.4 and 2.7 with target functions given in* (2.54) *and* (2.57) *in the next subsection.*

Remark 2. *The fact that* ψ *may be zero on* $\mathcal{S}$ *is what limits the error estimates for both the* FP_N^+ *approximation (Theorem 1) and the* UD_N *approximation (Theorem 2). However, if* ψ *is strictly positive and* $\mathcal{E}_{FP_N}[\mathbf{u}^N]$ *converges to* ψ *uniformly, then one can prove that both* $\mathcal{E}_{FP_N^+}$ *and* $\mathcal{E}_{UD_N}$ *recover the optimal rate for the* FP_N *approximation. Indeed, uniform convergence to a strictly positive function implies that* $\mathcal{E}_{FP_N}[\mathbf{u}^N] > 0$ *for all* N *greater than some* $\tilde{N}$*. In this case,* $\mathcal{E}_{FP_N^+}[\mathbf{u}^N] = \mathcal{E}_{UD_N}[\mathbf{u}^N] = \mathcal{E}_{FP_N}[\mathbf{u}^N]$.

2.3.2 Convergence Tests

In this subsection, we present numerical convergence results for the FP_N^+ and UD_N approximations. These results suggest that the stronger assumptions for the FP_N^+ approximation about the underlying function (C^q vs. H^q) in Theorem 1 may not be necessary. Meanwhile, the convergence rates for the UD_N approximation in Theorem 2 appear to be sharp.

We begin with one-dimensional tests for functions defined on $[-1, 1]$. For an expansion of degree N, we use for $\mathcal{Q}$ (cf. (2.20)) a Gauss-Legendre quadrature rule with $N + 1$ points, which has degree of precision $2N + 1$. Figures 2.1–2.5 illustrate convergence results for several functions, each with different regularity properties. Corresponding results for the P_N and FP_N approximation are included for reference. In each figure, we plot the L^2 approximation errors versus the moment order N. The observed convergence rates are shown in the legend.[7]

The target functions are as follows:

- Step function (Figure 2.1). The function

$$\psi(\mu) = \begin{cases} 1, & \mu \in (\hat{\mu}, 1] \\ 0, & \mu \in [-1, \hat{\mu}] \end{cases} \tag{2.51}$$

  where $\hat{\mu} = 0.75$, is in $H^q([-1, 1])$, $\forall q < 0.5$. From Figure 2.1, it can be seen that the P_N^+ (FP_N^+ with no spectral filter) and FP_N^+ approximations converge roughly at the same rate as the P_N and FP_N approximation. The UD_N ap-

[7] For oscillatory data, we compute the convergence rate based on its upper envelope. Otherwise, the convergence rate is given by the slope of least square linear fitting of the data.

proximations, on the other hand, have a slower convergence rate, which is consistent with result of Theorem 2. Note that $\hat{\mu}$ can be arbitrarily chosen from $(-1, 1)$. However, for some choices of $\hat{\mu}$, the approximation errors may converge faster than the (worst case) error estimates given in Theorems 1 and 2.

- Singular function (Figure 2.2). The function

$$\psi(\mu) = \begin{cases} (\mu - 0.75)^{-0.1}, & \mu \in (\hat{\mu}, 1] \\ 0, & \mu \in [-1, \hat{\mu}] \end{cases} \tag{2.52}$$

is an L^2 function with a singularity at $\hat{\mu} = 0.75$. For this function, the UD_N approximation does not converge, while the FP_N^+ approximation still converges roughly at the same rate as the FP_N approximation.

- Smooth function (Figure 2.3). The function

$$\psi(\mu) = \exp(5\mu \sin(10\mu)) \tag{2.53}$$

is in $C^\infty([-1, 1])$. Here we observe, as is expected from Theorems 1 and 2, that the FP_N^+ and UD_N approximations to converge with the order of the spectral filter used to define them. If no filter is applied, both approximations converge spectrally.

- Sobolev function (Figure 2.4 and 2.5). The function

$$\psi(\mu) = \begin{cases} (\mu - \hat{\mu})^r, & \mu \in (\hat{\mu}, 1] \\ 0, & \mu \in [-1, \hat{\mu}] \end{cases} \tag{2.54}$$

belongs to $H^q([-1,\,1])$ for all $q < r + \frac{1}{2}$ and all $\hat{\mu} \in (-1, 1)$. For such functions, the UD_N approximations typically converge at slower rates than the P_N and P_N^+ approximations. We consider two specific cases: in Figure 2.4, $(r, \hat{\mu}) = (0.5, 0.975)$; in Figure 2.5, $(r, \hat{\mu}) = (3, 0.75)$. In the first case, we select $\hat{\mu}$ in order to show that the estimate in Theorem 2 is most likely sharp. Indeed, in Figure 2.4, the convergence rate of the UD_N ansatz is around 0.25, which matches the error estimate provided in Theorem 2. In the second case, r is chosen to illustrate the effect of the spectral filters on the convergence rate. In Figure 2.5, we observe that a loss in order occurs for the UD_N approximation when $r + 1/2 \leq p$—that is, when the order of the filter is greater than the regularity of ψ.

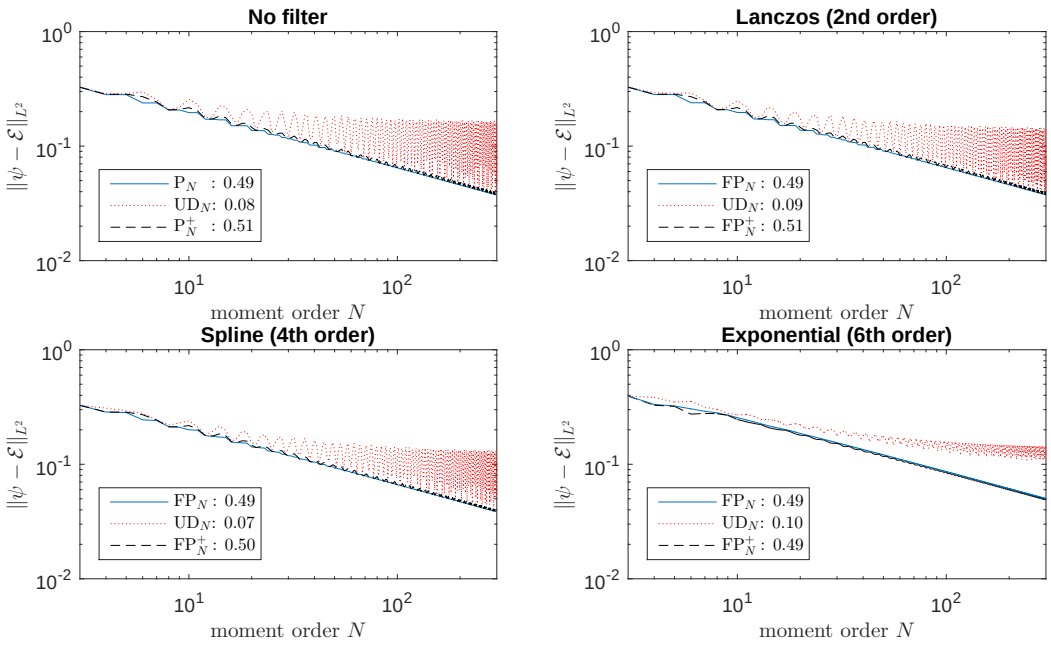


Figure 2.1: Step function ψ on $[-1,\,1]$. $\psi \in H^q([-1,\,1])$ for all $q < 0.5$; see (2.51). The observed convergence rates, as defined in Footnote 7, are listed in the legend.

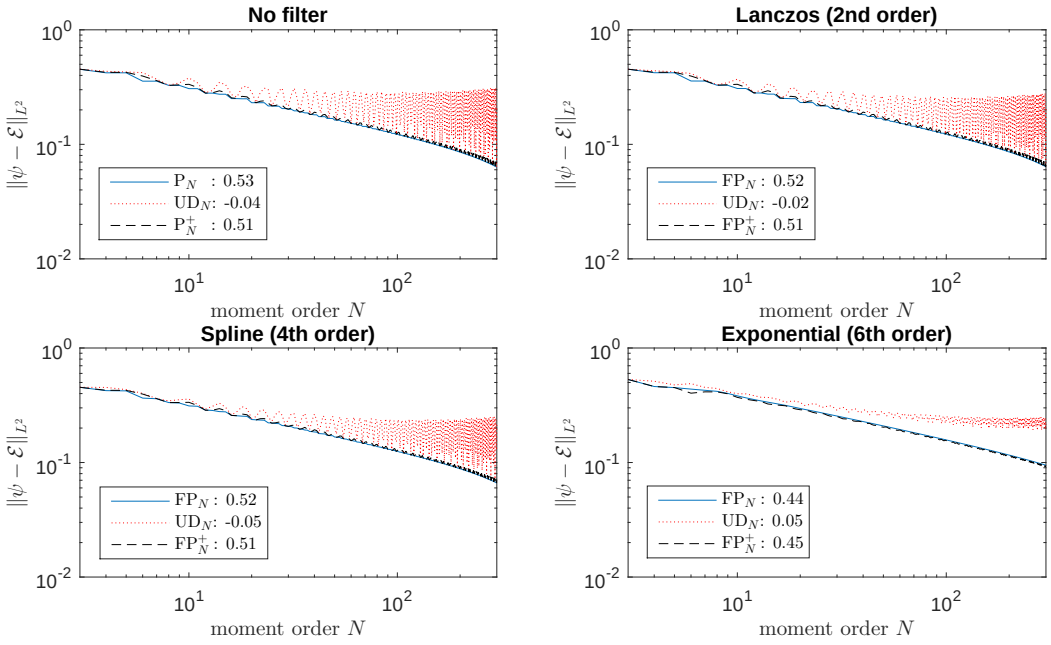


Figure 2.2: Singular function ψ on $[-1, 1]$. $\psi \in L^2([-1, 1])$; see (2.52). The observed convergence rates, as defined in Footnote 7, are listed in the legend.

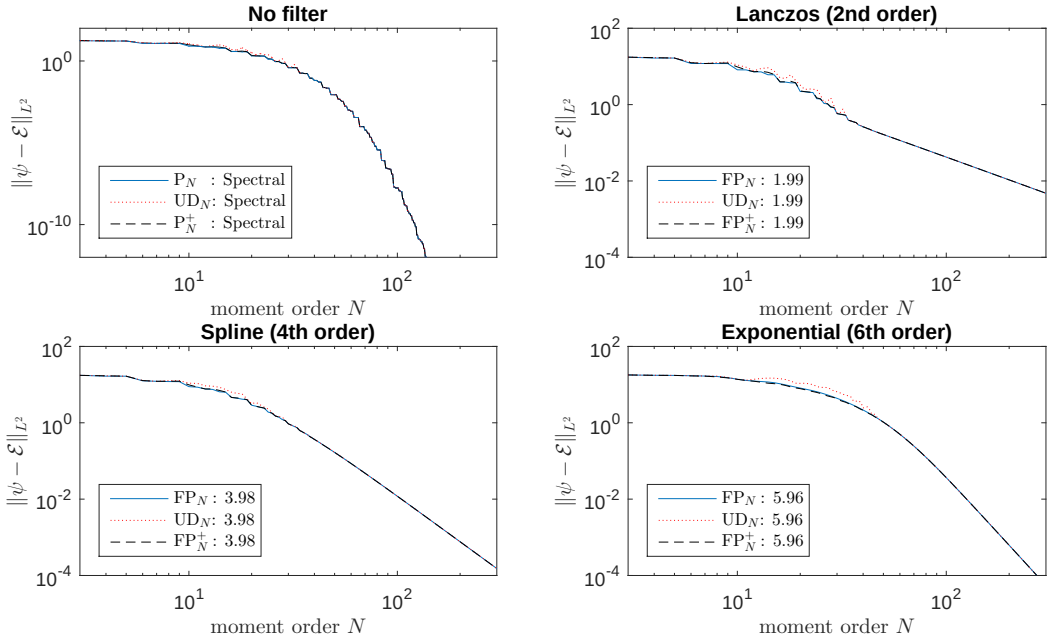


Figure 2.3: Smooth function ψ on $[-1, 1]$. $\psi \in C^\infty([-1, 1])$; see (2.53). The observed convergence rates, as defined in Footnote 7, are listed in the legend.

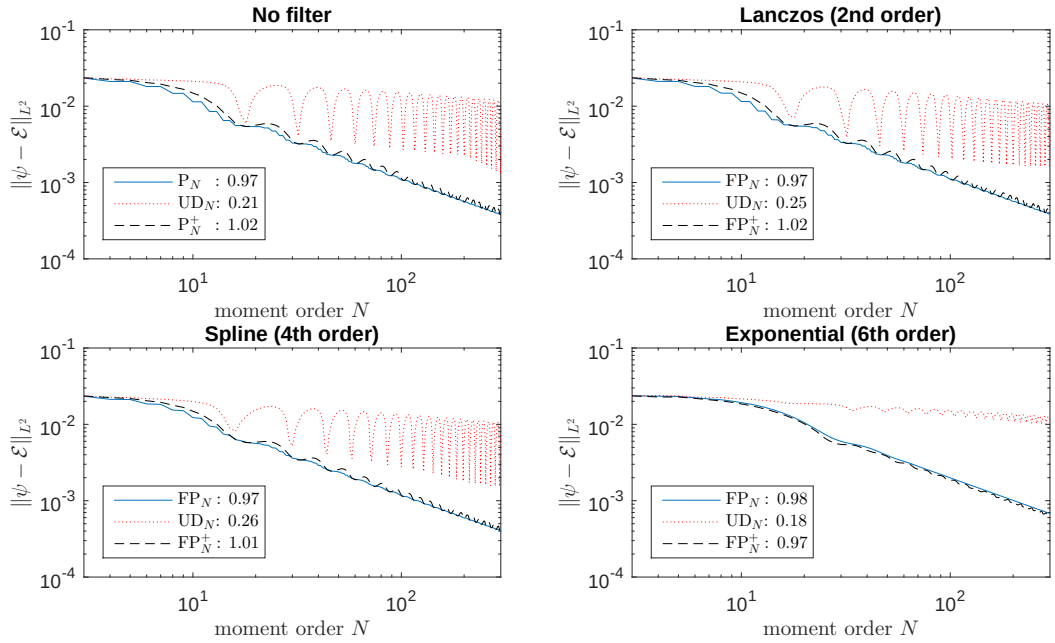


Figure 2.4: Sobolev function ψ on $[-1, 1]$. $\psi \in H^q([-1, 1])$ for all $q < 1$; see (2.54), $r = 0.5$, $\hat{\mu} = 0.975$. The observed convergence rates, as defined in Footnote 7, are listed in the legend.

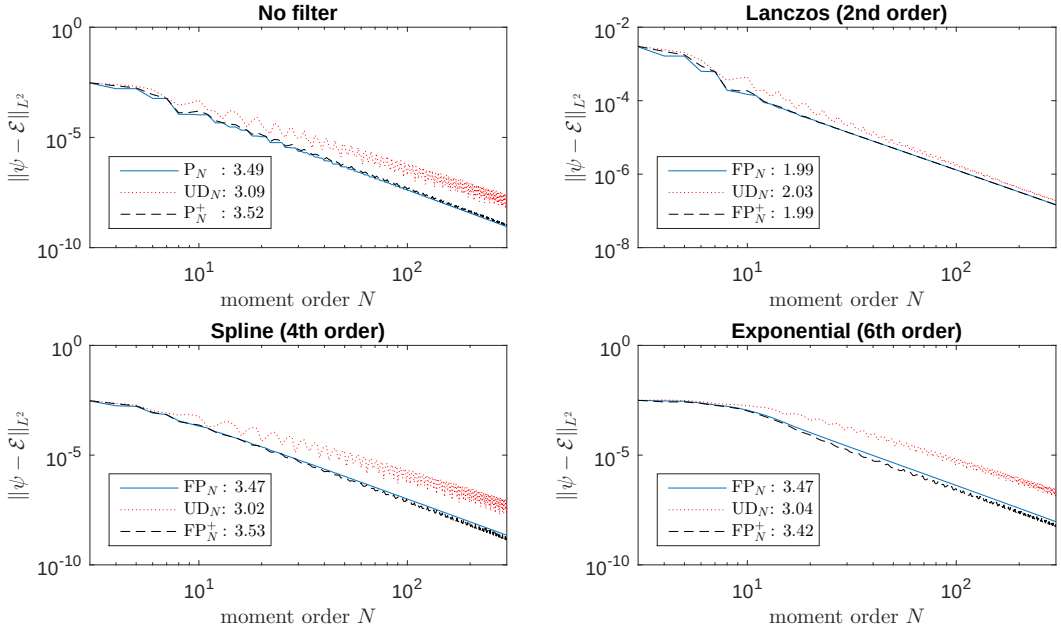


Figure 2.5: Sobolev function ψ on $[-1, 1]$. $\psi \in H^q([-1, 1])$ for all $q < 3.5$; see (2.54), $r = 3$, $\hat{\mu} = 0.75$. The observed convergence rates, as defined in Footnote 7, are listed in the legend.

We next consider target functions Ψ on $\mathbb{S}^2$ that are simple extensions of functions ψ on $[-1, 1]$:

$$\Psi(\mu, \phi) := \psi(\mu), \quad \forall(\mu, \phi) \in [-1, 1] \times [0, 2\pi] \,. \tag{2.55}$$

Due to behavior at the poles of $\mathbb{S}^2$, these extensions may not have the same regularity on $\mathbb{S}^2$ as the original function does on $[-1, 1]$. However, because of the tensor product construction, we expect the same convergence rates. For approximations of degree N, we use for $\mathcal{Q}$ (cf. (2.20)) a product quadrature rule on $\mathbb{S}^2$ that has degree of precision $2N + 1$. (This quadrature is defined in Section 2.2.2.3.) To ensure that our results do not depend on a special alignment of the quadrature with the coordinate axes, we rotate the points about the x and y axes by one and two radians, respectively.

Table 2.1 lists the observed L^2 convergence rates for functions of the form (2.55) with ψ defined in (2.51)–(2.54). The convergence rates for functions (2.51)–(2.54) on $[-1, 1]$, as seen in the legend of Figures 2.1–2.5, are also included for reference. We observe that, for most cases, the rates for the extended functions with rotated quadrature are close to the rates for the corresponding functions on $[-1, 1]$. Larger variations occur with the UD_N approximation, most noticeably for the singular function given in (2.52).

Finally, we consider general functions on $\mathbb{S}^2$. Convergence rates for these functions are presented in Figures 2.6 and 2.7. In Figure 2.6, the target function Ψ

Filter Order	**Approx. Type**	**Step** (2.51)		**Singular** (2.52)		**Smooth** (2.53), $q = \infty$		**Sobolev** (2.54), $q < 1$		**Sobolev** (2.54), $q < 3.5$	
		[-1, 1]	$\mathbb{S}^2$	[-1, 1]	$\mathbb{S}^2$	[-1, 1]	$\mathbb{S}^2$	[-1, 1]	$\mathbb{S}^2$	[-1, 1]	$\mathbb{S}^2$
None	P_N	0.49	0.51	0.53	0.50	∞	∞	0.97	1.33	3.49	3.47
	UD_N	0.08	0.06	-0.04	-0.22	∞	∞	0.21	0.06	3.09	2.92
	P_N^+	0.51	0.51	0.51	0.49	∞	∞	1.02	1.15	3.52	3.49
2	FP_N	0.49	0.51	0.52	0.50	1.99	1.95	0.97	1.32	1.99	1.96
	UD_N	0.09	0.10	-0.02	-0.23	1.99	1.95	0.25	0.05	2.03	2.20
	FP_N^+	0.51	0.51	0.51	0.49	1.99	1.95	1.02	1.15	1.99	1.96
4	FP_N	0.49	0.50	0.52	0.49	3.98	3.90	0.97	1.27	3.47	3.43
	UD_N	0.07	0.15	-0.05	-0.19	3.98	3.89	0.26	0.08	3.02	2.77
	FP_N^+	0.51	0.51	0.51	0.48	3.98	3.90	1.01	1.15	3.53	3.61
6	FP_N	0.49	0.47	0.44	0.40	5.96	5.84	0.98	1.07	3.47	3.41
	UD_N	0.10	0.23	0.05	0.00	5.96	5.81	0.18	0.11	3.04	2.86
	FP_N^+	0.49	0.47	0.45	0.41	5.96	5.81	0.97	1.05	3.42	3.39

Table 2.1: Convergence Rates – The observed L^2 convergence rates for the P_N, FP_N, UD_N, and FP_N^+ approximations to functions defined in (2.51) – (2.54) and their extensions on $\mathbb{S}^2$. Note that the index q express the regularity of the associated function on $[-1, 1]$.

is defined as

$$\Psi(\mu,\phi) = \begin{cases} 1, & \Omega_x \in [-0.2,\, 0.4], \Omega_y \in [0.5,\, 0.9] \\ 0, & \text{otherwise} \end{cases}, \tag{2.56}$$

where $\Omega_x = \sqrt{1-\mu^2}\cos\phi$ and $\Omega_y = \sqrt{1-\mu^2}\sin\phi$. This function is in $H^q(\mathbb{S}^2)$ for all $q < 0.5$. The location of the support for Ψ can be arbitrarily chosen; some choices may lead to faster convergence rates. For this particular choice, we observe that the UD_N approximation does not converge (or does so very slowly), while the FP_N^+ approximation converges with rate ≈ 0.5, just as the FP_N approximation does.

The next target function is given by

$$\Psi(\mu,\phi) = \psi_1(\mu)\psi_2(\phi), \tag{2.57}$$

where

$$\psi_1(\mu) = \begin{cases} 0.25, & |\mu| \in [0,\, 0.25) \\ 0.5 - |\mu|, & |\mu| \in [0.25,\, 0.5) \\ 0, & \text{else} \end{cases}, \psi_2(\phi) = \begin{cases} 0.25\pi, & |\phi| \in [0,\, 0.25\pi) \\ 0.5\pi - |\phi|, & |\phi| \in [0.25\pi,\, 0.5\pi) \\ 0, & \text{else} \end{cases}, \tag{2.58}$$

respectively. This function Ψ is in $H^q(\mathbb{S}^2)$, for all $q < 2$. Results related to this function are given in Figure 2.7. The convergence rate of the UD_N approximation is near one, as predicted by the error estimate given in Theorem 2. Hence, (2.44) appears to be a sharp error estimate for the UD_N approximation. The FP_N^+ approximation still converges at roughly the same rate as the FP_N approximation.

Remark 3. *In all the convergence tests we performed, the FP_N^+ approximation always converges at roughly the same rate as the FP_N approximation, even if the*

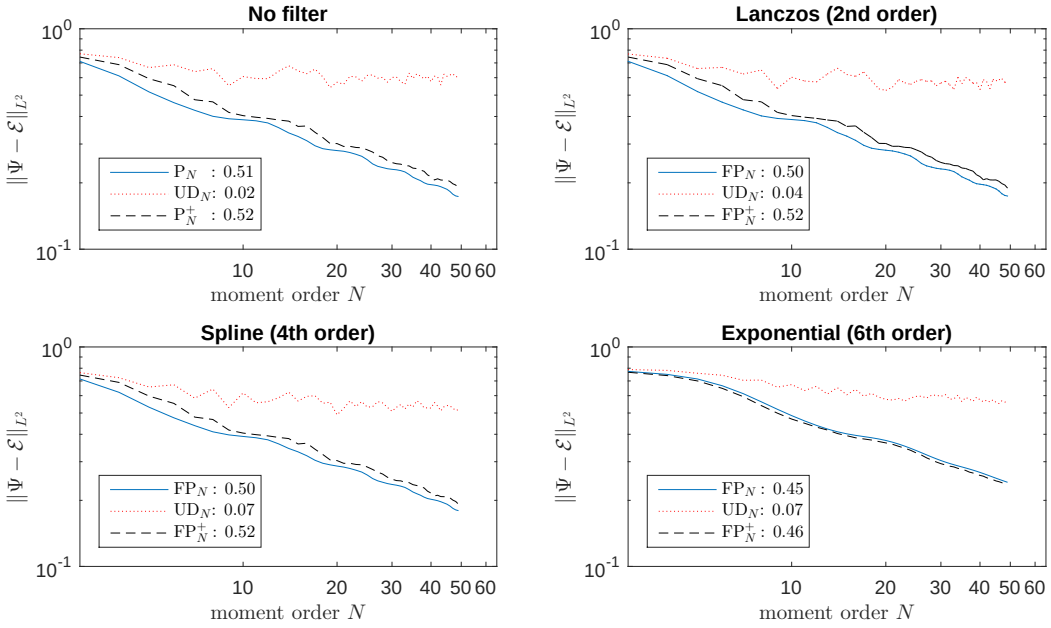


Figure 2.6: Step function Ψ on $\mathbb{S}^2$. $\Psi \in H^q(\mathbb{S}^2)$, for all $q < 0.5$; see (2.56). The observed convergence rates, as defined in Footnote 7, are listed in the legend.

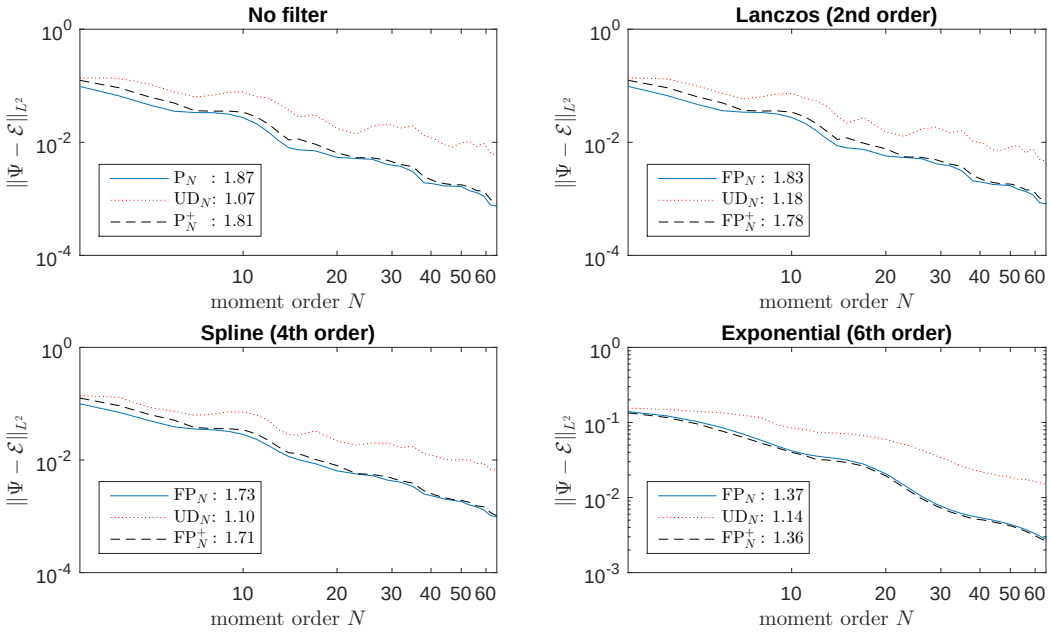


Figure 2.7: Sobolev function Ψ on $\mathbb{S}^2$. $\Psi \in H^q(\mathbb{S}^2)$, for all $q < 2$; see (2.57). The observed convergence rates, as defined in Footnote 7, are listed in the legend.

continuity assumption in Theorem 1 is violated, i.e., the target function belongs to H^q*, but not to* C^q*.*

2.4 Results on Line Source Benchmark Problem

In this section, we present solutions of the line source problem using the FP_N^+ closure and compare them to the results using P_N, FP_N, and PP_N closures (cf. Sections 2.1.2, 2.1.3, 2.1.4). Similar results for P_N, FP_N, and PP_N can be found in [62], [18] and [10], respectively. Results from the UD_N closure (cf. Section 2.1.5) are also included in the comparison.

2.4.1 The Line Source Benchmark

The line source benchmark problem was first formulated in [24], along with an exact solution. Since then, it has been used to study the behavior of various angular approximations for linear kinetic equations [9, 11, 18, 62]. It is a notoriously difficult problem that provides insight into the strengths and weaknesses of different approximations and how to pursue improvements.

The problem is as follows: Particles in an initial pulse are distributed isotropically along an infinite line in space and move through an infinite material medium with constant scattering cross-section. If this line is aligned with the z-axis, then f does not depend on z and the transport equation (2.1) reduces to

$$\partial_t f + \Omega_x \partial_x f + \Omega_y \partial_y = \frac{\sigma}{4\pi} \langle f \rangle - \sigma f \,, \tag{2.59}$$

with initial condition $f^{\mathrm{in}}(r, \Omega) = \frac{1}{4\pi} \delta(x, y)$.

2.4.2 Numerical Results

We simulate the line source problem with $\sigma = 1.0$. A steep Gaussian distribution with variance $\varsigma^2 = 9 \times 10^{-4}$ is used to approximate the delta function initial condition, and a small positive floor is added:

$$f^{\text{in}}(r, \Omega) \approx \frac{1}{4\pi} \left(\max \left(\frac{1}{2\pi\varsigma^2} e^{\frac{-(x^2+y^2)}{2\varsigma^2}}, f_{\text{floor}} \right) \right) . \tag{2.60}$$

This floor is only needed for the PP_N closure, which requires a strictly positive distribution. For our calculations, we set $f_{\text{floor}} = 10^{-4}$. We truncate the infinite spatial domain to a $[-1.5, 1.5] \times [-1.5, 1.5]$ square centered at the origin and impose artificial boundary condition equal to f_{floor}. The computation is run to a final time $t_{\text{final}} = 1.0$.

The calculations are performed using a 200×200 mesh ($\Delta x = \Delta y = 0.015$). The time step for the P_N and FP_N methods is $\Delta t = 0.45\Delta x$; for the UD_N, PP_N, and FP_N^+ methods is $\Delta t = 0.225\Delta x$ and a minmod-type slope limiter of $\vartheta = 2$ is used to enforce positivity in the kinetic scheme. See (A.7) and (A.8) in Appendix A.1. The more restrictive step is used to maintain positivity of the particle concentration for the FP_N^+, UD_N, and PP_N closures. In the numerical tests, we choose the filter coefficient $\sigma_F = 15$, which is used in FP_N, UD_N, and FP_N^+ closures.

Algorithm CR-MPC with Rule 2 (presented in Chapter 4) is used to solve the optimization problems in the FP_N^+ method. The algorithm parameter values used in the test are $\tau = 0.5$, $\zeta = 0.9$, $\varkappa = 0.98$, $\underline{\lambda} = 10^{-6}$, $\lambda^{\max} = 10^{30}$, and $\varepsilon = 10^{-4}$.

In Figures 2.8 and 2.9, we plot the particle concentration $\langle f \rangle$ for various meth-

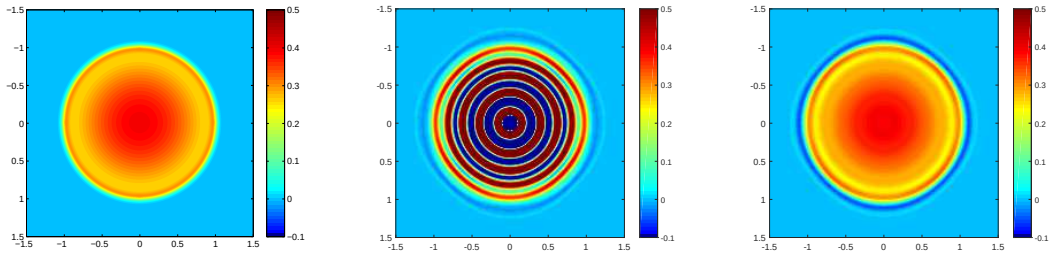
ods with moments of order $N = 11$ and quadrature precision of degree $N_{\mathcal{Q}} = 2N + 1 = 23$ (the minimum required precision) and $N_{\mathcal{Q}} = 47$. We consider both product and Lebedev quadrature rules; see Section 2.2.2.3 for details. Figure 2.8 shows the heat maps over the entire two-dimensional domain, and Figure 2.9 presents the one-dimensional line-outs along the x-axis. For comparison, the exact transport solution is included in all the line-out figures.

We observe the following qualitative features from the numerical results:

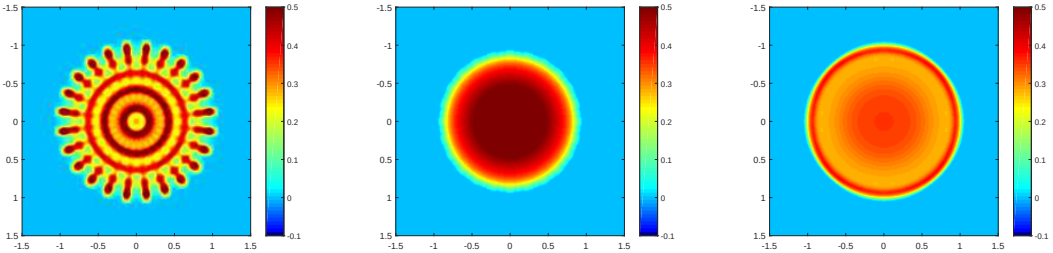
- P_N (Figures 2.8(b), 2.9(b)) The P_N method clearly suffers from severe oscillations that lead to particle concentrations with large negative values. The P_N solution preserves the rotational invariance of the exact line source solution and the quadrature has minimal effect on the P_N solution, as long as it has degree of precision $2N + 1$.

- FP_N (Figures 2.8(c), 2.9(c)) The FP_N solution contains only mild oscillations. Like the P_N method, the FP_N method maintains rotational invariance in the solution. However, it still suffers from the loss of positivity in the particle concentration, as can be seen near the wave front. Like the P_N solution, the FP_N solution is unaffected by the degree of quadrature precision $N_{\mathcal{Q}}$, as long as $N_{\mathcal{Q}} \geq 2N + 1$.

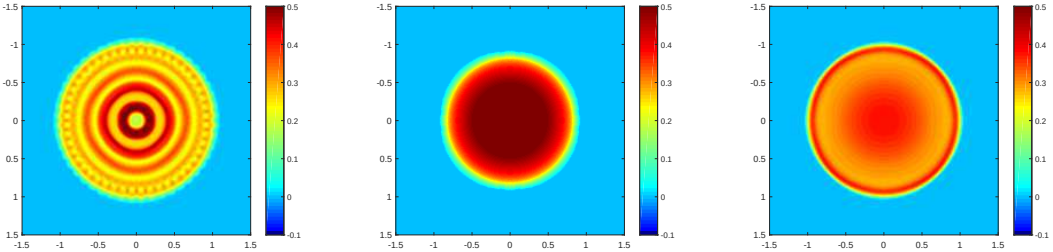
- PP_N (Figures 2.8(d), 2.8(g), 2.9(d), 2.9(g)) Oscillations still occur in the PP_N solution. However, they are much weaker than those occurring in the P_N solution. Because the PP_N closure uses a positive ansatz, the PP_N solution



(a) Exact solution (b) P_{11} (c) FP_{11}



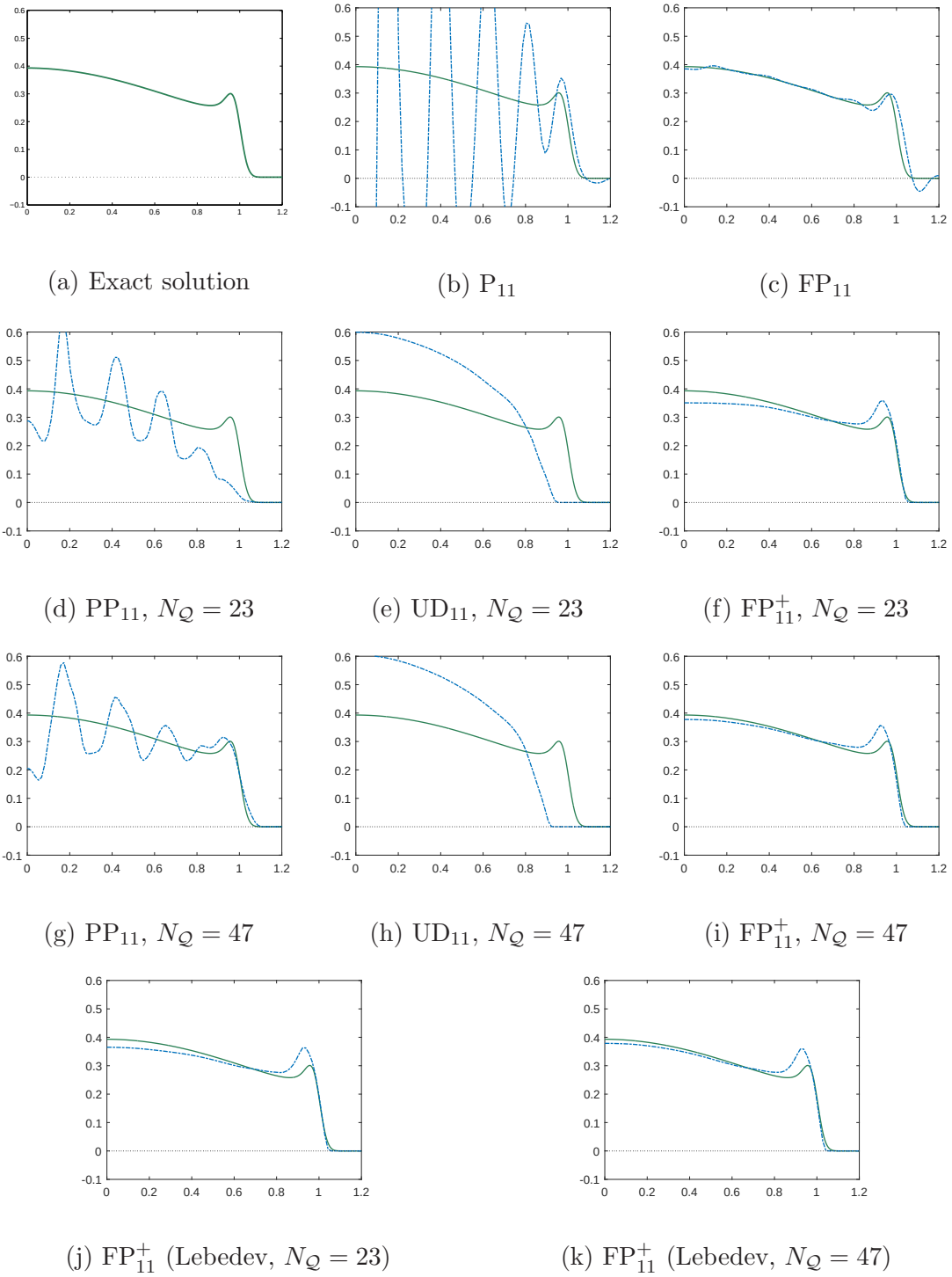
(d) PP_{11}, $N_{\mathcal{Q}} = 23$ (e) UD_{11}, $N_{\mathcal{Q}} = 23$ (f) FP_{11}^{+}, $N_{\mathcal{Q}} = 23$



(g) PP_{11}, $N_{\mathcal{Q}} = 47$ (h) UD_{11}, $N_{\mathcal{Q}} = 47$ (i) FP_{11}^{+}, $N_{\mathcal{Q}} = 47$

(j) FP_{11}^{+} (Lebedev, $N_{\mathcal{Q}} = 23$) (k) FP_{11}^{+} (Lebedev, $N_{\mathcal{Q}} = 47$)

Figure 2.8: Heat maps – the particle concentration $\langle f \rangle$ of the solutions to the line source benchmark for various methods.



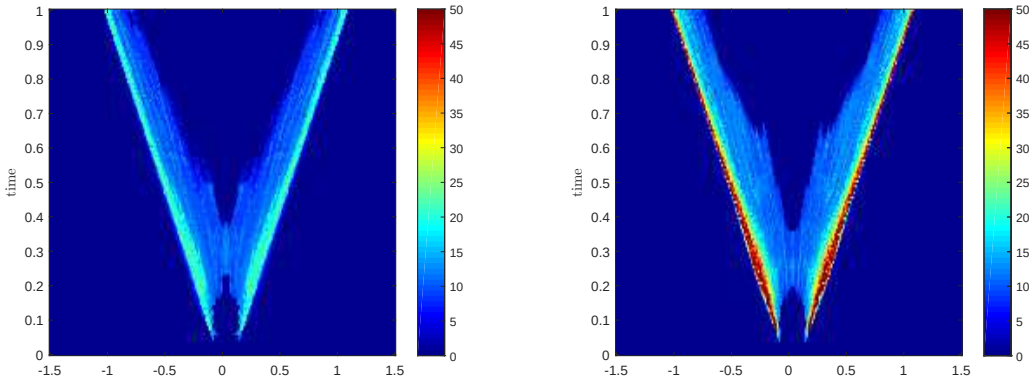
(a) Exact solution (b) P_{11} (c) FP_{11}

(d) PP_{11}, $N_{\mathcal{Q}} = 23$ (e) UD_{11}, $N_{\mathcal{Q}} = 23$ (f) FP_{11}^{+}, $N_{\mathcal{Q}} = 23$

(g) PP_{11}, $N_{\mathcal{Q}} = 47$ (h) UD_{11}, $N_{\mathcal{Q}} = 47$ (i) FP_{11}^{+}, $N_{\mathcal{Q}} = 47$

(j) FP_{11}^{+} (Lebedev, $N_{\mathcal{Q}} = 23$) (k) FP_{11}^{+} (Lebedev, $N_{\mathcal{Q}} = 47$)

Figure 2.9: Line-outs (along the x-axis) – the particle concentration $\langle f \rangle$ of the solutions to the line source benchmark for various methods.



(a) $N_{\mathcal{Q}} = 23$ (b) $N_{\mathcal{Q}} = 47$

Figure 2.10: The number of iterations needed to solve the optimization problem (2.23) for FP_{11}^{+} at each cell on the x-axis of the space and each time step.

maintains positivity in the particle concentration. However, because the ansatz is not polynomial, its moments cannot be evaluated exactly with a numerical quadrature rule. As a consequence, the PP_N solution loses rotational invariance and suffers from ray effects. Moreover, the accuracy of the PP_N solution is highly dependent on the quadrature precision.

- UD_N (Figures 2.8(e), 2.8(h), 2.9(e), 2.9(h)) The UD_N closure imposes strong damping which effectively removes all oscillations from the solution. The closure also maintains a positive particle concentration. However, the damping has a significant effect on accuracy; indeed, the UD_N solution completely misses the location of the wave front.

- FP_N^{+} (Figures 2.8(f), 2.8(i), 2.9(f), 2.9(i)) As expected, the FP_N^{+} solution preserves the positivity of the particle concentration. It contains only tiny oscil-

lations that are barely visible in the figures, which indicates that the nonlinear filter (constrained optimization) in the FP_N^+ method not only maintains the positivity of the ansatz, but also slightly damps the oscillations. This damping does reduce the accuracy of the solution near the origin, when compared to the FP_N results. Like the P_N and FP_N solutions, the FP_N^+ solution is also rotationally invariant. The accuracy of the FP_N^+ solution is slightly improved by using quadrature with a higher degree of precision. However, the computational cost of solving problem (2.20) may become prohibitive. (See Table 2.2 in Section 2.4.3 below.)

Remark 4 (Lebedev Quadrature)**.** *The Lebedev quadrature [50] requires fewer quadrature points than the product quadrature (see Section 2.2.2.3) does to achieve the same degree of precision. For comparison, we test the FP_N^+ closure with Lebedev quadrature rules that have degree of precision $N_{\mathcal{Q}} = 23$ and $N_{\mathcal{Q}} = 47$ on the line source problem, and the solutions are shown in Figures 2.8(j), 2.8(k), and 2.9(j), 2.9(k). With the Lebedev rule, the computation time is reduced by about 25%, due to the smaller number of constraints in optimization problem, as shown in Table 2.2.*

Remark 5 (Location of "hard" problems)**.** *In the numerical tests, we observed that most of the computation time of the FP_N^+ method is spent in solving the "hard" optimization problems that locate near the wave front, as seen in Figure 2.10 for quadrature precision $N_{\mathcal{Q}} = 23$ and $N_{\mathcal{Q}} = 47$.*

2.4.3 Computational Performance

In Table 2.2, we list the computation times for the line source calculations in Section 2.4.2. The P_N and FP_N methods are significantly faster because they (i) can take larger time steps, since positivity does not need to be enforced; (ii) have simpler flux evaluations; and (iii) most importantly, require no numerical optimization for their closure. The UD_N method has the least computation cost among all positive-preserving methods (UD_N, PP_N, FP_N^+), but still takes about twice the time of the P_N and FP_N methods. The PP_N method is by far the slowest. The computation time for the FP_N^+ method depends heavily on the number of quadrature points. Hence, the computation time using the Lebedev quadrature with degree of precision 23 and 47 is also reported in Table 2.2. As discussed in Remark 4, the Lebedev quadrature rule requires fewer points to reach the same degree of precision than the product quadrature, leading to lower computation time. Overall the FP_N^+ closure with Algorithm CR-MPC on the Lebedev quadrature is still slower than the simple UD_N closure. With degree of precision $N_{\mathcal{Q}} = 23$ (the minimum required), the computation time is about ten times that of the UD_N closure. In the next subsection, we compare efficiency of these methods, taking into account accuracy.

2.4.4 Efficiency

The ultimate goal in the development of the FP_N^+ closure is to generate an approximate solution of the transport equation that is accurate, preserves positivity of the particle concentration, and is efficient for challenging test problems when the

Quadrature Type	Product	Product	Lebedev	Lebedev
Degree	$N_{\mathcal{Q}} = 23$	$N_{\mathcal{Q}} = 47$	$N_{\mathcal{Q}} = 23$	$N_{\mathcal{Q}} = 47$
# of points	$\lvert\mathcal{Q}\rvert = 144$	$\lvert\mathcal{Q}\rvert = 576$	$\lvert\mathcal{Q}\rvert = 105$	$\lvert\mathcal{Q}\rvert = 401$
P_{11}	270	286	—	—
FP_{11}	272	287	—	—
UD_{11}	448	1732	—	—
PP_{11}	13798	49574	—	—
FP_{11}^{+}	5929	13925	4564	8963

Table 2.2: The computation times (sec) for the line source benchmark with various closures with $N = 11$. The optimization problems in the FP_N^+ closure are solved by Algorithm CR-MPC described in Chapter 4.

underlying solution lacks high regularity. To this end, we compare the efficiency of the FP_N^+ and UD_N closures by examining the cost and accuracy of solving the line source benchmark for different values of the moment order N. To allow for larger values of N, we use a smoother initial condition (a Gaussian distribution, as in (2.60), with variance $\varsigma^2 = 10^{-2}$), reduce the spatial mesh from 200×200 cells to 100×100 cells, and use only quadrature rules with $N_{\mathcal{Q}} = 2N + 1$ (the minimum required degree of precision). All other parameter values are identical to those listed in Section 2.4.2.

Figure 2.11 illustrates the efficiency comparison between the UD_N and FP_N^+ closures, the latter implemented with Algorithm CR-MPC. The FP_N^+ closure is

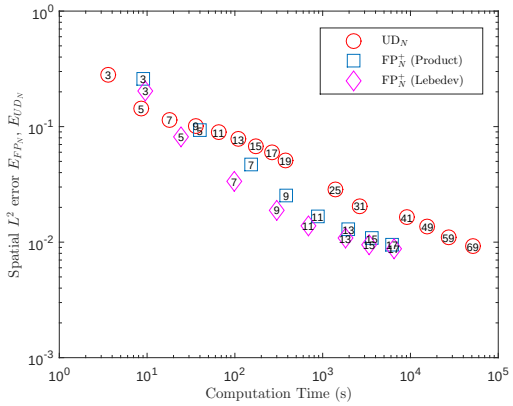


Figure 2.11: Efficiency Comparison – Each data point on the figure represents a solution of the moment equations, and the x-axis and y-axis are respectively the computation time and spatial error for the solution. The integers inside each symbol are the moment orders N. The FP_N^+ closure is implemented with Algorithm CR-MPC.

tested on both the product and Lebedev quadrature. We plot the spatial errors

$$E_{\mathrm{FP}_N^+} := \|\rho_{\text{exact}} - \rho_{\mathrm{FP}_N^+}\|_{L^2(\mathbb{R}^2)} \quad \text{and} \quad E_{\mathrm{UD}_N} := \|\rho_{\text{exact}} - \rho_{\mathrm{UD}_N}\|_{L^2(\mathbb{R}^2)}, \tag{2.61}$$

versus the computation time. Here ρ_{exact}, $\rho_{\mathrm{FP}_N^+}$, and ρ_{UD_N} are the particle concentration ($\rho := \frac{\langle f \rangle}{4\pi}$) at t_{final} of the exact, FP_N^+, and UD_N solutions, respectively. Each data point in Figure 2.11 represents a solution of the moment equations and is marked with a number that corresponds to the value of N. The data shows that, except for very low orders, the FP_N^+ solutions are two to ten times faster than the UD_N solutions to reach the same accuracy.

2.4.5 Space-Time Convergence

In this subsection, we compute space-time convergence rates of the finite volume scheme outlined in Appendix A when using the UD_N and FP_N^+ closures. Con-

vergence rates when using the FP_N closure are also included for reference. In the numerical tests reported in this section, the spectral filter is implemented in the filtered equation (2.14), and the FP_N, UD_N, and FP_N^+ closures are defined based on the moments $\mathbf{u}^*$ in (2.14). By doing so, we eliminate the influence of the spectral filter on the convergence properties of the numerical scheme (see [47]), so that the numerical results reflect only the effect of enforcing positivity in the UD_N and FP_N^+ closures.[8]

As before, we truncate the spatial domain to a $[-1.5, 1.5] \times [-1.5, 1.5]$ square centered at the origin and impose artificial boundary condition equal to f_{floor}. The computation is run to a final time $t_{\text{final}} = 1.0$. The numerical scheme is tested with initial condition

$$\rho^{\text{in}}(r) = \begin{cases} \cos^5(2\sqrt{x^2+y^2}), & \text{if } 2\sqrt{x^2+y^2} \leq \frac{\pi}{2}, \\ f_{\text{floor}}, & \text{otherwise}, \end{cases} \tag{2.62}$$

for the particle concentration. For $N > 0$, all moments are initially set to zero. All parameter values we used were identical to those used in the numerical tests reported in Section 2.4.2, except that the moment order N is chosen to be 5 and 7, instead of 11.

Since an analytic solution is not available in our problem, we define the space-

[8] We referred to this in Section 2.1.3 as the *continuous embedding* of the filter. With it, we expect (and observe) second-order space-time accuracy for the FP_N closure, whereas for the *embedded approach* that applies the filter at each time step, we expect (and observe) only first-order accuracy in time.

time error $E^p_{\Delta x}$ by

$$E^p_{\Delta x} := \|\mathbf{u}_{\Delta x} - \mathbf{u}_{\Delta x/2}\|_{L^p(\mathbb{R}^2, L^2(\mathbb{R}^n))}\,, \tag{2.63}$$

where $\mathbf{u}_{\Delta x}(r) \in \mathbb{R}^n$ is the computed solution to the moment equation with the finite volume scheme at $t_{\text{final}} = 1$, Δx denotes the side length of the square spatial cells, and the norm is defined as $\|\mathbf{v}\|_{L^p(\mathbb{R}^2, L^2(\mathbb{R}^n))} := \left(\int_{\mathbb{R}^2} \|\mathbf{v}(x)\|_2^p dx\right)^{1/p}$ for $p < \infty$, and $\|\mathbf{v}\|_{L^\infty(\mathbb{R}^2, L^2(\mathbb{R}^n))} := \max_{x\in\mathbb{R}^2} \|\mathbf{v}(x)\|_2$ for $p = \infty$.

Table 2.3 reports the space-time errors and observed convergence rates for FP_N, UD_N, and FP^+_N closures with $p = 1$ and $p = \infty$. The observed convergence rate ν is computed by

$$\nu := \log\left(\frac{E^p_{\Delta x_i}}{E^p_{\Delta x_{i+1}}}\right) \log\left(\frac{\Delta x_i}{\Delta x_{i+1}}\right)^{-1}. \tag{2.64}$$

The results for moment order $N = 5$ and $N = 7$ are presented. These results indicate that the FP^+_N closure has minimal effect on the convergence rate[9], while the UD_N closure causes a serious degradation in the convergence order.

[9]The only noticeable difference is the error $E^\infty_{\Delta x}$ for the FP^+_5 solution with the 320^2 mesh.

	FP_5		UD_5		FP_5^+		FP_7		UD_7		FP_7^+	
mesh	$E^1_{\Delta x}$	ν	$E^1_{\Delta x}$	ν	$E^1_{\Delta x}$	ν	$E^1_{\Delta x}$	ν	$E^1_{\Delta x}$	ν	$E^1_{\Delta x}$	ν
20^2	4.9e-3	—	1.5e-2	—	5.7e-3	—	5.8e-3	—	1.4e-2	—	6.2e-3	—
40^2	1.48e-3	1.7	1.4e-3	3.4	1.3e-3	2.1	1.8e-3	1.7	1.7e-3	3.0	1.6e-3	2.0
80^2	3.7e-4	2.0	6.9e-4	1.1	3.6e-4	1.9	4.4e-4	2.0	7.7e-4	1.2	4.3e-4	1.9
160^2	8.9e-5	2.0	1.3e-3	-0.9	8.7e-5	2.1	1.1e-4	2.0	8.6e-4	-0.2	1.0e-4	2.1
320^2	2.2e-5	2.0	2.6e-3	-1.0	2.2e-5	2.0	—	—	—	—	—	—
	$E^\infty_{\Delta x}$	ν	$E^\infty_{\Delta x}$	ν	$E^\infty_{\Delta x}$	ν	$E^\infty_{\Delta x}$	ν	$E^\infty_{\Delta x}$	ν	$E^\infty_{\Delta x}$	ν
20^2	1.1e-2	—	4.7e-2	—	1.7e-2	—	1.2e-2	—	4.4e-2	—	1.6e-2	—
40^2	4.0e-3	1.5	6.0e-3	3.0	5.0e-3	1.8	4.3e-3	1.5	7.2e-3	2.6	5.1e-3	1.7
80^2	1.0e-3	1.9	7.2e-3	-0.3	1.2e-3	2.0	1.1e-3	1.9	9.0e-3	-0.3	1.1e-3	2.2
160^2	2.5e-4	2.0	2.3e-2	-1.7	2.7e-4	2.2	2.8e-4	2.0	2.0e-2	-1.1	2.8e-4	2.0
320^2	6.2e-5	2.0	3.9e-2	-0.8	8.0e-5	1.8	—	—	—	—	—	—

Table 2.3: Convergence of space-time errors with $p = 1$ and $p = \infty$ for FP_N, UD_N, and FP_N^+ closures. The results for moment orders $N = 5$ and $N = 7$ are reported. The value N_x is the number of spatial cells in each direction of the square domain. In order to minimize the influence of the optimization tolerance in the FP_N^+ method, the tolerance ε is set to 10^{-8}.

Chapter 3: A Positive Asymptotic Preserving (AP) Scheme

In this chapter, we move our focus to the improvement on the efficiency of the numerical scheme used to integrate the moments equations. As discussed in [10], the second-order kinetic scheme (refer to Appendix A) used in Chapter 2 is very inefficient in diffusion regimes, where strong scattering occurs and long time scale is of interest. (See [15] and citations therein for more details.) More specifically, letting $\sigma_s \to \epsilon^{-1}\sigma_s$ and $t \to \epsilon^{-1}t$ in (2.1) for small $\epsilon > 0$ leads to the following scaled equation

$$\epsilon\partial_t f + \Omega \cdot \nabla_r f = \frac{\sigma_s}{\epsilon}\left(\frac{1}{4\pi}\langle f\rangle - f\right) . \tag{3.1}$$

Here we allow for the case that the scattering cross-section $\sigma_s = \sigma_s(r)$ is a function of spatial position r.

It is well-known [26, 63] that, when $\epsilon \ll 1$, the kinetic distribution f in (3.1) is given by $f = \rho + O(\epsilon)$, with $\rho := \langle f\rangle/4\pi$ the particle concentration. Meanwhile, ρ is governed by the diffusion equation

$$\partial_t\rho - \nabla_r \cdot (D\nabla_r\rho) = O(\epsilon) , \tag{3.2}$$

where the matrix of diffusion coefficients D is given by

$$D = \frac{1}{4\pi\sigma_s}\,\mathrm{diag}\left(\langle\Omega_x^2\rangle,\, \langle\Omega_y^2\rangle,\, \langle\Omega_z^2\rangle\right) = \frac{1}{3\sigma_s}I_{3\times 3} . \tag{3.3}$$

The diffusion equation (3.2) is referred as the diffusion limit.

For the second-order kinetic scheme used in Chapter 2, the space accuracy requirements dictate that the spatial mesh depends on ϵ, and the stability requirement for this explicit scheme is $\Delta t = O(\epsilon\Delta x, \epsilon^2)$, where $\epsilon\Delta x$ is required by the hyperbolic time scale, and ϵ^2 is required by the collisional time scale. These restrictions do not affect the numerical tests presented in Section 2.4 (where $\epsilon = 1$), but certainly limit the efficiency of the scheme when ϵ is small.

A variety of asymptotic preserving (AP) schemes are proposed to preserve stability and accuracy when solving transport equations near the diffusion limit, without resolving the size of the temporal-spatial mesh. Such schemes often require decompositions on the distribution function. A macro-micro decomposition approach is introduced in [35], and various AP schemes are proposed using the even-odd decomposition – a finite element approach is proposed in [64], while [32, 33, 65] introduced several finite difference approaches. A staggered finite difference method is used in a recent work [66].

In this chapter, we propose a positive preserving AP scheme for solving

$$\epsilon\partial_t \mathbf{u} + \nabla_r \cdot \langle \mathbf{m}\Omega\mathcal{E}[\mathbf{u}]\rangle = -\frac{\sigma_s}{\epsilon} R\mathbf{u}\,, \tag{3.4}$$

which is the moment equation associated with (3.1). We prove that the proposed scheme indeed computes the correct diffusion limit as $\epsilon \to 0$, without the strict restrictions on the spatial mesh size and time step.

In Section 3.1, we present the proposed AP scheme based on even-odd decomposition, and introduce the spatial discretization methods used in the scheme.

We prove stability conditions and positivity preserving properties in Section 3.2. We compare the proposed AP scheme to the second-order kinetic scheme proposed in [10, 15] on the line source problem under various temporal-spatial scales, and the results are reported in Section 3.3.

3.1 The AP Scheme

In this section, we first derive an AP scheme for the transport equation (3.1) using even-odd decomposition. We then show that our proposed AP scheme for the moment equation (3.4) can be obtained by taking moments from the first scheme. The details of the finite difference spatial discretization method used in the proposed AP scheme are also presented.

3.1.1 An AP Scheme for Transport Equations

Let us first define $f_+(r,\Omega,t) := f(r,\Omega,t)$ and $f_-(r,\Omega,t) := f(r,-\Omega,t)$, with $-\Omega := (-\Omega_x, -\Omega_y, -\Omega_z)$. The even and odd parity terms of f are then given by

$$f_E := \frac{1}{2}\left(f_+ + f_-\right) \ , \text{ and } f_O := \frac{1}{2}\left(f_+ - f_-\right) \ . \tag{3.5}$$

The scaled transport equation (3.1) can then be decomposed into the even equation

$$\epsilon\partial_t f_E + \Omega\cdot\nabla_r f_O = \frac{\sigma_s}{\epsilon}\left(\frac{1}{4\pi}\langle f_E\rangle - f_E\right) \ , \tag{3.6}$$

and the odd equation

$$\epsilon\partial_t f_O + \Omega\cdot\nabla_r f_E = -\frac{\sigma_s}{\epsilon} f_O \ . \tag{3.7}$$

Note that since f_E and f_O are even and odd functions on Ω, respectively, we have $\langle f_E \rangle = \langle f \rangle$ and $\langle f_O \rangle = 0$.

With initial time t_0 and temporal discretization $t^k := t_0 + k\Delta t$, (3.6) and (3.7) can be written in the following semi-discrete forms

$$\frac{f_E^{k+1} - f_E^k}{\Delta t/\epsilon} + \Omega \cdot \nabla_r f_O^{k+1} = \frac{\sigma_s}{\epsilon}\left(\frac{1}{4\pi}\langle f_E^{k+1}\rangle - f_E^{k+1}\right), \tag{3.8}$$

and

$$\frac{f_O^{k+1} - f_O^k}{\Delta t/\epsilon} + \Omega \cdot \nabla_r f_E^k = -\frac{\sigma_s}{\epsilon} f_O^{k+1}, \tag{3.9}$$

where f_E^k and f_O^k are the approximations to $f_E(\cdot,\cdot,t^k)$ and $f_O(\cdot,\cdot,t^k)$, respectively. Note that we treat the flux term in the odd equation (3.9) explicitly, and all the other terms implicitly. As shown in [67], such choice introduces a diffusion correction term into the transport scheme. Such property is verified in the following paragraphs.

After some trivial algebraic work, (3.8) and (3.9) can be rewritten as

$$\left(1 + \frac{\sigma_s \Delta t}{\epsilon^2}\right) f_E^{k+1} - \frac{\sigma_s \Delta t}{\epsilon^2}\frac{1}{4\pi}\langle f_E^{k+1}\rangle = f_E^k - \frac{\Delta t}{\epsilon}\Omega \cdot \nabla_r f_O^{k+1}, \tag{3.10}$$

and

$$\left(1 + \frac{\sigma_s \Delta t}{\epsilon^2}\right) f_O^{k+1} = f_O^k - \frac{\Delta t}{\epsilon}\Omega \cdot \nabla_r f_E^k, \tag{3.11}$$

respectively. Let $\gamma := \frac{\epsilon^2}{\epsilon^2 + \sigma_s \Delta t}$. Multiplying γ on both sides of (3.10) and (3.11) yields

$$f_E^{k+1} - (1-\gamma)\frac{1}{4\pi}\langle f_E^{k+1}\rangle = \gamma f_E^k - \frac{\Delta t}{\epsilon}\gamma\,\Omega \cdot \nabla_r f_O^{k+1}, \tag{3.12}$$

and

$$f_O^{k+1} = \gamma f_O^k - \frac{\Delta t}{\epsilon}\gamma\,\Omega \cdot \nabla_r f_E^k. \tag{3.13}$$

To avoid non-conservative products (see [68–70] for details), we perform a change of variables by letting $h := \gamma^{-1} f$, and h_+ and h_- are defined analogously. Thus, we have $h_E := \gamma^{-1} f_E$ and $h_O := \gamma^{-1} f_O$. Applying the change of variables on (3.12) and (3.13) leads to

$$h_E^{k+1} - (1-\gamma)\frac{1}{4\pi}\langle h_E^{k+1}\rangle = \gamma h_E^k - \frac{\Delta t}{\epsilon}\Omega\cdot\nabla_r \gamma h_O^{k+1}\,, \tag{3.14}$$

and

$$h_O^{k+1} = \gamma h_O^k - \frac{\Delta t}{\epsilon}\Omega\cdot\nabla_r \gamma h_E^k\,, \tag{3.15}$$

respectively. Note that (3.15) is already a fully explicit scheme for the odd equation, while there are still implicit terms involved in (3.14). Since the implicit term h_O^{k+1} on the right-hand-side of (3.14) is given in (3.15), we then plug (3.15) into (3.14), and obtain

$$h_E^{k+1} - (1-\gamma)\frac{\langle h_E^{k+1}\rangle}{4\pi} = \gamma h_E^k - \frac{\Delta t}{\epsilon}\Omega\cdot\nabla_r \gamma^2 h_O^k + \frac{\Delta t^2}{\epsilon^2}\Omega\cdot\nabla_r\left(\gamma\,\Omega\cdot\nabla_r \gamma h_E^k\right)\,. \tag{3.16}$$

Combining (3.15) and (3.16) forms an explicit, semi-discrete, even-odd parity scheme for solving the linear transport equation (3.1).

3.1.2 An AP Scheme for Moment Equations

In order to apply the scheme (3.15)–(3.16) on the moment equation (3.4), we define the even and odd moments as

$$\mathbf{u}_E := \langle \mathbf{m}_E f_E\rangle\,, \quad \text{and} \quad \mathbf{u}_O := \langle \mathbf{m}_O f_O\rangle\,, \tag{3.17}$$

where $\mathbf{m}_E$ and $\mathbf{m}_O$ are spherical harmonic functions of even and odd degrees, respectively. By the properties of spherical harmonic functions, it is not difficult to verify

that the odd degree moments of even functions f_E and the even degree moments of odd functions f_O always vanish. Hence they are omitted in the formulation. The scaled moments $\mathbf{v} := \gamma^{-1}\mathbf{u}$ are then analogously decomposed into

$$\mathbf{v}_E := \gamma^{-1}\mathbf{u}_E = \langle \mathbf{m}_E h_E \rangle \,, \quad \text{and} \quad \mathbf{v}_O := \gamma^{-1}\mathbf{u}_O = \langle \mathbf{m}_O h_O \rangle \,, \tag{3.18}$$

and the associated even and odd ansatzes are given by

$$\mathcal{E}_E[\mathbf{v}_E] := \mathbf{m}_E^T \mathbf{v}_E \,, \quad \text{and} \quad \mathcal{E}_O[\mathbf{v}_O] := \mathbf{m}_O^T \mathbf{v}_O \,. \tag{3.19}$$

With this setup, taking the moments from the transport scheme (3.15)–(3.16) leads to the following scheme for solving the moment equation (3.4)

$$\mathbf{v}_O^{k+1} = \gamma \mathbf{v}_O^k - \frac{\Delta t}{\epsilon} \nabla_r \cdot \gamma \left\langle \Omega \mathbf{m}_O \mathcal{E}_E[\mathbf{v}_E^k] \right\rangle \,, \tag{3.20}$$

and

$$\begin{aligned} T\mathbf{v}_E^{k+1} = \gamma \mathbf{v}_E^k - \frac{\Delta t}{\epsilon} \nabla_r \cdot \gamma^2 \left\langle \Omega \mathbf{m}_E \mathcal{E}_O[\mathbf{v}_O^k] \right\rangle \\ + \frac{\Delta t^2}{\epsilon^2} \nabla_r \cdot \left(\gamma \left\langle (\Omega \otimes \Omega) \nabla_r \gamma \mathbf{m}_E \mathcal{E}_E[\mathbf{v}_E^k] \right\rangle \right) \,, \end{aligned} \tag{3.21}$$

where the matrix $T = \mathrm{diag}(\gamma, 1, 1, \ldots, 1)$.

We next show that, when $\epsilon \to 0$, the limit of proposed scheme (3.20)–(3.21) is consistent with a semi-discrete scheme for the diffusion equation (3.2).

Asymptotic preserving property

Let us first take the limits on (3.20) and (3.21) when ϵ tends to zero. By definition of γ, we have

$$\gamma \to 0 \,, \quad \frac{\gamma}{\epsilon} \to 0 \,, \quad \text{and} \quad \frac{\gamma}{\epsilon^2} \to \frac{1}{\sigma_s \Delta t} \,, \quad \text{as } \epsilon \to 0 \,. \tag{3.22}$$

Thus, from (3.20), it is clear that $\mathbf{v}_O$ tends to zero as $\epsilon \to 0$. In addition, it follows from (3.21) that all components of $\mathbf{v}_E$ also tends to zero when $\epsilon \to 0$, with the

exception of the first component. Let v_0 be the first component of $\mathbf{v}_E$. When $\epsilon \to 0$, the first equation of (3.21) then becomes

$$\gamma\, v_0^{k+1} = \gamma\, v_0^k + \Delta t \nabla_r \cdot \left(\frac{1}{\sigma_s} \left\langle (\Omega \otimes \Omega) \nabla_r \gamma m_0 \mathcal{E}_E[\mathbf{v}_E^k] \right\rangle \right) , \tag{3.23}$$

where $m_0 = \frac{1}{\sqrt{4\pi}}$ is the first component of $\mathbf{m}$, which is the normalized spherical harmonic function of degree 0. Since all the other components of $\mathbf{v}_E$ tend to zero, $\mathcal{E}_E[\mathbf{v}_E^k]$ is simply given by $\mathcal{E}_E[\mathbf{v}_E^k] = m_0 v_0^k$. Plugging this into (3.23) yields

$$\gamma\, v_0^{k+1} = \gamma\, v_0^k + \Delta t \nabla_r \cdot \left(\frac{1}{\sigma_s} \left\langle \frac{(\Omega \otimes \Omega)}{4\pi} \right\rangle \nabla_r \gamma v_0^k \right) , \tag{3.24}$$

and it can be verified via direct evaluation that the integral of matrix $\frac{(\Omega \otimes \Omega)}{4\pi}$ over $\mathbb{S}^2$ takes value $\frac{1}{3} I$.

On the other hand, the particle concentration ρ in (3.2) is defined as $\rho := \langle f \rangle / 4\pi$. From the definitions of f_E, h_E, and $\mathbf{v}_E$, we have

$$\rho = \frac{\langle f_E \rangle}{4\pi} = \frac{\gamma \langle h_E \rangle}{4\pi} = \frac{\gamma\, v_0}{\sqrt{4\pi}} , \tag{3.25}$$

Thus, by dividing both sides of (3.24) by $\sqrt{4\pi}$, we obtain

$$\rho^{k+1} = \rho^k + \Delta t \nabla_r \cdot \left(\frac{1}{3\sigma_s} I\, \nabla_r \rho^k \right) , \tag{3.26}$$

which is exactly a semi-discrete scheme for the diffusion equation (3.2). Hence, we confirm that the proposed scheme indeed achieves the correct diffusion limit when ϵ tends to zero.

3.1.3 Spatial Discretization

For simplicity of exposition, we restrict ourselves to a reduced equation starting from this section. The reduced equation is used in the formulation of the line source

benchmark in Section 2.4.1, which is given by

$$\epsilon\partial_t f + \xi\partial_x f + \eta\partial_y f = \frac{\sigma_s}{\epsilon}\left(\frac{1}{4\pi}\langle f\rangle - f\right) , \tag{3.27}$$

where $(\xi, \eta) := (\Omega_x, \Omega_y)$. This equation is valid when $\partial_z f = 0$ [3], and the associated closed moment equation with ansatz $\mathcal{E}$ becomes

$$\epsilon\partial_t \mathbf{u} + \partial_x\langle \mathbf{m}\xi\mathcal{E}[\mathbf{u}]\rangle + \partial_y\langle \mathbf{m}\eta\mathcal{E}[\mathbf{u}]\rangle = -\frac{\sigma_s}{\epsilon}R\mathbf{u} . \tag{3.28}$$

Applying the moment scheme (3.20)–(3.21) on (3.28) leads to the reduced scheme

$$\mathbf{v}_O^{k+1} = \gamma\mathbf{v}_O^k - \frac{\Delta t}{\epsilon}\left\langle \mathbf{m}_O\xi\partial_x\gamma\mathcal{E}_E[\mathbf{v}_E^k] + \mathbf{m}_O\eta\gamma\partial_y\mathcal{E}_E[\mathbf{v}_E^k]\right\rangle , \tag{3.29}$$

and

$$\begin{aligned} T\mathbf{v}_E^{k+1} = \gamma\mathbf{v}_E^k &- \frac{\Delta t}{\epsilon}\left\langle \mathbf{m}_E\xi\partial_x\gamma^2\mathcal{E}_O[\mathbf{v}_O^k] + \mathbf{m}_E\eta\partial_y\gamma^2\mathcal{E}_O[\mathbf{v}_O^k]\right\rangle \\ &+\frac{\Delta t^2}{\epsilon^2}\left\langle \mathbf{m}_E\xi^2\partial_x\gamma\partial_x\gamma\mathcal{E}_E[\mathbf{v}_E^k] + \mathbf{m}_E\eta^2\partial_y\gamma\partial_y\gamma\mathcal{E}_E[\mathbf{v}_E^k]\right. \\ &\qquad \left.+\mathbf{m}_E\xi\eta\partial_x\gamma\partial_y\gamma\mathcal{E}_E[\mathbf{v}_E^k] + \mathbf{m}_E\eta\xi\partial_y\gamma\partial_x\gamma\mathcal{E}_E[\mathbf{v}_E^k]\right\rangle . \end{aligned} \tag{3.30}$$

The reduced scheme requires the first, second and mixed partial derivatives on directions x and y. We approximate these derivatives with a finite difference discretization on a uniform spatial mesh. For rectangular domains $[x_L, x_R] \times [y_L, y_R]$, each point on the mesh is defined as (x_i, y_j) where $x_i := x_L + i\Delta x$, and $y_j := y_L + j\Delta y$. The details on the discretization for the derivatives are presented in Section 3.1.3.1 and 3.1.3.2.

3.1.3.1 First Derivatives

We approximate the first derivatives in (3.29) and (3.30) with a second-order upwind scheme with the minmod flux limiter. For example, at point (x_i, y_j), the

first derivative terms in (3.29) are approximated by

$$(\xi\partial_x\gamma\mathcal{E}_E[\mathbf{v}_E])_{i,j} = \frac{\xi}{\Delta x}\left(\gamma_{i+1/2,j}\mathcal{E}_{E,i+1/2,j} - \gamma_{i-1/2,j}\mathcal{E}_{E,i-1/2,j}\right) , \tag{3.31}$$

and

$$(\eta\partial_y\gamma\mathcal{E}_E[\mathbf{v}_E])_{i,j} = \frac{\eta}{\Delta y}\left(\gamma_{i,j+1/2}\mathcal{E}_{E,i,j+1/2} - \gamma_{i,j-1/2}\mathcal{E}_{E,i,j-1/2}\right) , \tag{3.32}$$

where $\gamma_{i\pm1/2,\,j}$, $\gamma_{i,\,j\pm1/2}$ are given by

$$\gamma_{i+1/2,\,j} := \frac{\epsilon^2}{\epsilon^2 + \sigma_{s,i+1/2,j}\Delta t} , \text{ and } \gamma_{i,\,j+1/2} := \frac{\epsilon^2}{\epsilon^2 + \sigma_{s,i,j+1/2}\Delta t} , \tag{3.33}$$

with

$$\sigma_{s,i+1/2,j} = \frac{1}{2}(\sigma_{s,i+1,j} + \sigma_{s,i,j}) , \text{ and } \sigma_{s,i,j+1/2} = \frac{1}{2}(\sigma_{s,i,j+1} + \sigma_{s,i,j}) . \tag{3.34}$$

$\mathcal{E}_{E,i\pm1/2,j}$ and $\mathcal{E}_{E,i,j\pm1/2}$ are approximations to the ansatz $\mathcal{E}_E[\mathbf{v}_E]$. To compute $\mathcal{E}_{E,i\pm1/2,j}$ and $\mathcal{E}_{E,i,j\pm1/2}$ with the upwind scheme, we must first decompose the even ansatz $\mathcal{E}_E$ into $\mathcal{E}_+$ and $\mathcal{E}_-$, where $\mathcal{E}_+ := \mathcal{E}[\mathbf{v}](\Omega)$ and $\mathcal{E}_- := \mathcal{E}[\mathbf{v}](-\Omega)$. Note that since $\mathcal{E}_-$ takes values on $-\Omega$, the upwind direction for $\mathcal{E}_-$ is opposite to the upwind direction for $\mathcal{E}_+$. The upwind computation is then performed on values of $\mathcal{E}_+$ and $\mathcal{E}_-$. For example, in the computation of $\mathcal{E}_{E,i+1/2,j}$, the properties of spherical harmonic functions allow us to split it as

$$\mathcal{E}_{E,i+1/2,j} = \frac{1}{2}\left(\mathcal{E}_{+,i+1/2,j} + \mathcal{E}_{-,i+1/2,j}\right) , \tag{3.35}$$

with $\mathcal{E}_{+,i+1/2,j}$ and $\mathcal{E}_{-,i+1/2,j}$ defined via upwinding

$$\mathcal{E}_{+,i+1/2,j} := \begin{cases} \mathcal{E}_{+,i,j} + \frac{s^x_{+,i,j}}{2} , & \xi > 0 \\ \mathcal{E}_{+,i+1,j} - \frac{s^x_{+,i+1,j}}{2} , & \xi < 0 \end{cases} , \quad \mathcal{E}_{-,i+1/2,j} := \begin{cases} \mathcal{E}_{-,i+1,j} - \frac{s^x_{-,i+1,j}}{2} , & \xi > 0 \\ \mathcal{E}_{-,i,j} + \frac{s^x_{-,i,j}}{2} , & \xi < 0 \end{cases} , \tag{3.36}$$

where $s^x_{+,i,j}$ and $s^x_{-,i,j}$ approximates the spatial derivative on the x direction of $\mathcal{E}_+$ and $\mathcal{E}_-$, respectively.

In order to preserve positivity of the particle concentration, a minmod limiter is required on the approximation of spatial derivatives. Thus, the approximations are given as

$$s^x_{+,i,j} = \text{minmod}\left\{\vartheta(\mathcal{E}_{+,i,j} - \mathcal{E}_{+,i-1,j}), \frac{\mathcal{E}_{+,i+1,j} - \mathcal{E}_{+,i-1,j}}{2}, \vartheta(\mathcal{E}_{+,i+1,j} - \mathcal{E}_{+,i,j})\right\}, \tag{3.37}$$

and

$$s^x_{-,i,j} = \text{minmod}\left\{\vartheta(\mathcal{E}_{-,i,j} - \mathcal{E}_{-,i-1,j}), \frac{\mathcal{E}_{-,i+1,j} - \mathcal{E}_{-,i-1,j}}{2}, \vartheta(\mathcal{E}_{-,i+1,j} - \mathcal{E}_{-,i,j})\right\}, \tag{3.38}$$

where $1 \leq \vartheta \leq 2$ [71, 72][1] and

$$\text{minmod}(a,b,c) := \begin{cases} \min\{a,b,c\}\,, & a,b,c > 0 \\ \max\{a,b,c\}\,, & a,b,c < 0 \\ 0\,, & \text{otherwise} \end{cases}. \tag{3.39}$$

Following the steps (3.31)–(3.39), all other first derivative terms in (3.29) and (3.30) can be computed analogously.

3.1.3.2 Second Derivatives and Mixed Derivatives

In (3.30), there are second derivative and mixed partial derivative terms involved. In this section, we present the spatial discretization methods used for these

[1] Any value of $\vartheta \in [1,2]$ will yield a formally second-order scheme; roughly speaking, larger values of ϑ decrease numerical diffusion in the scheme.

terms.

For second derivatives on directions x and y, we use the standard second-order central difference approximation. For example, at point (x_i, y_j), the second derivative terms in (3.30) are approximated by

$$\begin{aligned}\left(\xi^2\partial_x\gamma\partial_x\gamma\mathcal{E}_E[\mathbf{v}_E]\right)_{i,j} = \frac{\xi^2}{\Delta x^2}(&\gamma_{i+1/2,j}(\gamma_{i+1,j}\mathcal{E}_{E,i+1,j} - \gamma_{i,j}\mathcal{E}_{E,i,j}) \\ &-\gamma_{i-1/2,j}(\gamma_{i,j}\mathcal{E}_{E,i,j} - \gamma_{i-1,j}\mathcal{E}_{E,i-1,j}))\,,\end{aligned} \tag{3.40}$$

and

$$\begin{aligned}\left(\eta^2\partial_y\gamma\partial_y\gamma\mathcal{E}_E[\mathbf{v}_E]\right)_{i,j} = \frac{\eta^2}{\Delta y^2}(&\gamma_{i,j+1/2}(\gamma_{i,j+1}\mathcal{E}_{E,i,j+1} - \gamma_{i,j}\mathcal{E}_{E,i,j}) \\ &-\gamma_{i,j-1/2}(\gamma_{i,j}\mathcal{E}_{E,i,j} - \gamma_{i,j-1}\mathcal{E}_{E,i,j-1}))\,,\end{aligned} \tag{3.41}$$

with $\gamma_{i\pm1/2,\,j}$, $\gamma_{i,\,j\pm1/2}$ defined in (3.33), and $\mathcal{E}_{E,i\pm1,j\pm1} = \mathcal{E}_E[\mathbf{v}_{E,i\pm1,j\pm1}]$.

For the mixed partial derivative terms, the second-order central difference approximation gives

$$\begin{aligned}(\xi\eta\partial_x\gamma\partial_y\gamma\mathcal{E}_E[\mathbf{v}_E])_{i,j} = \frac{\xi\eta}{4\Delta x\Delta y}(&\gamma_{i+1,j}(\gamma_{i+1,j+1}\mathcal{E}_{E,i+1,j+1} - \gamma_{i+1,j-1}\mathcal{E}_{E,i+1,j-1}) \\ &-\gamma_{i-1,j}(\gamma_{i-1,j+1}\mathcal{E}_{E,i-1,j+1} - \gamma_{i-1,j-1}\mathcal{E}_{E,i-1,j-1}))\,,\end{aligned} \tag{3.42}$$

and

$$\begin{aligned}(\eta\xi\partial_y\gamma\partial_x\gamma\mathcal{E}_E[\mathbf{v}_E])_{i,j} = \frac{\xi\eta}{4\Delta x\Delta y}(&\gamma_{i,j+1}(\gamma_{i+1,j+1}\mathcal{E}_{E,i+1,j+1} - \gamma_{i-1,j+1}\mathcal{E}_{E,i-1,j+1}) \\ &-\gamma_{i,j-1}(\gamma_{i+1,j-1}\mathcal{E}_{E,i+1,j-1} - \gamma_{i-1,j-1}\mathcal{E}_{E,i-1,j-1}))\,,\end{aligned} \tag{3.43}$$

However, the scheme does not preserve positivity of the solution if the second-order approximations given in (3.42) and (3.43) are used.

We proposed a first-order method to approximate the mixed partial derivative terms based on applying first-order upwind method direction by direction. Let

us take the term $\xi\eta\partial_x\gamma\partial_y\gamma\mathcal{E}_E[\mathbf{v}_E]$ as an example. At point (x_i, y_j), we first split $\mathcal{E}_{E,i,j} := \mathcal{E}_E[\mathbf{v}_{E,i,j}]$ as

$$\mathcal{E}_{E,i,j} = \frac{1}{2}(\mathcal{E}_{+,i,j} + \mathcal{E}_{-,i,j}) \,. \tag{3.44}$$

Note again that the upwind direction for $\mathcal{E}_-$ is opposite to the upwind direction for $\mathcal{E}_+$. In the case when $\xi > 0$ and $\eta > 0$, the proposed method approximates the partial derivatives on x by

$$\begin{aligned}(\partial_x\gamma\partial_y\gamma\mathcal{E}_E[\mathbf{v}_E])_{i,j} =& \frac{1}{\Delta x}\left(\gamma_{i,j}\left(\partial_y\gamma\mathcal{E}_+[\mathbf{v}_E]\right)_{i,j} - \gamma_{i-1,j}\left(\partial_y\gamma\mathcal{E}_+[\mathbf{v}_E]\right)_{i-1,j}\right) \\ & \frac{1}{\Delta x}\left(\gamma_{i+1,j}\left(\partial_y\gamma\mathcal{E}_-[\mathbf{v}_E]\right)_{i+1,j} - \gamma_{i,j}\left(\partial_y\gamma\mathcal{E}_-[\mathbf{v}_E]\right)_{i,j}\right) \,,\end{aligned} \tag{3.45}$$

and approximates the partial derivatives on y by

$$\left(\partial_y\gamma\mathcal{E}_+[\mathbf{v}_E]\right)_{i,j} = \frac{1}{\Delta y}\left(\gamma_{i,j}\mathcal{E}_{+,i,j} - \gamma_{i,j-1}\mathcal{E}_{+,i,j-1}\right) \,, \tag{3.46}$$

and

$$\left(\partial_y\gamma\mathcal{E}_-[\mathbf{v}_E]\right)_{i,j} = \frac{1}{\Delta y}\left(\gamma_{i,j+1}\mathcal{E}_{-,i,j+1} - \gamma_{i,j}\mathcal{E}_{-,i,j}\right) \,. \tag{3.47}$$

Thus, in such case, the proposed method approximates $\left(\xi\eta\partial_x\gamma\partial_y\gamma\mathcal{E}_E[\mathbf{v}_E]\right)_{i,j}$ by

$$\begin{aligned}\frac{\xi\eta}{2\Delta x\Delta y}(&\gamma_{i,j}(\gamma_{i,j}\mathcal{E}_{+,i,j} - \gamma_{i,j-1}\mathcal{E}_{+,i,j-1}) \\ &-\gamma_{i-1,j}(\gamma_{i-1,j}\mathcal{E}_{+,i-1,j} - \gamma_{i-1,j-1}\mathcal{E}_{+,i-1,j-1}) \\ &+\gamma_{i+1,j}(\gamma_{i+1,j+1}\mathcal{E}_{-,i+1,j+1} - \gamma_{i+1,j}\mathcal{E}_{-,i+1,j}) \\ &-\gamma_{i,j}(\gamma_{i,j+1}\mathcal{E}_{-,i,j+1} - \gamma_{i,j}\mathcal{E}_{-,i,j}))\,,\end{aligned} \tag{3.48}$$

and the approximations for the other three cases can be derived analogously.

In Section 3.2.2, we will show that the proposed method guarantees positivity preservation of the AP scheme (3.29)–(3.30).

3.2 Properties

In this section, we analyze properties of the proposed AP scheme (3.29)–(3.30) for the reduced moment equation (3.28), including stability and positive preservation.

The properties established in this section can be extended to the AP scheme (3.20)–(3.21) for the unreduced moment equation (3.4) with minor modification.

3.2.1 CFL Stability Condition

One important property for an AP scheme is that, when ϵ is away from zero, the scheme should take a hyperbolic Courant-Friedrichs-Lewy (CFL) stability condition, which, in the one-dimensional case, takes the form

$$\Delta t \leq C\epsilon\Delta x \,, \tag{3.49}$$

with some constant C. While (3.49) works well in the transport regime ($\epsilon = 1$), the time step size restriction soon becomes very restrictive while approaching the diffusion regime ($\epsilon \ll 1$). Such small size of time step makes the computational cost prohibitive. Hence, it is desirable to show that, when $\epsilon \to 0$, the CFL stability condition for the proposed scheme becomes a parabolic CFL condition, which, in the one-dimensional case, takes the form

$$\Delta t \leq C\Delta x^2 \,. \tag{3.50}$$

Also, (3.50) is the CFL stability condition that governs most of the explicit schemes for the diffusion equation (3.2).

In Proposition 1, we show that the proposed AP scheme (3.29)–(3.30) has the desirable CFL conditions.

Proposition 1. *In the case when ϵ is away from zero, the stability of proposed AP scheme (3.29)–(3.30) is governed by the hyperbolic CFL condition*

$$\Delta t \leq \frac{1}{2}\epsilon \left(\frac{\Delta x \Delta y}{\Delta x + \Delta y} \right) , \tag{3.51}$$

and, in the case when ϵ tends to zero, the stability of proposed AP scheme (3.29)–(3.30) is governed by the parabolic CFL condition

$$\Delta t \leq \frac{27}{16} \left(\frac{\Delta x \Delta y}{\Delta x + \Delta y} \right)^2 \sigma_{\min} , \tag{3.52}$$

where $\sigma_{\min} := \min_r \sigma_s(r)$.

Proof. From (3.29) and (3.30), the maximum wave speed of the equation is given by $\frac{1}{\epsilon}\gamma_{\max}^{3/2}$, with $\gamma_{\max} := \epsilon^2/(\epsilon^2 + \sigma_{\min}\Delta t)$. The CFL condition of the proposed scheme is then formulated as

$$\frac{1}{\epsilon}\gamma_{\max}^{3/2} \left(\frac{\Delta t}{\Delta x} + \frac{\Delta t}{\Delta y} \right) \leq \frac{1}{2} . \tag{3.53}$$

By using the definition of $\gamma_{\max}$, (3.53) can be rewritten as

$$\epsilon^2 \Delta t \leq \frac{1}{2} \left(\frac{\Delta x \Delta y}{\Delta x + \Delta y} \right) (\epsilon^2 + \sigma_{\min}\Delta t)^{3/2} . \tag{3.54}$$

Since $\sigma_{\min}\Delta t \geq 0$, it can be safely omitted from the right-hand-side of (3.54) when ϵ is away from zero. For any $\epsilon > 0$, (3.54) then becomes the hyperbolic CFL condition

$$\Delta t \leq \frac{1}{2}\epsilon \left(\frac{\Delta x \Delta y}{\Delta x + \Delta y} \right) . \tag{3.55}$$

For $\epsilon \to 0$, we need to take $\sigma_{\min}\Delta t$ back into consideration. We first let $\kappa = \frac{1}{2}\frac{\Delta x \Delta y}{\Delta x + \Delta y}$, and write (3.54) as

$$g(\epsilon) := \epsilon^{4/3}\Delta t^{2/3} - \kappa^{2/3}(\epsilon^2 + \sigma_{\min}\Delta t) \leq 0\,, \tag{3.56}$$

where g is defined as a function of ϵ. Note that the most strict CFL condition occurs when $\epsilon = \epsilon^* := \operatorname{argmax}_{\epsilon \geq 0} g(\epsilon)$. The value of ϵ^* can be computed by solving $g'(\epsilon) = 0$, which yields

$$\epsilon^* = \left(\frac{2}{3}\right)^{3/2} \frac{\Delta t}{\kappa}\,. \tag{3.57}$$

To confirm that ϵ^* is indeed a maximizer, one can easily verify that $g''(\epsilon^*) \leq 0$. When ϵ is chosen to be ϵ^*, (3.56) then becomes

$$g(\epsilon^*) = \left(\frac{2}{3}\right)^2 \frac{\Delta t^2}{\kappa^{4/3}} - \left(\frac{2}{3}\right)^3 \frac{\Delta t^2}{\kappa^{4/3}} + \kappa^{2/3}\sigma_{\min}\Delta t \leq 0\,. \tag{3.58}$$

After some algebraic work, it is not hard to show that (3.58) leads to

$$\Delta t \leq \frac{27}{4}\kappa^2\sigma_{\min} = \frac{27}{16}\left(\frac{\Delta x \Delta y}{\Delta x + \Delta y}\right)^2 \sigma_{\min}\,, \tag{3.59}$$

which takes the form of CFL conditions for diffusion schemes.

□

3.2.2 Positivity

In the following proposition, we show that the proposed AP scheme preserves positivity of the particle concentration.

Proposition 2. *Suppose $\mathcal{E}[\mathbf{v}_{i,j}^k] \geq 0$, i.e., $\mathcal{E}[\mathbf{v}_{i,j}^k]$ is a non-negative function on Ω, for all (x_i, y_j) on the spatial mesh. Assume that, at step k, for all (x_i, y_j) on the*

spatial mesh, there exists some positive constant C such that

$$\|\mathcal{E}[\mathbf{v}_{i,j}^k] - \left\langle \mathcal{E}[\mathbf{v}_{i,j}^k] \right\rangle / 4\pi\|_{L^\infty(\mathbb{S}^2)} \leq C\epsilon\mathcal{E}[\mathbf{v}_{i,j}^k] \,, \tag{3.60}$$

then the particle concentration $\rho_{i,j}^{k+1} := \sqrt{4\pi}\gamma_{i,j}v_{0,i,j}^{k+1}$, where $v_{0,i,j}^{k+1}$ is the zeroth moment of $\mathbf{v}_{i,j}^{k+1}$, is non-negative under the time-step restriction

$$\Delta t \leq \min\left\{ \frac{\Delta x^2 \Delta y^2}{\Delta x^2 + \Delta y^2}\sigma_{\min},\, C^{-1}\frac{\vartheta}{2}\frac{\sigma_{\min}^2}{\sigma_{\max}^2}(\Delta x + \Delta y) \right\} \,, \tag{3.61}$$

where $\sigma_{\min} := \min_r \sigma_s(r)$, $\sigma_{\max} := \max_r \sigma_s(r)$, and $\vartheta \in [1,\, 2]$ is the parameter in (3.37) and (3.38) used for in the minmod flux limiter.

Proof. By definition, $\rho_{i,j}^{k+1}$ has the same sign as $v_{0,i,j}^{k+1}$ Thus, to show that $\rho_{i,j}^{k+1}$ is non-negative, we only need to consider the update of $v_{0,i,j}^{k+1}$, i.e., the first equation in (3.30).

In the following proof, we use the spatial discretizations described in Sections 3.1.3.1 and 3.1.3.2 on the proposed scheme, and provide bounds for each term on the right-hand-side in the first equation of (3.30) after spatial discretization. Then we utilize the bounds to show non-negativity of $\rho_{i,j}^{k+1}$.

In this proof, we only consider the case that $\xi > 0$ and $\eta > 0$. Note that, for other cases, the following argument follows with minor modification on the upwind direction in the spatial discretization.

At point $(x_i,\, y_j)$, the discretized version of the first derivative term on x direction is written as

$$d_x := \frac{\xi}{2\Delta x}\left(\gamma_{i+1/2,j}^2(\mathcal{E}_{+,i+1/2,j}^k - \mathcal{E}_{-,i+1/2,j}^k) - \gamma_{i-1/2,j}^2(\mathcal{E}_{+,i-1/2,j}^k - \mathcal{E}_{-,i-1/2,j}^k)\right) . \tag{3.62}$$

Stripping the negative terms from (3.62) and applying (3.36) on the remaining terms yields

$$d_x \leq \frac{\xi}{2\Delta x}\left(\gamma^2_{i+1/2,j}(\mathcal{E}^k_{+,i,j} + s^x_{+,i,j}/2) + \gamma^2_{i-1/2,j}(\mathcal{E}^k_{-,i,j} - s^x_{-,i,j}/2)\right) . \tag{3.63}$$

Suppose $s^x_{+,i,j} > 0$ and $s^x_{-,i,j} < 0$, it then follows from (3.37) and (3.38) that

$$s^x_{+,i,j} \leq \vartheta(\mathcal{E}^k_{+,i,j} - \mathcal{E}^k_{+,i-1,j}) , \quad \text{and} \quad s^x_{-,i,j} \geq \vartheta(\mathcal{E}^k_{-,i+1,j} - \mathcal{E}^k_{-,i,j}) . \tag{3.64}$$

By applying both inequalities on (3.63), we have

$$\begin{aligned} d_x \leq & \frac{\xi}{2\Delta x}\frac{2+\vartheta}{2}\left(\gamma^2_{i+1/2,j}\mathcal{E}^k_{+,i,j} + \gamma^2_{i-1/2,j}\mathcal{E}^k_{-,i,j}\right) \\ & - \frac{\xi}{2\Delta x}\frac{\vartheta}{2}\left(\gamma^2_{i+1/2,j}\mathcal{E}^k_{+,i-1,j} + \gamma^2_{i-1/2,j}\mathcal{E}^k_{-,i+1,j}\right) . \end{aligned} \tag{3.65}$$

If $s^x_{+,i,j} \leq 0$ or $s^x_{-,i,j} \geq 0$, they can be directly taken out from (3.63) to form an upper bound on d_x, and it is not difficult to verify that this upper bound is tighter than the bound provided in (3.65). In other words, (3.65) holds for any $s^x_{+,i,j}$ and $s^x_{-,i,j}$.

Similarly, let us define

$$d_y := \frac{\eta}{2\Delta y}\left(\gamma^2_{i,j+1/2}(\mathcal{E}^k_{+,i,j+1/2} - \mathcal{E}^k_{-,i,j+1/2}) - \gamma^2_{i,j-1/2}(\mathcal{E}^k_{+,i,j-1/2} - \mathcal{E}^k_{-,i,j-1/2})\right) . \tag{3.66}$$

Then an upper bound of d_y can be obtained with the same strategy. The resulting upper bound is given by

$$\begin{aligned} d_y \leq & \frac{\eta}{2\Delta y}\frac{2+\vartheta}{2}\left(\gamma^2_{i,j+1/2}\mathcal{E}^k_{+,i,j} + \gamma^2_{i,j-1/2}\mathcal{E}^k_{-,i,j}\right) \\ & - \frac{\eta}{2\Delta y}\frac{\vartheta}{2}\left(\gamma^2_{i,j+1/2}\mathcal{E}^k_{+,i,j-1} + \gamma^2_{i,j-1/2}\mathcal{E}^k_{-,i,j+1}\right) . \end{aligned} \tag{3.67}$$

For second derivative terms, we define

$$\begin{aligned} d_{xx} := \frac{\xi^2}{4\Delta x^2}(\gamma_{i+1/2,j}(\gamma_{i+1,j}\mathcal{E}^k_{E,i+1,j} - \gamma_{i,j}\mathcal{E}^k_{E,i,j}) \\ -\gamma_{i-1/2,j}(\gamma_{i,j}\mathcal{E}^k_{E,i,j} - \gamma_{i-1,j}\mathcal{E}^k_{E,i-1,j})) , \end{aligned} \tag{3.68}$$

and

$$d_{yy} := \frac{\eta^2}{4\Delta y^2}(\gamma_{i,j+1/2}(\gamma_{i,j+1}\mathcal{E}_{E,i,j+1} - \gamma_{i,j}\mathcal{E}^k_{E,i,j}) \\ -\gamma_{i,j-1/2}(\gamma_{i,j}\mathcal{E}_{E,i,j} - \gamma_{i,j-1}\mathcal{E}_{E,i,j-1}))\,. \tag{3.69}$$

Then taking all the non-negative terms out yields

$$d_{xx} \geq -\frac{\xi^2}{4\Delta x^2}\left(\gamma_{i+1/2,j} + \gamma_{i-1/2,j}\right)\gamma_{i,j}\mathcal{E}_{E,i,j}\,, \tag{3.70}$$

and

$$d_{yy} \geq -\frac{\eta^2}{4\Delta y^2}\left(\gamma_{i,j+1/2} + \gamma_{i,j-1/2}\right)\gamma_{i,j}\mathcal{E}_{E,i,j}\,. \tag{3.71}$$

Finally, we define the mixed partial derivative terms as

$$d_{xy} := \frac{\xi\eta}{2\Delta x\Delta y}(\gamma_{i,j}(\gamma_{i,j}\mathcal{E}^k_{+,i,j} - \gamma_{i,j-1}\mathcal{E}^k_{+,i,j-1}) \\ -\gamma_{i-1,j}(\gamma_{i-1,j}\mathcal{E}^k_{+,i-1,j} - \gamma_{i-1,j-1}\mathcal{E}^k_{+,i-1,j-1}) \\ +\gamma_{i+1,j}(\gamma_{i+1,j+1}\mathcal{E}^k_{-,i+1,j+1} - \gamma_{i+1,j}\mathcal{E}^k_{-,i+1,j}) \\ -\gamma_{i,j}(\gamma_{i,j+1}\mathcal{E}^k_{-,i,j+1} - \gamma_{i,j}\mathcal{E}^k_{-,i,j}))\,, \tag{3.72}$$

and

$$d_{yx} := \frac{\eta\xi}{2\Delta y\Delta x}(\gamma_{i,j}(\gamma_{i,j}\mathcal{E}^k_{+,i,j} - \gamma_{i-1,j}\mathcal{E}^k_{+,i-1,j}) \\ -\gamma_{i,j-1}(\gamma_{i,j-1}\mathcal{E}^k_{+,i,j-1} - \gamma_{i-1,j-1}\mathcal{E}^k_{+,i-1,j-1}) \\ +\gamma_{i,j+1}(\gamma_{i+1,j+1}\mathcal{E}^k_{-,i+1,j+1} - \gamma_{i,j+1}\mathcal{E}^k_{-,i,j+1}) \\ -\gamma_{i,j}(\gamma_{i+1,j}\mathcal{E}^k_{-,i+1,j} - \gamma_{i,j}\mathcal{E}^k_{-,i,j}))\,. \tag{3.73}$$

Then the lower bounds on d_{xy} and d_{yz} can be obtained by stripping all non-negative terms from (3.72) and (3.73). The bounds are given by

$$d_{xy} \geq \frac{-\xi\eta}{2\Delta x\Delta y}(\gamma_{i,j}(\gamma_{i,j-1}\mathcal{E}^k_{+,i,j-1} + \gamma_{i,j+1}\mathcal{E}^k_{-,i,j+1}) \\ +\gamma^2_{i-1,j}\mathcal{E}^k_{+,i-1,j} + \gamma^2_{i+1,j}\mathcal{E}^k_{-,i+1,j})\,, \tag{3.74}$$

and

$$d_{yx} \geq \frac{-\eta\xi}{2\Delta y\Delta x}(\gamma_{i,j}(\gamma_{i-1,j}\mathcal{E}^k_{+,i-1,j} + \gamma_{i+1,j}\mathcal{E}^k_{-,i+1,j}) \\ +\gamma^2_{i,j-1}\mathcal{E}^k_{+,i,j-1} + \gamma^2_{i,j+1}\mathcal{E}^k_{-,i,j+1}) \,. \tag{3.75}$$

Finally, we obtained all necessary bounds on the derivative terms in the first equation of (3.30). With proper scaling, the first equation of (3.30) can be written as

$$\rho^{k+1}_{i,j} := \frac{\gamma_{i,j}\, v^{k+1}_{0,i,j}}{m_0} = \frac{\gamma_{i,j} v^k_{0,i,j}}{m_0} - \frac{\Delta t}{\epsilon}\langle d_x + d_y\rangle + \frac{\Delta t^2}{\epsilon^2}\langle d_{xx} + d_{yy} + d_{xy} + d_{yx}\rangle \,. \tag{3.76}$$

Let us split the bounds provided in (3.65) and (3.67) into positive and negative parts and define each part as

$$\begin{aligned} d_{x,+} &:= \frac{\xi}{2\Delta x}\frac{2+\vartheta}{2}\left(\gamma^2_{i+1/2,j}\mathcal{E}^k_{+,i,j} + \gamma^2_{i-1/2,j}\mathcal{E}^k_{-,i,j}\right) \,, \\ d_{x,-} &:= \frac{\xi}{2\Delta x}\frac{\vartheta}{2}\left(\gamma^2_{i+1/2,j}\mathcal{E}^k_{+,i-1,j} + \gamma^2_{i-1/2,j}\mathcal{E}^k_{-,i+1,j}\right) \,, \\ d_{y,+} &:= \frac{\eta}{2\Delta y}\frac{2+\vartheta}{2}\left(\gamma^2_{i,j+1/2}\mathcal{E}^k_{+,i,j} + \gamma^2_{i,j-1/2}\mathcal{E}^k_{-,i,j}\right) \,, \\ d_{y,-} &:= \frac{\eta}{2\Delta y}\frac{\vartheta}{2}\left(\gamma^2_{i,j+1/2}\mathcal{E}^k_{+,i,j-1} + \gamma^2_{i,j-1/2}\mathcal{E}^k_{-,i,j+1}\right) \,. \end{aligned} \tag{3.77}$$

Then, (3.76) can be separated into two parts as follows:

$$\rho^{k+1}_{i,j} \geq \left(\frac{\gamma_{i,j} v^k_{0,i,j}}{m_0} - \frac{\Delta t}{\epsilon}\langle d_{x,+} + d_{y,+}\rangle + \frac{\Delta t^2}{\epsilon^2}\langle d_{xx} + d_{yy}\rangle\right) \\ + \left(\frac{\Delta t}{\epsilon}\langle d_{x,-} + d_{y,-}\rangle + \frac{\Delta t^2}{\epsilon^2}\langle d_{xy} + d_{yx}\rangle\right) \,. \tag{3.78}$$

We prove $\rho^{k+1}_{i,j} \geq 0$ by showing that the two parts in (3.78) are both non-negative.

Since $v^k_{0,i,j} := \langle m_0 \mathcal{E}[\mathbf{v}^k_{i,j}]\rangle$, the first term on the right-hand-side of (3.78) can be written as

$$\frac{\gamma_{i,j} v^k_{0,i,j}}{m_0} = \frac{\gamma_{i,j}\langle m_0\mathcal{E}[\mathbf{v}^k_{i,j}]\rangle}{m_0} = \langle \gamma_{i,j}\mathcal{E}[\mathbf{v}^k_{i,j}]\rangle \,. \tag{3.79}$$

Thus, for the first part in (3.78), it suffices to show that

$$\gamma_{i,j}\mathcal{E}[\mathbf{v}_{i,j}^k] - \frac{\Delta t}{\epsilon}\left(d_{x,+} + d_{y,+}\right) + \frac{\Delta t^2}{\epsilon^2}\left(d_{xx} + d_{yy}\right) \geq 0\,. \tag{3.80}$$

Using the bounds (3.77), (3.70), (3.71), it can be verified after some algebraic work that (3.80) holds under condition

$$\Delta t \leq \frac{\Delta x^2 \Delta y^2}{\Delta x^2 + \Delta y^2}\sigma_{\min}\,. \tag{3.81}$$

On the other hand, by using bounds (3.77), (3.74), (3.75), together with the assumption (3.60) on the second part in (3.78), we observe that the second part in (3.78) is non-negative if

$$\frac{\Delta t}{\epsilon}\left(\frac{\xi}{2\Delta x} + \frac{\eta}{2\Delta y}\right)\frac{\vartheta}{2}\sigma_{\min}^2 - \frac{\Delta t^2}{\epsilon^2}\frac{\xi\eta}{\Delta x \Delta y}\sigma_{\max}^2 C\epsilon \geq 0\,. \tag{3.82}$$

Thus, the second part in (3.78) is non-negative given the condition

$$\Delta t \leq C^{-1}\frac{\vartheta}{2}\frac{\sigma_{\min}^2}{\sigma_{\max}^2}(\Delta x + \Delta y)\,. \tag{3.83}$$

Combining (3.81) and (3.83) completes the proof for case $\xi > 0$ and $\eta > 0$. The proof for other cases follows analogously with the modified upwinding direction with respect to the signs of ξ and η. □

Proposition 2 shows that the proposed AP scheme achieves the positive preserving property, under a mild condition (3.60). From (3.30), it can be observed that all even moments, except the zeroth one, are decaying at rate $O(\epsilon^2)$ for small ϵ. Thus, even if (3.60) is violated in the initial step, it will eventually be satisfied for any positive constant C as time evolves. In addition, the following remark provides a time step restriction that guarantees positive preserving even when (3.60) does not hold.

Remark 6. *If condition* (3.60) *does not hold, the proposed AP scheme still preserves positivity of the particle concentration, but under a more strict time step restriction*

$$\Delta t \leq \epsilon \frac{\vartheta}{2} \frac{\sigma_{\min}^2}{\sigma_{\max}^2} (\Delta x + \Delta y) \,, \tag{3.84}$$

where $\sigma_{\min}$, $\sigma_{\max}$, *and* ϑ *are defined in Proposition 2.*

Condition (3.84) can be easily obtained following the proof for Proposition 2, with minor modification on (3.82). Without assumption (3.60), (3.82) is rewritten as

$$\frac{\Delta t}{\epsilon} \left(\frac{\xi}{2\Delta x} + \frac{\eta}{2\Delta y} \right) \frac{\vartheta}{2} \sigma_{\min}^2 - \frac{\Delta t^2}{\epsilon^2} \frac{\xi \eta}{\Delta x \Delta y} \sigma_{\max}^2 \geq 0 \,. \tag{3.85}$$

Thus, condition (3.84) follows.

3.3 Numerical Results

In this section, we compare the proposed AP scheme with the second-order (non-AP) kinetic scheme introduced in [10, 15]. The numerical schemes are tested with the FP_N^+ method on the line source benchmark problem described in Section 2.4.1. The initial condition is given by a steep Gaussian distribution with variance 9×10^{-4}, and an artificial zero boundary condition is imposed. The calculations are performed on a $[-1.5, 1.5] \times [-1.5, 1.5]$ square domain centered at the origin, with a 100×100 spatial mesh.

The parabolic time step restriction (3.61) is used to guarantee positivity on the AP scheme. In the case that the particle concentration becomes negative (due to violation of (3.60)), we take one step back in time and apply the more restrictive

hyperbolic condition (3.84) to enforce positivity. In our numerical tests, we never encountered such a situation.

For the second-order non-AP scheme, the time step is given by

$$\Delta t \leq \frac{2}{2+\vartheta}\frac{\Delta x \Delta y}{\Delta x \Delta y}\epsilon\,, \tag{3.86}$$

where $\vartheta = 2$ is the parameter for the minmod-type limiter, which is also used to enforce positivity in the kinetic scheme. See (A.7) and (A.8) in Appendix A.1.

For both schemes, we perform integration in time using Heun's method, which is the optimal[2], second-order strong-stability-preserving Runge-Kutta (SSP-RK2) method [73]. The detailed formulation for the Heun's method can be found in Appendix A.2.

The spectral filter in the FP_N^+ method is implemented with discrete embedding for simplicity (see Section 2.1.3 for details). Algorithm CR-MPC (presented in Chapter 4) is used to solve the optimization problems in the FP_N^+ method. The algorithm parameter values used in the test are $\varepsilon = 10^{-4}$, $\tau = 0.5$, $\zeta = 0.9$, $\varkappa = 0.98$, $\underline{\lambda} = 10^{-6}$, and $\lambda^{\max} = 10^{30}$.

In Figures 3.1 – 3.10, we plot the particle concentration $\rho := \frac{\langle f \rangle}{4\pi}$ for the solutions generated by the FP_N^+ method with the non-AP and AP schemes. The moment order $N = 7$ is used in the test, and the degree of precision of the quadrature rules are $N_{\mathcal{Q}} = 15$ and $N_{\mathcal{Q}} = 16$, the minimum required precision for the non-AP and AP scheme, respectively. Figures 3.1, 3.2, 3.5, 3.6, and 3.9 show the heat maps over the entire two-dimensional domain and Figures 3.3, 3.4, 3.7, 3.8, and 3.10 present

[2]It is optimal in the sense that it allows for the largest possible time step in the SSP-RK2 class.

the one-dimensional line-outs along the x-axis. For comparison, the exact transport solution (green solid line) is included in Figures 3.3 and 3.4. For Figures 3.7, 3.8, and 3.10, a reference diffusion solution (green solid line) is included. Such reference solution is generated by solving the diffusion equation (3.2) with the corresponding initial condition.

From Figures 3.1 – 3.4, we observe that, in the transport regime ($\epsilon = 1$), the solution generated by the proposed AP scheme is comparable to the solution generated by the non-AP scheme. When ϵ decreases to 0.1, both schemes produce reasonably good approximations to the reference diffusion solution, as shown in Figures 3.5 – 3.8. If ϵ is further decreased to 0.001, the AP scheme gives pretty accurate approximation to the diffusion solution, as illustrated in Figures 3.9 and 3.10. However, the exceedingly strict time step size restriction (3.86) prevents the non-AP scheme from solving this problem in a reasonable amount of time.

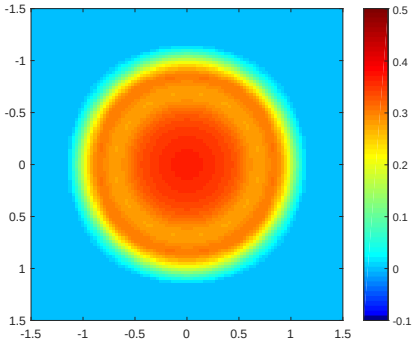


Figure 3.1: Line source solution ($\epsilon = 1$) – heat map of particle concentration ρ at $t = 1$ using the non-AP solver.

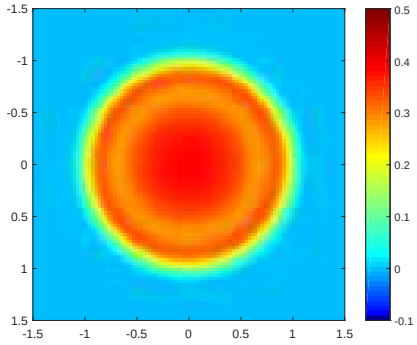


Figure 3.2: Line source solution ($\epsilon = 1$) – heat map of particle concentration ρ at $t = 1$ using the AP solver.

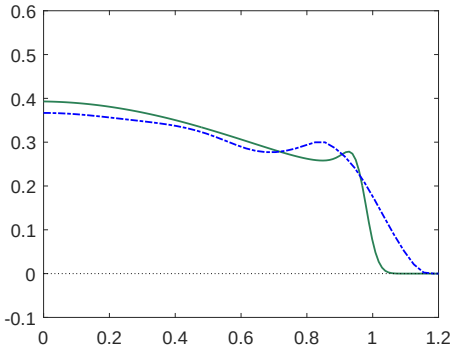


Figure 3.3: Line source solution ($\epsilon = 1$) – line-outs of particle concentration ρ along the x-axis at $t = 1$ using the non-AP solver.

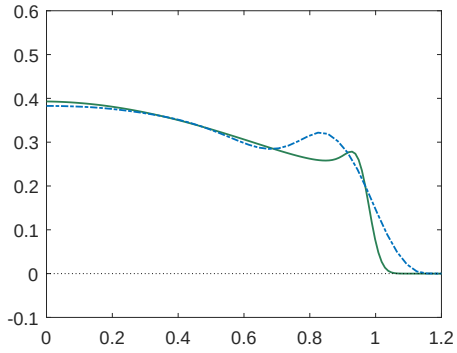


Figure 3.4: Line source solution ($\epsilon = 1$) – line-outs of particle concentration ρ along the x-axis at $t = 1$ using the AP solver.

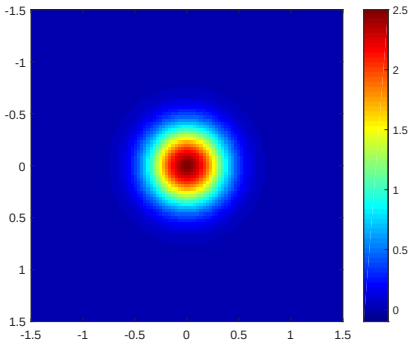


Figure 3.5: Line source solution ($\epsilon = 0.1$) – heat map of particle concentration ρ at $t = 0.1$ using the non-AP solver.

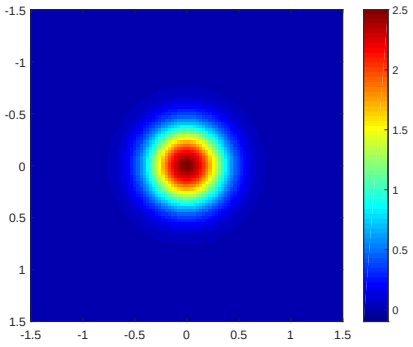


Figure 3.6: Line source solution ($\epsilon = 0.1$) – heat map of particle concentration ρ at $t = 0.1$ using the AP solver.

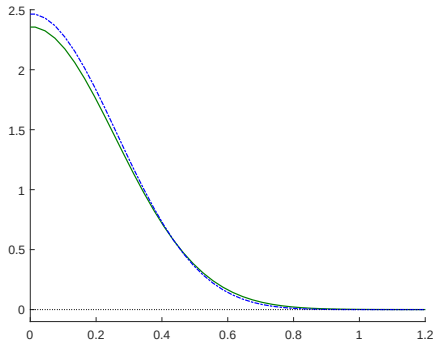


Figure 3.7: Line source solution ($\epsilon = 0.1$) – line-outs of particle concentration ρ along the x-axis at $t = 0.1$ using the non-AP solver.

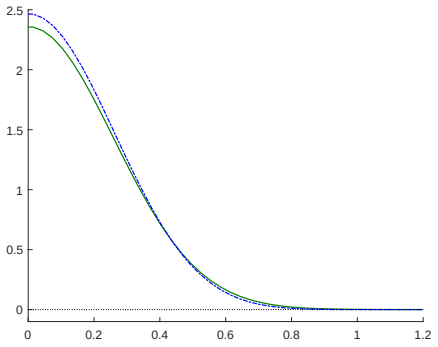


Figure 3.8: Line source solution ($\epsilon = 0.1$) – line-outs of particle concentration ρ along the x-axis at $t = 0.1$ using the AP solver.

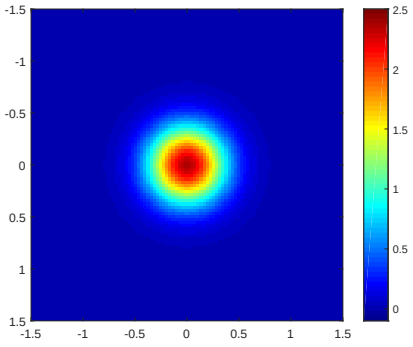


Figure 3.9: Line source solution ($\epsilon = 0.001$) – heat map of particle concentration ρ at $t = 0.1$ using the AP solver.

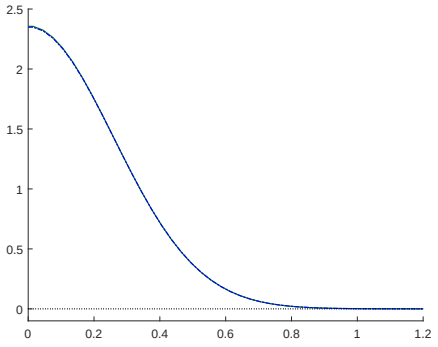


Figure 3.10: Line source solution ($\epsilon = 0.001$) – heat map of particle concentration ρ at $t = 0.1$ using the AP solver.

Chapter 4: A Mehrotra-Predictor-Corrector (MPC) Algorithm for Convex Quadratic Programming – a Constraint-Reduced Variant

In this chapter, we propose Algorithm CR-MPC. The algorithm was initially developed to solve convex quadratic programming problems (CQPs) of the form (2.23) in the FP_N^+ method. (See Section 2.2 for details.) However, Algorithm CR-MPC is actually an efficient solver for general CQPs with a large number of inequality constraints. Hence, in this chapter, we consider the CQP in standard inequality form

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\ & f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^T H\mathbf{x} + \mathbf{c}^T\mathbf{x} \\ \text{subject to}\ & A\mathbf{x} \geq \mathbf{b}\,, \end{aligned} \tag{P}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the optimization variable, $f : \mathbb{R}^n \to \mathbb{R}$ is the objection function with $\mathbf{c} \in \mathbb{R}^n$ and a symmetric and positive semidefinite Hessian matrix $H \in \mathbb{R}^{n\times n}$, and $A \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$ define the m linear inequality constraints in this problem.

The dual problem associated to (P) is given as

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n,\,\boldsymbol{\lambda}\in\mathbb{R}^m}{\text{maximize}}\ & -\frac{1}{2}\mathbf{x}^T H\mathbf{x} + \mathbf{b}^T\boldsymbol{\lambda} \\ \text{subject to}\ & H\mathbf{x} + \mathbf{c} - A^T\boldsymbol{\lambda} = 0 \\ & \boldsymbol{\lambda} \geq 0\,, \end{aligned} \tag{D}$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. Since the objective function f is convex and the constraints are linear, solving (P)–(D) is equivalent to solving the Karush-Kuhn-Tucker (KKT) system

$$\begin{aligned} H\mathbf{x} - A^T\boldsymbol{\lambda} + \mathbf{c} &= 0, \\ A\mathbf{x} - \mathbf{b} - \mathbf{s} &= 0, \\ S\boldsymbol{\lambda} &= 0, \\ \mathbf{s}, \boldsymbol{\lambda} &\geq 0, \end{aligned} \tag{4.1}$$

where $\mathbf{s}$ is a vector of slack variables associated to the inequality constraints in the primal problem, and $S = \text{diag}(\mathbf{s})$.

Primal-dual interior-point methods (PDIPMs) are commonly used for the solution of CQPs by iteratively solving the KKT system (4.1). Mehrotra's predictor-corrector (MPC) algorithm [44] and its variants [36] compute the search direction for solving the KKT system by constructing two directions – the predictor direction and the corrector/centering direction. With the help of the corrector direction, the MPC algorithm usually generates better search directions compared to other interior-point methods, and has proved to be significantly more efficient in practice.

In Section 4.1, we introduce a modified MPC algorithm for solving CQPs, and propose a constraint-reduced variant, Algorithm CR-MPC, which is an extension of the MPC algorithm analyzed in [37] for solving linear programming problems.[1] Further, we provide conditions on the constraint selection rules used in Algorithm CR-MPC, and prove that, under these conditions and appropriate assumptions, Algorithm CR-MPC converges globally to the solution at a locally q-quadratic

[1]The affine-scaling version analyzed in [37] was similarly extended to CQP in [40].

rate. In Section 4.2, we propose a new rule for selecting the constraints, and show that the proposed rule satisfies the required conditions for the convergence properties of Algorithm CR-MPC. Numerical results for experiments on randomly generated problems and the CQPs from the FP_N^+ closure are both reported in Section 4.3.

4.1 MPC Algorithm – A Constraint-Reduced Variant

4.1.1 Definitions

Here we first define some of the important sets which are frequently used in the remainder of this chapter.

Definition 1. *The primal feasible set* $\mathcal{F}_P := \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} \geq \mathbf{b}\}$.

Definition 2. *The primal strictly feasible set* $\mathcal{F}_P^o := \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} > \mathbf{b}\}$.

Definition 3. *The primal solution set* $\mathcal{F}_P^* := \{\mathbf{x}^* \in \mathcal{F}_P : f(\mathbf{x}^*) \leq f(\mathbf{x}), \forall \mathbf{x} \in \mathcal{F}_P\}$.

Definition 4. *The index set* $\mathcal{I} := \{1, \ldots, m\}$, *and for* $\mathbf{x} \in \mathcal{F}_P$, *the active constraint set* $\mathcal{A}(\mathbf{x}) := \{i \in \mathcal{I} : \mathbf{a}_i^T \mathbf{x} = b_i\}$.

4.1.2 A Modified MPC Algorithm

Given $(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$ with $\boldsymbol{\lambda} > \mathbf{0}$, $\mathbf{s} > \mathbf{0}$, the KKT system (4.1) can be solved by computing the *primal-dual affine-scaling* direction $(\Delta\mathbf{x}^{\text{a}}, \Delta\boldsymbol{\lambda}^{\text{a}}, \Delta\mathbf{s}^{\text{a}})$ at $(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$, which is given by the solution of the Newton linear system associated with the first

three lines of (4.1), i.e.,

$$\begin{bmatrix} H & -A^T & 0 \\ A & 0 & -I \\ 0 & S & \Lambda \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}^{\mathrm{a}} \\ \Delta\boldsymbol{\lambda}^{\mathrm{a}} \\ \Delta\mathbf{s}^{\mathrm{a}} \end{bmatrix} = \begin{bmatrix} -(H\mathbf{x}+\mathbf{c}) + A^T\boldsymbol{\lambda} \\ \mathbf{0} \\ -S\boldsymbol{\lambda} \end{bmatrix}, \tag{4.2}$$

where $\Lambda = \mathrm{diag}(\boldsymbol{\lambda})$. Using two steps of block Gaussian elimination, it can be seen that the affine-scaling direction equivalently satisfies the normal equation system

$$\begin{aligned} M\Delta\mathbf{x}^{\mathrm{a}} &= -(H\mathbf{x}+\mathbf{c}), \\ \Delta\mathbf{s}^{\mathrm{a}} &= A\Delta\mathbf{x}^{\mathrm{a}}, \\ \Delta\boldsymbol{\lambda}^{\mathrm{a}} &= -\boldsymbol{\lambda} - S^{-1}\Lambda\Delta\mathbf{s}^{\mathrm{a}}\,, \end{aligned} \tag{4.3}$$

where the "normal" matrix M is given by

$$M := H + A^T S^{-1}\Lambda A = H + \sum_{i=1}^{m} \frac{\lambda_i}{s_i}\mathbf{a}_i\mathbf{a}_i^T\,, \tag{4.4}$$

with $\mathbf{a}_i$ the transpose of the i-th row of A.

For CQPs with a larger number of inequality constraints than the number of variables, solving the normal system (4.3) is generally preferable to solving the original linear system (4.2), since M is of a smaller order n than the original matrix which is of order $n + 2m$. However, a major numerical difficulty for solving (4.3) is that the cost of *forming* M is high; see Section 4.1.3 for details.

MPC algorithms improve on the affine-scaling search direction by introducing a *centering/corrector* direction [36, 44]. Similar to the direction used in [37] in the linear-programming case, the centering/corrector direction ($\Delta\mathbf{x}^{\mathrm{c}}$, $\Delta\boldsymbol{\lambda}^{\mathrm{c}}$, $\Delta\mathbf{s}^{\mathrm{c}}$) in our

extended MPC algorithm is obtained by solving

$$\begin{bmatrix} H & -A^T & 0 \\ A & 0 & -I \\ 0 & S & \Lambda \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}^{\mathrm{c}} \\ \Delta\boldsymbol{\lambda}^{\mathrm{c}} \\ \Delta\mathbf{s}^{\mathrm{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \sigma\mu\mathbf{1} - \Delta S^{\mathrm{a}}\Delta\boldsymbol{\lambda}^{\mathrm{a}} \end{bmatrix}, \tag{4.5}$$

where $\mathbf{1}$ is a vector of all ones, the centering parameter $\sigma = (1 - \alpha^{\mathrm{a}})^3$, and duality measure $\mu = \mathbf{s}^T\boldsymbol{\lambda}/m$. The affine-scaling step size α^{a} is given below in (4.16). Following [37], our MPC search direction at $(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$ is given by

$$(\Delta\mathbf{x}, \Delta\boldsymbol{\lambda}, \Delta\mathbf{s}) = (\Delta\mathbf{x}^{\mathrm{a}}, \Delta\boldsymbol{\lambda}^{\mathrm{a}}, \Delta\mathbf{s}^{\mathrm{a}}) + \gamma(\Delta\mathbf{x}^{\mathrm{c}}, \Delta\boldsymbol{\lambda}^{\mathrm{c}}, \Delta\mathbf{s}^{\mathrm{c}}), \tag{4.6}$$

where the "mixing" parameter $\gamma \in [0, 1]$ is chosen to guarantee that the search direction is a descent direction and the magnitude of the centering/corrector direction is not too large compared to the magnitude of the affine-scaling direction. Specifically,

$$\gamma := \min\left\{\gamma_1, \tau\frac{\|\Delta\mathbf{x}^{\mathrm{a}}\|}{\|\Delta\mathbf{x}^{\mathrm{c}}\|}, \tau\frac{\|\Delta\mathbf{x}^{\mathrm{a}}\|}{\sigma\mu}\right\}, \tag{4.7}$$

where $\tau \in [0, 1)$ is an algorithm parameter for controlling the magnitude of the centering/corrector component, and

$$\gamma_1 := \operatorname{argmax}\left\{\tilde{\gamma} \in [0, 1] \,|\, f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^{\mathrm{a}} + \tilde{\gamma}\Delta\mathbf{x}^{\mathrm{c}}) \geq \zeta(f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^{\mathrm{a}}))\right\}, \tag{4.8}$$

where $\zeta \in (0, 1)$ is an algorithm parameter.

4.1.3 A Constraint-Reduced MPC Algorithm

In the modified MPC algorithm described in Section 4.1.2, the main computational cost is incurred in forming the normal matrix M ((4.4)), which requires

approximately $mn^2/2$ multiplications if A is dense. The cost may become prohibitive when solving CQPs with a large number of inequality constraints. The constraint reduction mechanism in [40] modifies M by limiting the sum in (4.4) to a wisely selected small subset of all constraints, referred as the "working set." In this chapter, we denote the working set by Q, and the rules for selecting Q are presented later in this section.

Given an index set $Q \subseteq \{1, \ldots, m\}$ of constraints, the constraint reduction technique produces an *approximate* affine-scaling direction by solving a "reduced" version of the Newton system (4.2), which is given by

$$\begin{bmatrix} H & -A_Q^T & 0 \\ A_Q & 0 & -I \\ 0 & S_Q & \Lambda_Q \end{bmatrix} \begin{bmatrix} \Delta \mathbf{x}^{\mathrm{a}} \\ \Delta \boldsymbol{\lambda}_Q^{\mathrm{a}} \\ \Delta \mathbf{s}_Q^{\mathrm{a}} \end{bmatrix} = \begin{bmatrix} -(H\mathbf{x} + \mathbf{c}) + A_Q^T \boldsymbol{\lambda}_Q \\ \mathbf{0} \\ -S_Q \boldsymbol{\lambda}_Q \end{bmatrix}, \tag{4.9}$$

where A_Q is a sub-matrix of A consisting of those rows with index in Q, $\mathbf{s}_Q$ and $\boldsymbol{\lambda}_Q$ are sub-vectors of $\mathbf{s}$ and $\boldsymbol{\lambda}$ defined accordingly, $S_Q = \mathrm{diag}\,(\mathbf{s}_Q)$, and $\Lambda_Q = \mathrm{diag}\,(\boldsymbol{\lambda}_Q)$. Similar to (4.2), the reduced system (4.9) can be solved via solving the reduced normal system

$$\begin{aligned} M_{(Q)} \Delta \mathbf{x}^{\mathrm{a}} &= -(H\mathbf{x} + \mathbf{c}) \,, \\ \Delta \mathbf{s}_Q^{\mathrm{a}} &= A_Q \Delta \mathbf{x}^{\mathrm{a}} \,, \\ \Delta \boldsymbol{\lambda}_Q^{\mathrm{a}} &= -\boldsymbol{\lambda}_Q - S_Q^{-1} \Lambda_Q \Delta \mathbf{s}_Q^{\mathrm{a}} \,, \end{aligned} \tag{4.10}$$

where $M_{(Q)}$ is defined as

$$M_{(Q)} := H + A_Q^T S_Q^{-1} \Lambda_Q A_Q = H + \sum_{i \in Q} \frac{\lambda_i}{s_i} \mathbf{a}_i \mathbf{a}_i^T \,. \tag{4.11}$$

The following lemma from [40] (see Lemma 2.1 of [40] for proofs) gives conditions

that guarantee the nonsingularity of $M_{(Q)}$, which is necessary for a successful iteration.

Lemma 4 (Corresponds to Lemma 2.1 of [40]). *Let* $\boldsymbol{\lambda} > \mathbf{0}$, $\mathbf{s} > \mathbf{0}$, *and* $Q \subseteq \mathcal{I}$. *Then* $M_{(Q)}$ *is positive definite if and only if* $\operatorname{rank}([H\ A_Q^T]) = n$.

From (4.11), the cost for forming $M_{(Q)}$ is $\mathcal{O}(qn^2)$, where q equals the size of Q. As the working set generally is of size $q \ll m$, forming $M_{(Q)}$ is much less expensive than forming M in (4.4), which costs $\mathcal{O}(mn^2)$ operations. Matrix $M_{(Q)}$ is also used in computing a modified centering/corrector direction, by solving

$$\begin{bmatrix} H & -A_Q^T & 0 \\ A_Q & 0 & -I \\ 0 & S_Q & \Lambda_Q \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}^{\mathrm{c}} \\ \Delta\boldsymbol{\lambda}_Q^{\mathrm{c}} \\ \Delta\mathbf{s}_Q^{\mathrm{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \sigma\mu_{(Q)}\mathbf{1} - \Delta S_Q^{\mathrm{a}}\Delta\boldsymbol{\lambda}_Q^{\mathrm{a}} \end{bmatrix}, \tag{4.12}$$

where $\mu_{(Q)} := \mathbf{s}_Q^T\boldsymbol{\lambda}_Q/q$, and the corresponding normal equation system is given by

$$\begin{aligned} M_{(Q)}\Delta\mathbf{x}^{\mathrm{c}} &= A_Q^T S_Q^{-1}(\sigma\mu_{(Q)}\mathbf{1} - \Delta S_Q^{\mathrm{a}}\Delta\boldsymbol{\lambda}_Q^{\mathrm{a}}), \\ \Delta\mathbf{s}_Q^{\mathrm{c}} &= A_Q\Delta\mathbf{x}^{\mathrm{c}}, \\ \Delta\boldsymbol{\lambda}_Q^{\mathrm{c}} &= S_Q^{-1}(-\Lambda_Q\Delta\mathbf{s}_Q^{\mathrm{c}} + \sigma\mu_{(Q)}\mathbf{1} - \Delta S_Q^{\mathrm{a}}\Delta\boldsymbol{\lambda}_Q^{\mathrm{a}})\,. \end{aligned} \tag{4.13}$$

The search direction for the constraint-reduced MPC algorithm at $(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$ is then given by

$$(\Delta\mathbf{x}, \Delta\boldsymbol{\lambda}_Q, \Delta\mathbf{s}_Q) = (\Delta\mathbf{x}^{\mathrm{a}}, \Delta\boldsymbol{\lambda}_Q^{\mathrm{a}}, \Delta\mathbf{s}_Q^{\mathrm{a}}) + \gamma(\Delta\mathbf{x}^{\mathrm{c}}, \Delta\boldsymbol{\lambda}_Q^{\mathrm{c}}, \Delta\mathbf{s}_Q^{\mathrm{c}}), \tag{4.14}$$

where γ is given by (4.7), with μ replaced by $\mu_{(Q)}$.

The constraint-reduced MPC algorithm is stated in Algorithm CR-MPC below,[2] and a discussion on the guidelines for selecting the working set Q will follow.

[2]The "modified MPC algorithm" described in Section 4.1.2 is obtained as a special case by

ALGORITHM CR-MPC: A Constraint-Reduced variant of MPC Algorithm for CQP

Parameters: $\varepsilon \geq 0$, $\tau \in [0,1)$, $\zeta \in (0,1)$, $\varkappa \in (0,1)$, $\underline{\lambda} \in (0,\infty]$ and $\lambda^{\max} > 0$.

Data: Strictly feasible starting point $(\mathbf{x}, \boldsymbol{\lambda})$ for (P)–(D) with $\mathbf{s} := A\mathbf{x} - \mathbf{b} > \mathbf{0}$ and $\boldsymbol{\lambda} > \mathbf{0}$. Initialize $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}$.

Step 1. Set $\tilde{\mu} := \mathbf{s}^T\tilde{\boldsymbol{\lambda}}/m$. Terminate if

$$E := \max\left\{\frac{\|H\mathbf{x} + \mathbf{c} - A^T\tilde{\boldsymbol{\lambda}}\|_\infty}{\max\{\|A\|_\infty,\, \|H\|_\infty,\, \|\mathbf{c}\|_\infty\}},\quad \tilde{\mu}\right\} \leq \varepsilon. \tag{4.15}$$

Step 2. Select a working set Q. Define $q = |Q|$.

Step 3. Compute approximate normal matrix $M_{(Q)} = H + \sum_{i\in Q} \frac{\lambda_i}{s_i}\mathbf{a}_i\mathbf{a}_i^T$.

Step 4. Solve (4.9) for the affine-scaling direction $(\Delta\mathbf{x}^{\mathrm{a}}, \Delta\boldsymbol{\lambda}_Q^{\mathrm{a}}, \Delta\mathbf{s}_Q^{\mathrm{a}})$, and set $\Delta\mathbf{s}^{\mathrm{a}} := A\Delta\mathbf{x}^{\mathrm{a}}$.

Step 5. Compute the affine-scaling step

$$\alpha^{\mathrm{a}} := \operatorname{argmax}\{\alpha \in [0,\, 1] \,|\, \mathbf{s} + \alpha\Delta\mathbf{s}^{\mathrm{a}} \geq 0,\ \boldsymbol{\lambda}_Q + \alpha\Delta\boldsymbol{\lambda}_Q^{\mathrm{a}} \geq 0\}\,. \tag{4.16}$$

Step 6. Set $\mu_{(Q)} := \mathbf{s}_Q^T\boldsymbol{\lambda}_Q/q$, compute centering parameter $\sigma = (1-\alpha^{\mathrm{a}})^3$.

Step 7. Solve (4.12) for the corrector direction $(\Delta\mathbf{x}^{\mathrm{c}}, \Delta\boldsymbol{\lambda}_Q^{\mathrm{c}}, \Delta\mathbf{s}_Q^{\mathrm{c}})$, and set $\Delta\mathbf{s}^{\mathrm{c}} := A\Delta\mathbf{x}^{\mathrm{c}}$.

Step 8. Compute the mixing parameter γ as in (4.7), with μ replaced by $\mu_{(Q)}$. Compute the search direction

$$(\Delta\mathbf{x},\, \Delta\boldsymbol{\lambda}_Q,\, \Delta\mathbf{s}) := (\Delta\mathbf{x}^{\mathrm{a}},\, \Delta\boldsymbol{\lambda}_Q^{\mathrm{a}},\, \Delta\mathbf{s}^{\mathrm{a}}) + \gamma(\Delta\mathbf{x}^{\mathrm{c}},\, \Delta\boldsymbol{\lambda}_Q^{\mathrm{c}},\, \Delta\mathbf{s}^{\mathrm{c}})\,. \tag{4.17}$$

setting $Q = \mathcal{I}$, the entire set of constraints, in Step 2 of Algorithm CR-MPC.

Step 9. Set the primal and dual step sizes $\alpha_{\rm p}$ and $\alpha_{\rm d}$ by

$$\begin{aligned}
\bar{\alpha}_{\rm p} &:= \mathrm{argmax}\{\alpha : \mathbf{s} + \alpha\Delta\mathbf{s} \geq 0\},\\
\bar{\alpha}_{\rm d} &:= \mathrm{argmax}\{\alpha : \boldsymbol{\lambda}_Q + \alpha\Delta\boldsymbol{\lambda}_Q \geq 0\},\\
\alpha_{\rm p} &:= \min\{1,\, \max\{\varkappa\bar{\alpha}_{\rm p},\, \bar{\alpha}_{\rm p} - \|\Delta\mathbf{x}\|\}\},\\
\alpha_{\rm d} &:= \min\{1,\, \max\{\varkappa\bar{\alpha}_{\rm d},\, \bar{\alpha}_{\rm d} - \|\Delta\mathbf{x}\|\}\}\,.
\end{aligned} \tag{4.18}$$

Step 10. Update variables:

$$(\mathbf{x}^+, \hat{\boldsymbol{\lambda}}_Q, \mathbf{s}^+) := (\mathbf{x}, \boldsymbol{\lambda}_Q, \mathbf{s}) + (\alpha_{\rm p}\Delta\mathbf{x}, \alpha_{\rm d}\Delta\boldsymbol{\lambda}_Q, \alpha_{\rm p}\Delta\mathbf{s})\,. \tag{4.19}$$

Set $\tilde{\boldsymbol{\lambda}}^{\rm a}_Q := \boldsymbol{\lambda}_Q + \Delta\boldsymbol{\lambda}^{\rm a}_Q$, $\quad \chi := \|\Delta\mathbf{x}^{\rm a}\|^2 + \|[\tilde{\boldsymbol{\lambda}}^{\rm a}_Q]_-\|^2$, $\quad$ and $\quad \tilde{\lambda}_i = \begin{cases} \lambda_i + \Delta\lambda_i, & i \in Q \\ 0, & i \notin Q \end{cases}$.

Update

$$\lambda_i^+ := \min\{\lambda^{\max},\, \max\{\hat{\lambda}_i,\, \min\{\underline{\lambda},\, \chi\}\}\},\ \forall i \in Q\,. \tag{4.20}$$

Set $\mu^+_{(Q)} = (\mathbf{s}^+_Q)^T(\boldsymbol{\lambda}^+_Q)/q$.

Update

$$\hat{\lambda}_i := \mu^+_{(Q)}/s_i^+, \quad \lambda_i^+ := \min\{\lambda^{\max},\, \max\{\hat{\lambda}_i,\, \min\{\underline{\lambda},\, \chi\}\}\},\ \forall i \notin Q\,. \tag{4.21}$$

In the rest of this section, we will consider the case when $\varepsilon = 0$. Suppose that, at each iteration k, the working set Q^k satisfies the following rank condition, which implies the positive definiteness of $M_{(Q^k)}$[3] (See Lemma 4).

[3]In the case that Condition 1 is not satisfied or cannot be easily verified, the regularization scheme proposed in [39] can be applied as an alternative.

Condition 1.

$$\operatorname{rank}([H, A_{Q^k}^T]) = n\,. \tag{4.22}$$

Then, we show in Proposition 6 below that, given a strictly feasible starting point and under some additional assumptions on the problem, Algorithm CR-MPC either attains the solution after finitely many iterations, or generates well defined infinite sequences $\{\mathbf{x}^k\}$, $\{\boldsymbol{\lambda}^k\}$, and $\{Q^k\}$. In the rest of this chapter, we assume that infinite sequences are generated.

4.1.4 Guidelines for Selecting the Working Set Q

As discussed in [37, 38, 40], the quality of the search directions of constraint-reduced algorithms is highly dependent on the choice of the working set Q. Consider the case that the solution $\mathbf{x}^*$ to (P) is unique and we have prior knowledge of the active constraint set $\mathcal{A}(\mathbf{x}^*)$ at solution $\mathbf{x}^*$. Then, if the working set Q is such that $\operatorname{rank}([H, A_Q^T]) = n$ and $\mathcal{A}(\mathbf{x}^*) \subseteq Q$, solving (P) is equivalent to solving the reduced problem

$$\begin{aligned} &\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\ \frac{1}{2}\mathbf{x}^T H\mathbf{x} + \mathbf{c}^T\mathbf{x} \\ &\text{subject to } A_Q\mathbf{x} \geq \mathbf{b}_Q\,. \end{aligned} \tag{4.23}$$

In this case, (4.9) and (4.12) give the "exact" affine-scaling and corrector directions for solving (4.23), at a much lower computation cost if $q \ll m$.

However, obtaining $\mathcal{A}(\mathbf{x}^*)$ is as difficult as solving the problem (P). Hence, at iteration k, the constraint reduction mechanism approximates $\mathcal{A}(\mathbf{x}^*)$ by a working

set Q^k, which is updated at each iteration by some wisely formulated constraint selection rule.

We provide the following condition on constraint selection rules which guarantees the global convergence property and the local q-quadratic convergence rate for Algorithm CR-MPC.

Condition 2. *Let $\{\mathbf{x}^k\}$ be the sequence constructed by Algorithm CR-MPC with the constraint selection rule under consideration, and let Q^k be the working set generated by the constraint selection rule at iteration k. Then for all $\mathbf{x}' \notin \mathcal{F}_P^*$, there exists $\epsilon > 0$ such that $\mathcal{A}(\mathbf{x}') \subseteq Q^k$ for all k such that $\|\mathbf{x}^k - \mathbf{x}'\| < \epsilon$.*

Condition 2 gives a desirable property that when the iterate is sufficiently close to some non-optimal point, the working set includes the set of active constraints at the non-optimal point.

The details of the constraint selection rules will be discussed in Section 4.2. However, the following convergence and convergence rate analysis are valid for all constraint selection rules satisfying Conditions 1 and 2.

4.1.5 Convergence Analysis

In this section, we summarize the convergence properties of Algorithm CR-MPC. The detailed proofs for global convergence and local convergence rate are included in Appendices B and C, respectively.

4.1.5.1 Global Convergence

The following assumptions are made to prove the global convergence property for Algorithm CR-MPC. The first assumption is necessary to guarantee the existence of a working set Q that satisfies Condition 1.

Assumption 1. $[H \ A^T]$ *has full row rank.*

The second assumption provides the existence of a strictly feasible starting point for Algorithm CR-MPC.

Assumption 2. $\mathcal{F}_P^o \neq \emptyset$.

The third assumption implies that the solution set to the primal problem (P) is nonempty, and that the infinite sequence $\{\mathbf{x}^k\}$ generated by Algorithm CR-MPC is bounded.

Assumption 3. $\mathcal{F}_P^* \neq \emptyset$, *and bounded.*

The fourth assumption assumes that the gradients of active constraints are linearly independent.

Assumption 4. $\forall \mathbf{x} \in \mathcal{F}_P$, $\{\mathbf{a}_i : i \in \mathcal{A}(\mathbf{x})\}$ *is a linearly independent set.*

The last assumption gives desirable properties on the constraint selection rule and the working set, as discussed in Section 4.1.4.

Assumption 5. *At each iteration k, the working set Q^k is selected based on some constraint selection rule such that Conditions 1 and 2 are satisfied.*

With these assumptions, the global convergence property of the primal variable $\mathbf{x}$ for Algorithm CR-MPC is given in the following Theorem.

Theorem 3. *$\{\mathbf{x}^k\}$ converges to $\mathcal{F}_P^*$.*

Proof. See Appendix B. □

The following proposition shows that, under Assumptions 1–5, for any tolerance $\varepsilon > 0$, the stopping criterion (4.15) in Algorithm CR-MPC will eventually be achieved.

Proposition 3. *Under Assumptions 1–5, given any $\varepsilon > 0$, the iterate generated by Algorithm CR-MPC eventually meets the stopping criterion (4.15).*

The proof for Proposition 3 is included in Appendix B, after the proof for Theorem 3.

Note that, in addition to Proposition 3, the global convergence of the primal-dual iterate $(\mathbf{x}^k, \tilde{\boldsymbol{\lambda}}^k)$ generated by Algorithm CR-MPC can be proved under one additional assumption as following.

Assumption 6. *$\mathcal{F}_P^*$ is a singleton.*

This assumption gives uniqueness of the primal solution, and the uniqueness of the dual solution is then a direct consequence of Assumption 4. The global convergence property of $(\mathbf{x}^k, \tilde{\boldsymbol{\lambda}}^k)$ is given in the following proposition.

Proposition 4. *Under Assumptions 1–6, the sequence of primal-dual iterates $\{(\mathbf{x}^k, \tilde{\boldsymbol{\lambda}}^k)\}$ generated by Algorithm CR-MPC converges to $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$, where $\mathbf{x}^* \in \mathcal{F}_P^*$ is the unique*

primal solution, and $\boldsymbol{\lambda}^*$ *is the multiplier associated with* $\mathbf{x}^*$*. Also, for sufficiently large* k*, the working set* Q^k *contains* $\mathcal{A}(\mathbf{x}^*)$*.*

Proof. See Appendix B. □

4.1.5.2 Local q-quadratic Convergence Rate

To obtain the local q-quadratic convergence rate for Algorithm CR-MPC, the following additional assumption is needed.

Assumption 7. *The Lagrange multipliers* $\boldsymbol{\lambda}^*$ *associated with the optimal solution* $\mathbf{x}^* \in \mathcal{F}_P^*$ *are strictly complementary to* $\mathbf{s}^* := A^T\mathbf{x}^* - \mathbf{b}$*, i.e.,* $\lambda_i^* s_i^* = 0$ *and* $\lambda_i^* + s_i^* > 0$ *for all* $i \in \mathcal{I}$*.*

This assumption gives the strict complementary property. With Assumptions 1–7, the following theorem gives the local q-quadratic convergence rate of Algorithm CR-MPC.

Theorem 4. *Let* $\mathbf{x}^* \in \mathcal{F}_P^*$ *be the unique solution to* (P) *and* $\boldsymbol{\lambda}^*$ *be the Lagrange multiplier associated to* $\mathbf{x}^*$*. If* $\lambda_i^* < \lambda_{\max}$ *for all* $i \in \mathcal{I}$*, the infinite sequence* $\{(\mathbf{x}^k, \boldsymbol{\lambda}^k)\}$ *generated by Algorithm CR-MPC then converges locally q-quadratically to* $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$*. In other words,* $(\mathbf{x}^k, \boldsymbol{\lambda}^k) \to (\mathbf{x}^*, \boldsymbol{\lambda}^*)$ *and there exist* $k' > 0$ *and* $c^* > 0$ *such that, for all* $k > k'$*, we have*

$$\|(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}) - (\mathbf{x}^*, \boldsymbol{\lambda}^*)\| \leq c^* \|(\mathbf{x}^k, \boldsymbol{\lambda}^k) - (\mathbf{x}^*, \boldsymbol{\lambda}^*)\|^2 . \tag{4.24}$$

Proof. See Appendix C. □

4.2 A Constraint Selection Rule

Several constraint selection rules were proposed for constraint-reduced algorithms on various classes of optimization problems, such as LP [37–39], CQP [40,41], and SDP [42, 43]. We propose a new constraint selection rule and apply it on Algorithm CR-MPC for solving CQPs. Our proposed rule selects the working set Q^k such that Algorithm CR-MPC satisfies Conditions 1 and 2. Thus the convergence properties in Section 4.1.5 hold for Algorithm CR-MPC with the proposed constraint selection rule.

The proposed rule first computes a threshold value based on the decrement of the error, and then selects constraints by including all constraints with slack values less than the computed threshold. The details of the proposed constraint selection rule are presented below.

In the remainder of this section, we will consider the case that Condition 1 is satisfied, and the techniques used to guarantee Condition 1 will be discussed in Section 4.3. If Condition 1 holds, the following lemma then shows that Rule 1 satisfies Condition 2.

Lemma 5. *Algorithm CR-MPC with Rule 1 satisfies Condition 2.*

Proof. Let $\{\mathbf{x}^k\}$ be the sequence generated by Algorithm CR-MPC using Rule 1, let $\mathbf{x}' \notin \mathcal{F}_P^*$ be a non-optimal point, and, for any $\epsilon > 0$, let $K(\epsilon) := \{k : \|\mathbf{x}^k - \mathbf{x}'\| < \epsilon\}$. Since E^k is continuous in $\mathbf{x}^k$ and $\boldsymbol{\lambda}^k$, it cannot vanish unless $\mathbf{x}^k$ is optimal. Thus, there exists $\epsilon_1 > 0$ such that $\{E^k\}$ is bounded away from zero on $K(\epsilon_1)$. Hence, it

Rule 1 Constraint selection rule

Parameters: $\bar{\delta} > 0$, $\beta \in (0, 1)$, $\theta \in (0, 1)$, and $\gamma \geq 1$.

Input: Iteration: k, Slack variable: $\mathbf{s}^k$, Error: $E_{\min}$, E^k, Threshold: δ^{k-1}

Output: Working set: Q^k, Threshold: δ^k, Error: $E_{\min}$

if $k = 1$ **then**

  $\delta^k := \bar{\delta}$

  $E_{\min} := E^k$

else if $E^k \leq \beta E_{\min}$ **then**

  $\delta^k := \theta \delta^{k-1}$

  $E_{\min} := E^k$

else

  $\delta^k := \gamma \delta^{k-1}$

end if

Select Q^k to include $\{i \in \mathcal{I} \mid s_i^k \leq \delta^k\}$ and to satisfy Condition 1.

follows from Rule 1 that the sequence $\{\delta^k\}_{k\in K(\epsilon_1)}$ is also bounded away from zero. Thus, there exists some $\delta' > 0$ such that $\delta^k > \delta' > 0$ for all $k \in K(\epsilon_1)$.

On the other hand, let $\mathbf{s}^k := A\mathbf{x}^k - \mathbf{b}$ for all k, and $\mathbf{s}' := A\mathbf{x}' - \mathbf{b}$. Definition 4 implies that $\mathbf{s}'_{\mathcal{A}(\mathbf{x}')} = \mathbf{0}$. Thus, by continuity, there exists $\epsilon_2 > 0$ such that, for all $i \in \mathcal{A}(\mathbf{x}')$, $s_i^k < \delta'$ for all $k \in K(\epsilon_2)$.

Let $\epsilon := \min\{\epsilon_1, \epsilon_2\}$, then $K(\epsilon) \subseteq K(\epsilon_1)$ and $K(\epsilon) \subseteq K(\epsilon_2)$. We have, for all $i \in \mathcal{A}(\mathbf{x}')$,

$$s_i^k < \delta' < \delta^k\,, \quad \forall k \in K(\epsilon)\,. \tag{4.25}$$

Since Rule 1 requires that, at each iteration k, the working set Q^k must include $\{i \in \mathcal{I} \,|\, s_i^k \leq \delta^k\}$, we then conclude that $\mathcal{A}(\mathbf{x}') \subseteq Q^k$ for all $k \in K(\epsilon)$. □

Lemma 5 implies that Rule 1 satisfies Assumption 5. Thus, under Assumptions 1–5, the following corollary is a direct consequence of Theorems 3.

Corollary 1. *The sequence of primal iterates $\{\mathbf{x}^k\}$ generated by Algorithm CR-MPC with Rule 1 converges globally to the primal optimal set $\mathcal{F}_P^*$.*

In addition, under Assumptions 1–7, the following corollary is a direct consequence of Theorem 4.

Corollary 2. *The sequence of primal-dual iterates $\{(\mathbf{x}^k, \boldsymbol{\lambda}^k)\}$ generated by Algorithm CR-MPC with Rule 1 converges to the unique primal-dual optimal solution $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ with a locally q-quadratic rate.*

4.3 Numerical Experiments

In this section, we compare the computational performance of Algorithm CR-MPC and the constraint-reduced primal-dual affine-scaling algorithm (CR-AS) proposed in [40], which is similar to the affine-scaling special case of Algorithm CR-MPC by choosing parameter $\tau = 0$, while the dual variable $\boldsymbol{\lambda}$ is updated in a slightly different way.

Both algorithms are implemented in both MATLAB and C++ with various constraint selection rules, including the proposed Rule 1, the adaptive constraint selection rule proposed in [40], and the rule that selects all constraints, i.e., no reduction. For ease of reference, the adaptive constraint selection rule of [40] is stated below as Rule 2. Note that Rule 2 also satisfies Condition 1 and 2. However, it requires the size of the working set to be at least n, while Rule 1 allows working sets to contain fewer than n constraints.

Rule 2 Constraint selection rule

Input: Iteration: k, Slack variable: $\mathbf{s}^k$, Duality measure: μ^k

Output: Working set: Q^k

Set

$$q := \begin{cases} n & \text{, if } \mu^{1/4} m \leq n \\ \mu^{1/4} m & \text{, if } n < \mu^{1/4} m \leq m \\ m & \text{, if } m < \mu^{1/4} m \end{cases}$$

Select Q^k to include indexes of constraints with the q smallest slack values $\mathbf{s}^k$ and to satisfy Condition 1.

To highlight the effect of constraint selection rules, a dense direct solver is used to solve normal equations (4.10) and (4.13). The algorithms are set to terminate when the stopping criterion (4.15) is satisfied or when the iteration count reaches 200. The algorithm parameter values used in the tests are $\varepsilon = 10^{-4}$, $\tau = 0.5$, $\zeta = 0.9$, $\varkappa = 0.98$, $\underline{\lambda} = 10^{-6}$, and $\lambda^{\max} = 10^{30}$. For the randomly generated problems, the parameters in Rule 1 take values $\beta = 0.5$, $\theta = 0.5$, $\gamma = 1$, and $\bar{\delta}$ is defined as the n-th smallest slack value at the first iteration. For the problems in the FP_N^+ closure, we choose $\beta = 0.3$ and $\theta = 0.3$ to reflect the smaller size of the problems. Following [38] and [40], we impose a safeguard on the slack variable $\mathbf{s}$ by taking $\mathbf{s} = \max\{\mathbf{s}, 10^{-14}\}$ when generating the approximate normal matrix $M_{(Q)}$ defined in (4.11). Such safeguard prevents $M_{(Q)}$ from being too ill-conditioned.

For Rule 1, if Condition 1 is not satisfied, i.e., the approximate normal matrix $M_{(Q)}$ is ill-conditioned (see Lemma 4), causing the MATLAB factorization routine `chol` or the LAPACK factorization routine `dpotrf` to fail, we double δ, and select a new working set Q based on the new threshold δ, repeatedly as necessary. For Rule 2, q, instead of δ, is doubled in such situation. In rare cases, the matrix $M_{(Q)}$ could still be ill-conditioned even when the working set contains all constraints. The regularization technique proposed in [39] can be applied to resolve the issue.

4.3.1 Randomly Generated Problems

To illustrate the strength of the constraint reduction technique, we first compare the algorithms and constraint selection rules on randomly generated problems

with a much larger number of inequality constraints than the number of variables, i.e., $m \gg n$. We consider two types of such problems, each with numbers of constraints $m := 1000$ and $m := 10000$, and numbers of variables n varies from 1 to 200.

The problems are generated in a similar way as the randomly generated problems tested in [38] and [40]. Each problem takes the form (P). The elements in A and $\mathbf{c}$ are taken from a standard normal distribution $\mathcal{N}(0, 1)$. We generate $\mathbf{x}_0$ and $\mathbf{s}_0$ taking random numbers from uniform distributions $\mathcal{U}(0, 1)$ and $\mathcal{U}(1, 2)$, respectively, and set $\mathbf{b} := A\mathbf{x}_0 - \mathbf{s}_0$, which guarantees that the feasible set is nonempty and $\mathbf{x}_0$ is strictly feasible. For strictly convex problems, the Hessian matrix H is a diagonal matrix with diagonal elements taken from uniform distributions $\mathcal{U}(0, 1)$. For (non-strictly) convex problems, we first generate an n-by-n' matrix B, where n' is the smallest integer greater than $0.9n$, and set the Hessian matrix $H := B^T B$. In order to guarantee convergence, we keep generating the Hessian matrix until Assumption 1 is satisfied.

In our experiment, each randomly generated problem is solved in MATLAB using both Algorithm CR-MPC and Algorithm CR-AS with the three constraint selection rules – Rule 1, Rule 2, and full constraints.

Figure 4.1–4.16 illustrate the iteration counts, average size of the working set per iteration, and the timing results for each pair of algorithm and rule on both strictly and non-strictly convex problems with number of constraints $m = 1000$, 10000 and various numbers of variables n. We solve 1000 randomly generated problems for each pair of m and n, and each data point in the figures is the average

over the 1000 solved problems. In the figures, blue diamonds denote Algorithm CR-AS with Rule 1; red squares denote Algorithm CR-MPC with Rule 1; green stars denote Algorithm CR-AS with Rule 2; purple crosses denote Algorithm CR-MPC with Rule 2; black circles denote the unreduced AS algorithm; cyan triangles denote the unreduced MPC algorithm.

Figures 4.1, 4.5, 4.9, and 4.13 show the average iteration counts to solve one problem, strictly or non-strictly convex, of various sizes, and Figures 4.2, 4.6, 4.10, and 4.14 are the corresponding zoomed-in versions for methods that take less than 20 iterations in average. From these figures, we first notice that the iteration counts for the MPC algorithm and its variants are less than the iteration counts for the corresponding AS algorithms. This observation confirms that the additional corrector direction used in the MPC type algorithms indeed provides a better search direction, thus reduces the number of iterations needed to solve the problem. The second observation is that, in most cases, the unreduced MPC and AS algorithms takes the least number of iterations, and both Algorithm CR-MPC and Algorithm CR-AS with Rule 1 require more iterations than the corresponding algorithms with Rule 2. This result is not surprising, since for most problems, Rule 2 includes more constraints in the working set than Rule 1, and the unreduced algorithms use all m constraints, as shown in Figures 4.3, 4.7, 4.11, and 4.15. Including fewer constraints in the working set gives less computation cost per iteration, while it also suggests that the search direction is computed based on less information, thus the number of iterations may rise.

Figures 4.4, 4.8, 4.12, and 4.16 report the average computation time needed for

each algorithm to solve one randomly generated problem. It can be observed from these figures that, in most cases, Algorithm CR-MPC with Rule 1 steadily gives the best timing result among all other algorithms, except for the cases when $n \leq 10$. In those cases, Algorithm CR-AS with Rule 1 performs better, but the iteration counts increases rapidly as n increases. Overall, for the cases when $m = 10000$, the proposed Algorithm CR-MPC with the new selection rule Rule 1 gives roughly 10 times speedup comparing to Algorithm CR-AS with Rule 2 proposed in [40], and when $m = 1000$, Algorithm CR-MPC performs at least as well as Algorithm CR-AS. In addition, we do not observe any notable differences between the results for strictly and non-strictly convex problems.

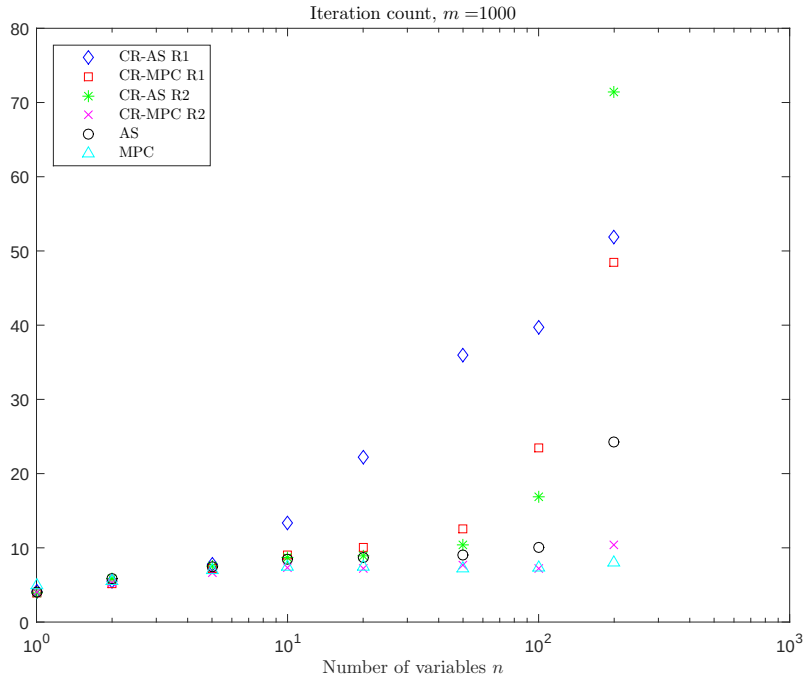


Figure 4.1: Average iteration counts for solving one strictly convex problem with $m = 1000$ and various n.

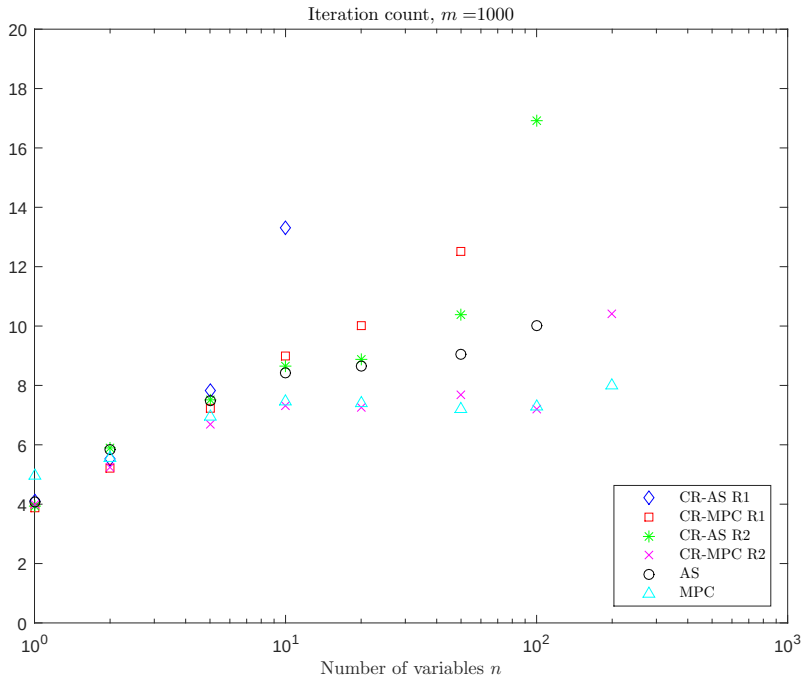


Figure 4.2: Average iteration counts (zoomed-in) for solving one strictly convex problem with $m = 1000$ and various n.

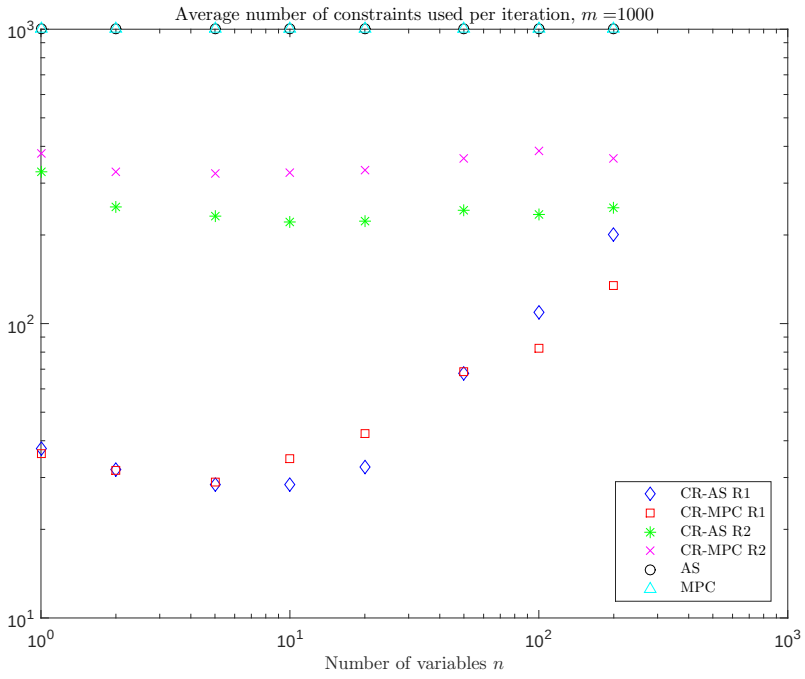


Figure 4.3: Average size of the working set in solving one strictly convex problem with $m = 1000$ and various n.

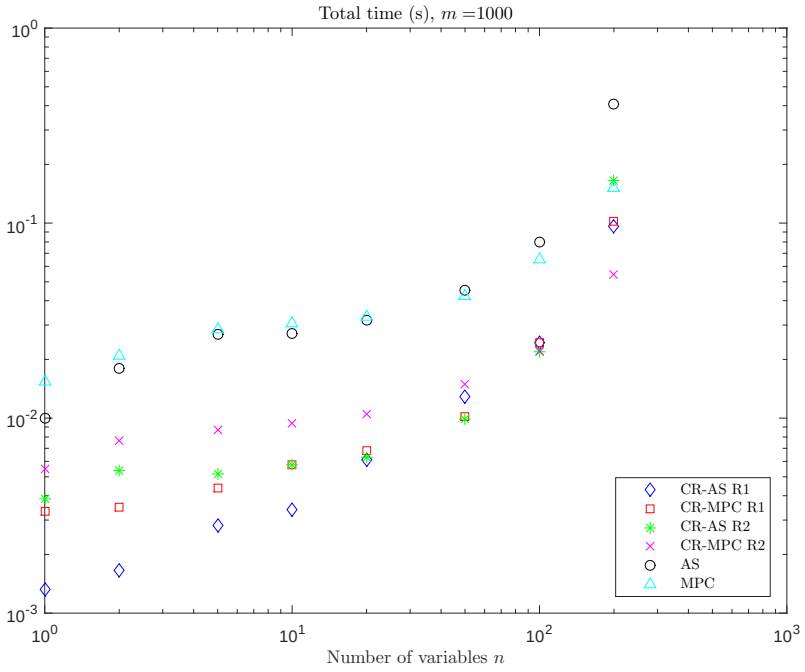


Figure 4.4: Average computation time for solving one strictly convex problem with $m = 1000$ and various n.

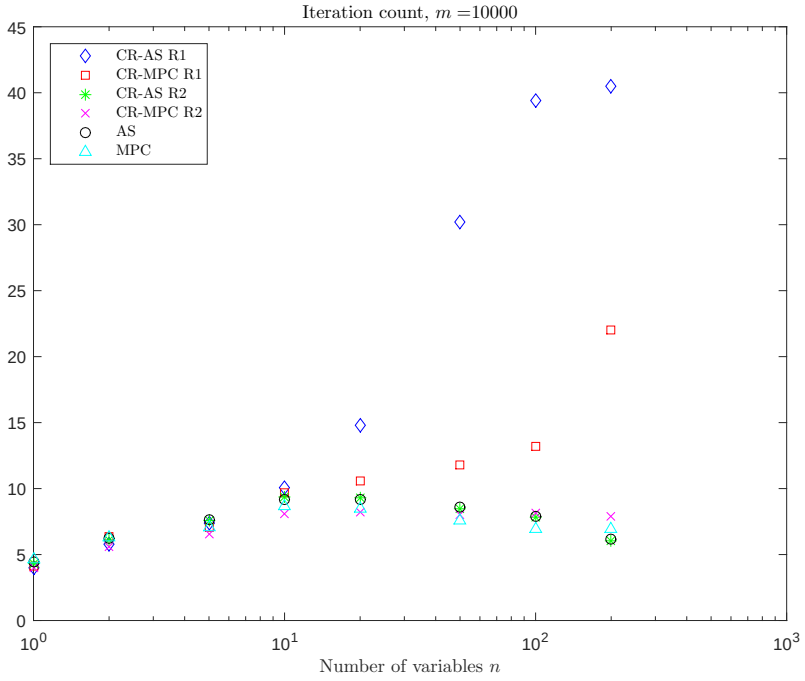


Figure 4.5: Average iteration counts for solving one strictly convex problem with $m = 10000$ and various n.

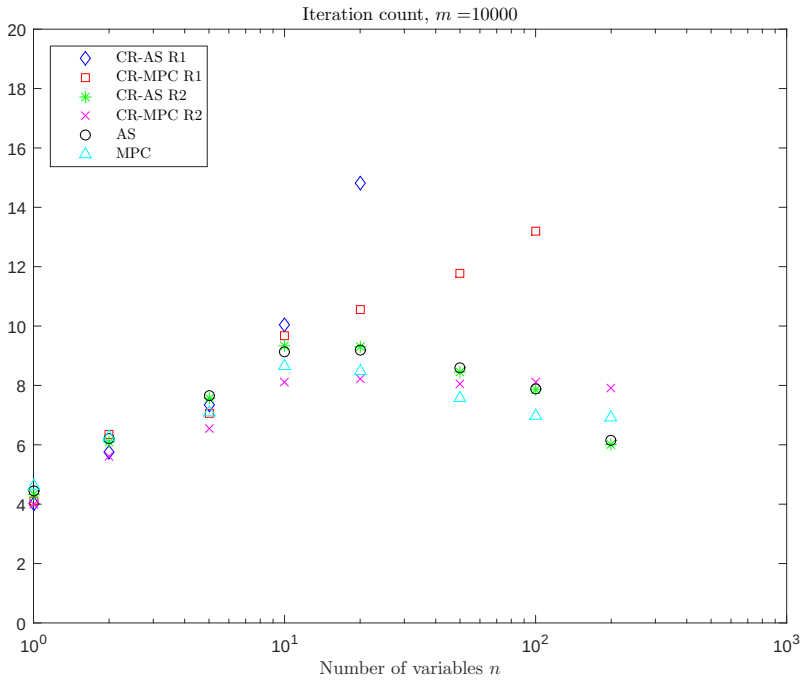


Figure 4.6: Average iteration counts (zoomed-in) for solving one strictly convex problem with $m = 10000$ and various n.

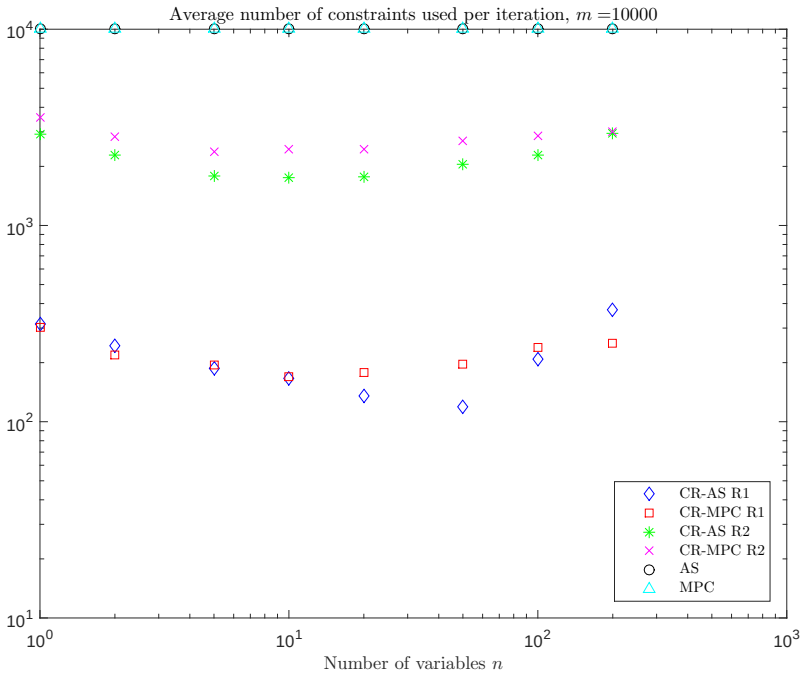


Figure 4.7: Average size of the working set in solving one strictly convex problem with $m = 10000$ and various n.

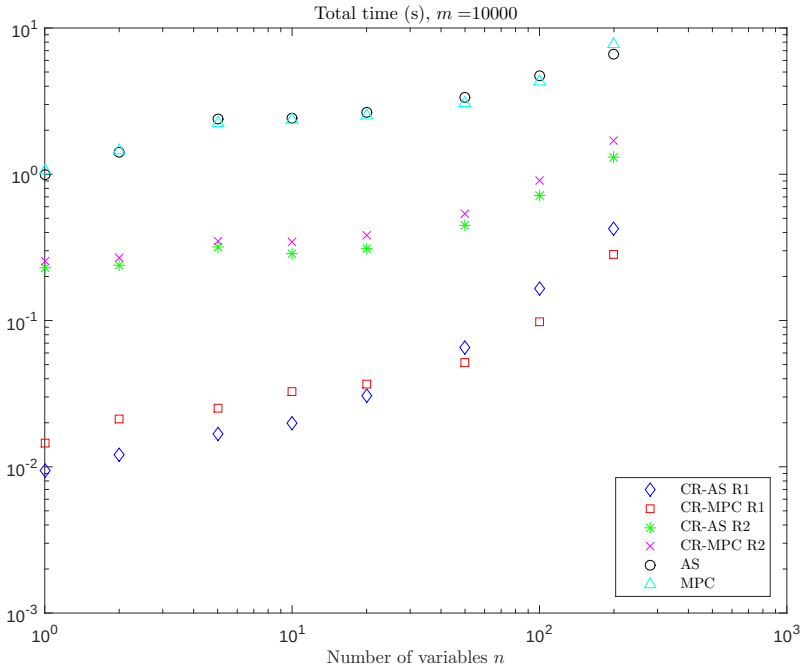


Figure 4.8: Average computation time for solving one strictly convex problem with $m = 10000$ and various n.

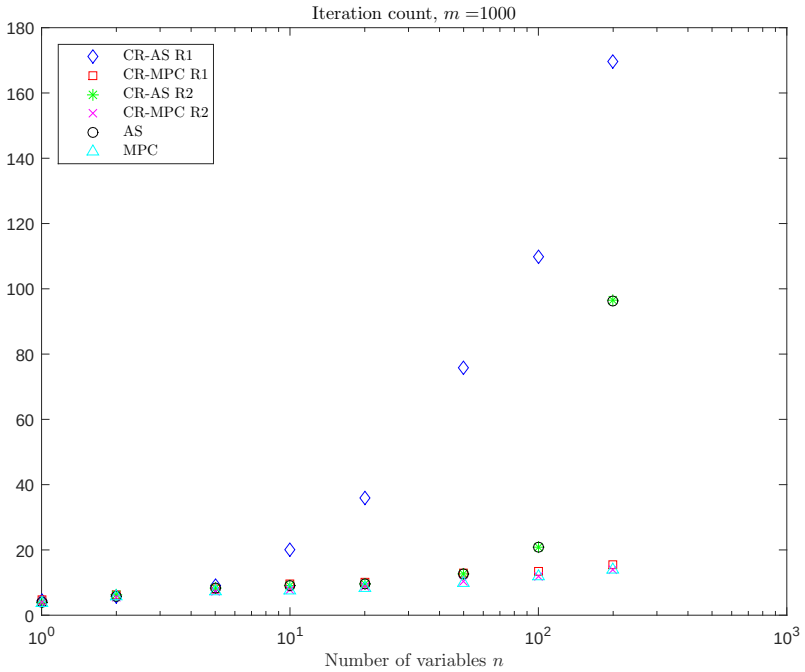


Figure 4.9: Average iteration counts for solving one non-strictly convex problem with $m = 1000$ and various n.

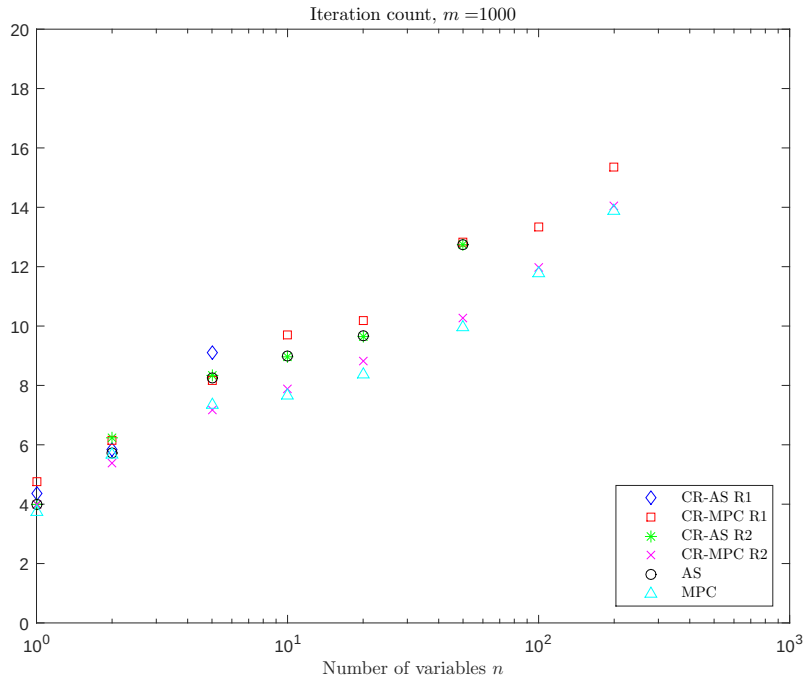


Figure 4.10: Average iteration counts (zoomed-in) for solving one non-strictly convex problem with $m = 1000$ and various n.

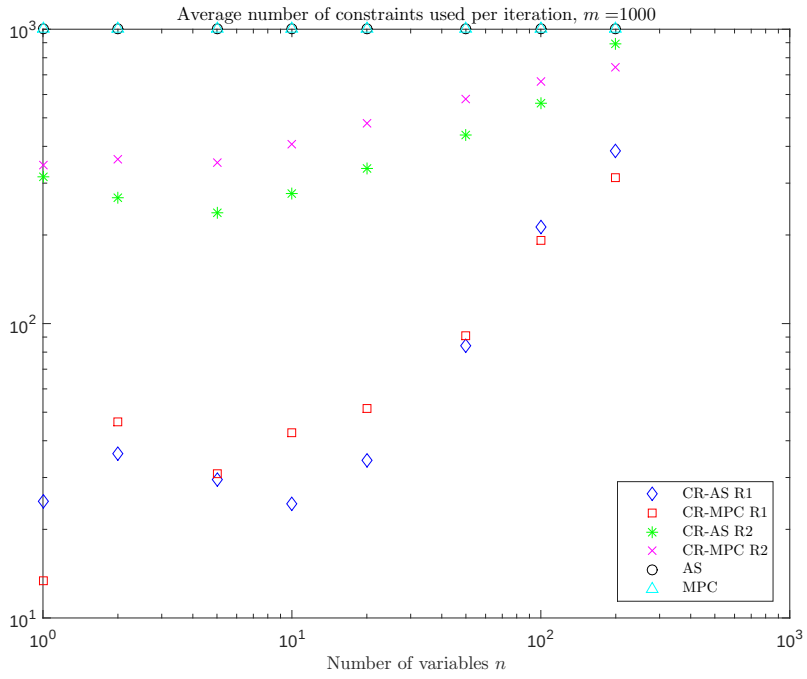


Figure 4.11: Average size of the working set in solving one non-strictly convex problem with $m = 1000$ and various n.

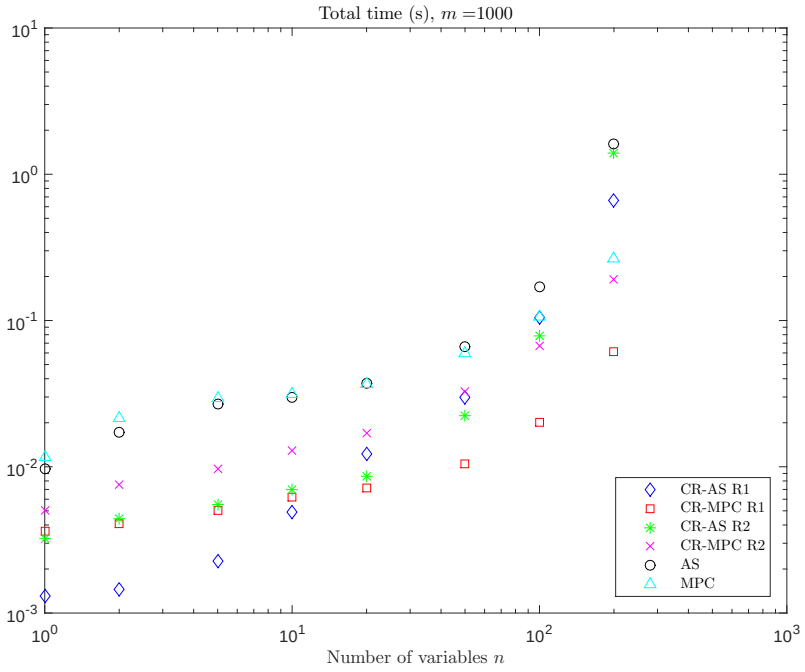


Figure 4.12: Average computation time for solving one non-strictly convex problem with $m = 1000$ and various n.

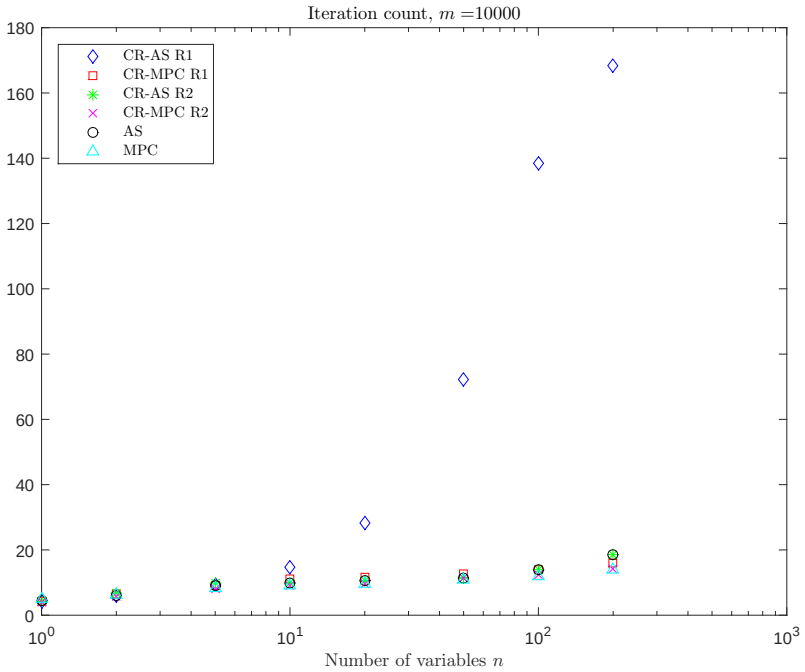


Figure 4.13: Average iteration counts for solving one non-strictly convex problem with $m = 10000$ and various n.

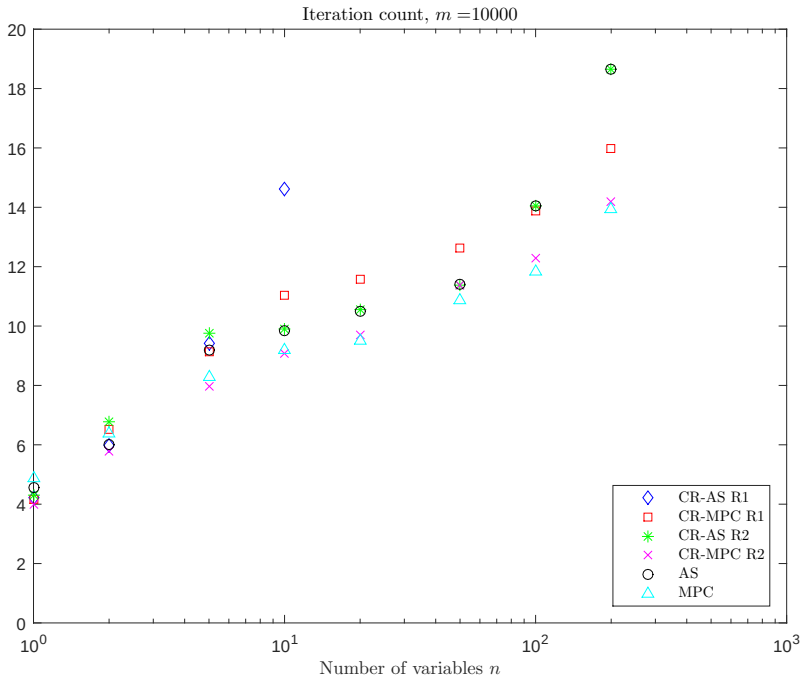


Figure 4.14: Average iteration counts (zoomed-in) for solving one non-strictly convex problem with $m = 10000$.

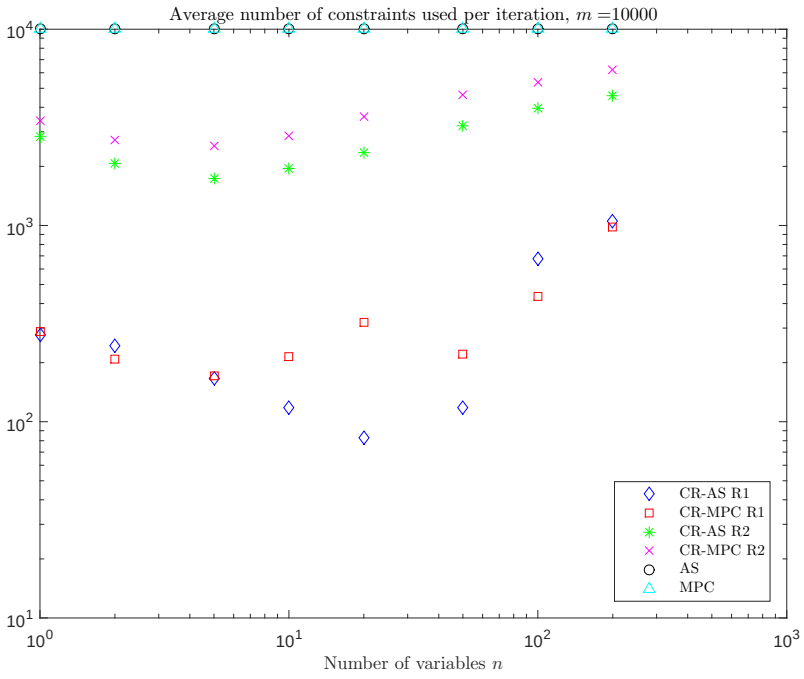


Figure 4.15: Average size of the working set in solving one non-strictly convex problem with $m = 10000$ and various n.

4.3.2 The FP_N^+ Closure for Linear Transport Equation

In this section, we test Algorithm CR-MPC and Algorithm CR-AS with the three constraint selection rules on solving the CQPs in the FP_N^+ closure for the line source problems introduced in Section 2.4.1. The calculation of the solution to the line source problem is formulated exactly the same as the test presented in Section 2.4.1 and 2.4.3. The computation times for the line source calculations with various algorithms and constraint selection rules are listed in Table 4.1. The test is performed for the case when moment order $N = 11$, which makes the total number of optimization variables to be $n = (N + 1)(N + 2)/2 - 1 = 77$. The tested quadrature rules are with degrees of precision 23 and 47, and the number of inequality constraints is given by the number of quadrature points: 144 and 576, respectively.

From Table 4.1, it can be observed that the computation time for the FP_N^+ method depends heavily on the optimization algorithm and the number of quadrature points. For $N_{\mathcal{Q}} = 47$, Algorithm CR-MPC and Algorithm CR-AS with Rule 1 and Rule 2 both reduce the computation time for the FP_N^+ method by about a factor of two comparing to the unreduced algorithms. For $N_{\mathcal{Q}} = 23$, the benefit of constraint reduction is less significant ($10 \sim 20\%$), as the number of constraints in the optimization problem is lower. Overall, the MPC algorithms performs better than the AS algorithms in most cases, but there is not much difference observed between algorithms using Rule 1 and Rule 2.

Quadrature Type	Product	Product
Degree	$N_{\mathcal{Q}} = 23$	$N_{\mathcal{Q}} = 47$
# of quadrature points	$\lvert\mathcal{Q}\rvert = 144$	$\lvert\mathcal{Q}\rvert = 576$
CR-AS R1	5838	21358
CR-MPC R1	5424	16515
CR-AS R2	5731	16277
CR-MPC R2	5929	13925
AS	7726	32941
MPC	6600	27319

Table 4.1: The computation times (sec) for the line source benchmark with the FP_N^+ closures with $N = 11$. The optimization problems in the FP_N^+ closure are solved by Algorithm CR-MPC and Algorithm CR-AS, with all three constraint selection rules.

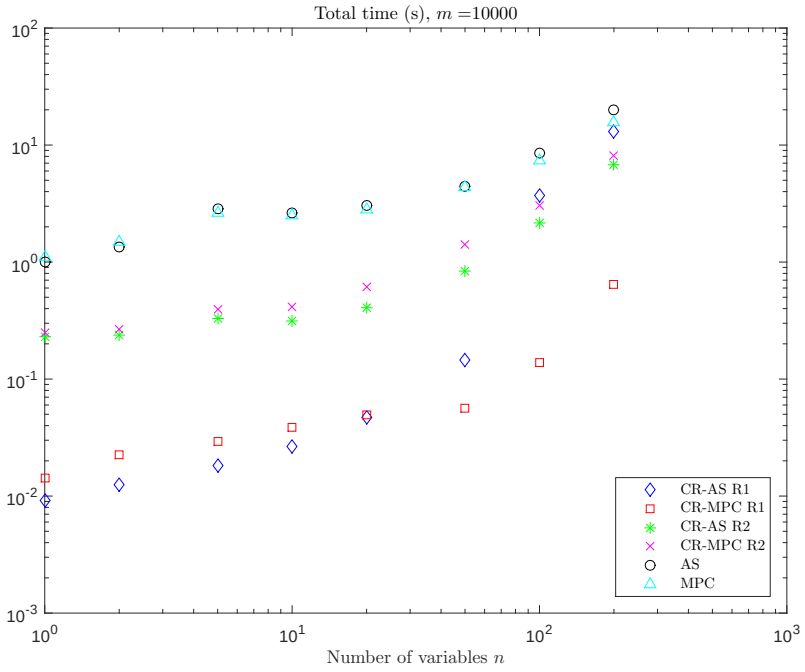


Figure 4.16: Average computation time for solving one non-strictly convex problem with $m = 10000$ and various n.

Chapter 5: Conclusions and Future Work

In this work, we have presented a new moment closure, the FP_N^+ closure, for generating approximate solutions of linear transport equations. The FP_N^+ closure preserves the positivity of particle concentration, while reducing oscillations that occur in many other positive closures, such as PP_N and M_N. The new closure is based on the solution of an optimization problem that modifies the coefficients in the usual filtered spherical harmonic expansion by enforcing positivity on a properly chosen quadrature set.

We have proven that, for target functions in the space C^q, where $q \geq 0$ is an integer, the FP_N^+ approximation converges in L^2 at the same (optimal) rate as the FP_N approximation. In practice, such convergence rate is observed for target functions in the larger space H^q. For one-dimensional problems with quadrature rules of the minimum required degree of precision, we are able to prove the optimal convergence rate for space H^q. However, the proof for the optimal convergence rate of the FP_N^+ approximation in general H^q spaces will be the subject of future work.

We have also investigated a simpler positive closure, which we refer to as the UD_N closure, based on a spatial limiter developed in [21] for finite volume schemes. For functions in H^q, we prove suboptimal convergence rates for the UD_N

approximation. Based on numerical tests, we believe that these rates are sharp. For problems with less regularity, we expect that the additional accuracy of the FP_N^+ closure compared to that of the UD_N approach will outweigh the additional cost. Our simulation results confirm this fact in the case of the line source benchmark. They also show that the UD_N closure degrades the space-time convergence rate of the PDE solver for the moment equations. For the FP_N^+ closure, we observe minimal, if any, such effect. For more regular problems, we expect the accuracy of the two closures to be comparable. In fact, we have observed this for other test problem results not reported here. For these problems, the UD_N closure may be more efficient, and a more careful comparison will be performed in later work.

We have proposed a positive preserving AP scheme for solving the transport equation near the diffusion limit. We have shown that the AP scheme requires a hyperbolic CFL stability condition in the transport regime and a parabolic CFL stability condition near the diffusion limit. The positive preserving property of the AP scheme is proved under a hyperbolic time step restriction, which can be relaxed to a parabolic time step restriction under mild assumption. The numerical results on the comparison between the AP scheme and the second-order kinetic scheme in [10] show that the proposed AP scheme gives solutions that are compatible with those obtained with the second-order kinetic scheme in the transport regime, and it indeed computes the correct diffusion limit without excessively refining the temporal-spatial mesh. On the other hand, the numerical results also confirm that the exceedingly restrictive mesh size requirements on the second-order kinetic scheme makes the scheme impractical near the diffusion limit. Some possible generalizations of the

proposed AP scheme, such as allowing for absorption and external sources, are of interest for future work.

We have presented Algorithm CR-MPC, a constraint-reduced variant of a Mehrotra-predictor-corrector algorithm. Algorithm CR-MPC was developed to solve the CQPs in the FP_N^+ method, but it can be applied to general CQPs. We have provided conditions for constraint selection rules used in Algorithm CR-MPC, and we have proved that, under those conditions, Algorithm CR-MPC enjoys global convergence and local q-quadratic convergence rate. We then proposed a new constraint selection rule, Rule 1, and proved that Rule 1 satisfies these conditions. Algorithm CR-MPC using Rule 1 is compared with Algorithm CR-AS using Rule 2 proposed in [40] on both randomly generated CQPs and the CQPs from the FP_N^+ closure. The numerical results suggest that the combination of Algorithm CR-MPC and Rule 1 is especially powerful when the CQP has many more inequality constraints than variables. In such cases, Algorithm CR-MPC with Rule 1 solved the CQP in approximately 10% of the computation time needed for Algorithm CR-AS with Rule 2, and in less than 1% of the computation time needed for the unreduced AS and MPC algorithms.

Appendix A: Numerical Integration of the Moment System

In this appendix, we present details of the kinetic scheme from Section 2.2.2.1, which was originally proposed in [15] for one-dimensional problems, and extended to two-dimensional problems in [10]. For simplicity of exposition, we restrict ourselves to the reduced equation that is used in the formulation of the line source benchmark. The reduced equation is given by

$$\partial_t f + \xi \partial_x f + \eta \partial_y f = \frac{\sigma}{4\pi} \langle f \rangle - \sigma f \,, \tag{A.1}$$

where $(\xi, \eta) = (\Omega_x, \Omega_y)$. This equation is valid when $\partial_z f = 0$ [3], and the associated closed moment equation with ansatz $\mathcal{E}$ becomes

$$\partial_t \mathbf{u} + \partial_x \langle \mathbf{m} \xi \mathcal{E} \rangle + \partial_y \langle \mathbf{m} \eta \mathcal{E} \rangle = -\sigma R \mathbf{u} \,, \tag{A.2}$$

with $\mathbf{u} = \langle \mathbf{m} \mathcal{E} \rangle$, $\mathbf{m}$ the spherical harmonic basis, and $R = \mathrm{diag}(0, 1, \ldots, 1)$ as defined in Section 2.1. The presentation below is based on the original description given in [10], which in turn is based on the algorithm in [15] for the even simpler case of slab geometry.

A.1 Spatial Discretization – Finite Volume Method

A finite volume method is used in the spatial discretization, which discretizes the rectangular spatial domain $[x_L,\, x_R]\times[y_L,\, y_U]$ on a Cartesian mesh with constant cell size. Let N_x, N_y be the number of cells on x, y direction respectively, and $\Delta x = (x_R - x_L)/N_x$ and $\Delta y = (y_U - y_L)/N_y$ be the dimensions of spatial cells. The spatial mesh is then defined as $\{x_i\}_{i=-1}^{N_x+2} \times \{y_j\}_{j=-1}^{N_y+2}$ where $x_i := x_L + (i - 0.5)\Delta x$ and $y_j := y_L + (j - 0.5)\Delta y$, with cells $C_{i,j} = [x_{i-1/2},\, x_{i+1/2}] \times [y_{j-1/2},\, y_{j+1/2}]$, where $x_{i\pm1/2} := x_i \pm \Delta x/2$ and $y_{j\pm1/2} := y_j \pm \Delta y/2$. Cells $C_{i,j}$ such that $i \in \{-1, 0, N_x + 1, N_x + 2\}$ or $j \in \{-1, 0, N_y + 1, N_y + 2\}$ are "ghost cells", which are used only to implement boundary conditions.

With the spatial mesh, the moment equation (A.2) is then discretized by approximating $\mathbf{u}$ numerically via its the cell averages, i.e.,

$$\mathbf{u}_{i,j} \simeq \frac{1}{\Delta x \Delta y} \int_{C_{i,j}} \mathbf{u}(x, y)\, dxdy\,, \tag{A.3}$$

which yields the semi-discrete moment equation

$$\partial_t \mathbf{u}_{i,j} + \left\langle \mathbf{m}\frac{\xi}{\Delta x}(\mathcal{E}_{i+1/2,\,j} - \mathcal{E}_{i-1/2,\,j}) + \mathbf{m}\frac{\eta}{\Delta y}(\mathcal{E}_{i,\,j+1/2} - \mathcal{E}_{i,\,j-1/2}) \right\rangle = -\sigma R \mathbf{u}_{i,j}\,, \tag{A.4}$$

where $\mathcal{E}_{i\pm1/2,\,j}$, $\mathcal{E}_{i,\,j\pm1/2}$ are approximations of the ansatz $\mathcal{E}(\mathbf{u})$ at the cell edges. These edge values are computed via upwinding:

$$\mathcal{E}_{i+1/2,j} := \begin{cases} \mathcal{E}_{i,j} + \frac{s^x_{i,j}}{2}\,, & \xi > 0 \\ \mathcal{E}_{i+1,j} - \frac{s^x_{i+1,j}}{2}\,, & \xi < 0 \end{cases}, \quad \mathcal{E}_{i,j+1/2} := \begin{cases} \mathcal{E}_{i,j} + \frac{s^y_{i,j}}{2}\,, & \eta > 0 \\ \mathcal{E}_{i,j+1} - \frac{s^y_{i,j+1}}{2}\,, & \eta < 0 \end{cases}, \tag{A.5}$$

where $s^x_{i,j}$, $s^y_{i,j}$ are approximations of the spatial derivative of $\mathcal{E}(\mathbf{u})$ in x, y directions, respectively. For P_N and FP_N closures, the spatial derivatives are approximated by the centered difference, i.e.,

$$s^x_{i,j} = \frac{\mathcal{E}_{i+1,j} - \mathcal{E}_{i-1,j}}{2}, \quad s^y_{i,j} = \frac{\mathcal{E}_{i,j+1} - \mathcal{E}_{i,j-1}}{2}, \tag{A.6}$$

which yield a simpler computation for the flux term; see [10] for details. For the FP^+_N closure, a minmod limiter is required in the approximation of spatial derivatives in order to preserve positivity. The approximations are given as

$$s^x_{i,j} = \text{minmod}\left\{\vartheta(\mathcal{E}_{i,j} - \mathcal{E}_{i-1,j}), \frac{\mathcal{E}_{i+1,j} - \mathcal{E}_{i-1,j}}{2}, \vartheta(\mathcal{E}_{i+1,j} - \mathcal{E}_{i,j})\right\}, \tag{A.7}$$

$$s^y_{i,j} = \text{minmod}\left\{\vartheta(\mathcal{E}_{i,j} - \mathcal{E}_{i,j-1}), \frac{\mathcal{E}_{i,j+1} - \mathcal{E}_{i,j-1}}{2}, \vartheta(\mathcal{E}_{i,j+1} - \mathcal{E}_{i,j})\right\}, \tag{A.8}$$

where $1 \leq \vartheta \leq 2$ [71, 72][1] and the minmod limiter is given in (3.39). This limiter ensures the non-negativity for FP^+_N, but requires explicit evaluation of the edge values of $\mathcal{E}_{\mathrm{FP}^+_N}$, which are computed via a cross-product quadrature described in Section 2.2.2.3.

A.2 Updates in Time

We integrate (A.4) in time using Heun's method. This method preserves non-negativity of the solution with FP^+_N closures under certain time step restriction;

[1] Any value of $\vartheta \in [1, 2]$ will yield a formally second-order scheme; roughly speaking, larger values of ϑ decrease numerical diffusion in the scheme. When $\vartheta = 1$, monotonic cell averages yield monotonic reconstructions. When $\vartheta = 2$, edge values are bounded by neighboring cell averages.

see Theorem 5 for details. Let $t^k = t_0 + k\Delta t$, with initial time t_0, and $\mathbf{u}_{i,j}^k$ be the approximation of $\mathbf{u}_{i,j}(t^k)$. For (A.4) in the abstract form $\partial_t \mathbf{u} = L(\mathbf{u})$ at the initial stage $\mathbf{u}^{(0)} := \mathbf{u}^k$, Heun's method is given as

$$\mathbf{u}^{(1)} := \bar{\mathbf{u}}^{(0)} + \Delta t L(\bar{\mathbf{u}}^{(0)}), \quad \mathbf{u}^{(2)} := \bar{\mathbf{u}}^{(1)} + \Delta t L(\bar{\mathbf{u}}^{(1)}), \quad \mathbf{u}^{k+1} := \frac{1}{2}\left(\bar{\mathbf{u}}^{(0)} + \mathbf{u}^{(2)}\right), \tag{A.9}$$

where $\bar{\mathbf{u}}^{(i)} = \langle \mathbf{m}\mathcal{E}[\mathbf{u}^{(i)}]\rangle$ for $i = 0, 1$. Here $\mathcal{E}$ may be a general ansatz; for example, it can be $\mathcal{E}_{\mathrm{P}_N}$, $\mathcal{E}_{\mathrm{FP}_N}$, or $\mathcal{E}_{\mathrm{FP}_N^+}$. However, if $\bar{\mathbf{u}}^{(i)} \neq \mathbf{u}^{(i)}$, the formal order of the method may be reduced.

A.3 Positivity

The following theorem shows that the non-negative ansatz generated by the FP_N^+ closure leads to a non-negative particle concentration in the solution generated by the kinetic scheme.

Theorem 5. *Suppose that* $\mathcal{E}[\mathbf{u}_{i,j}^k] \geq 0$ *for all* $(i,j) \in \{-1,\ldots,N_x+2\}\times\{-1,\ldots,N_y+2\}$. *Then the particle concentration* $\rho_{i,j}^{k+1} := \sqrt{4\pi}u_{0,i,j}^{k+1}$, *where* $u_{0,i,j}^{k+1}$ *is the zeroth moment of* $\mathbf{u}_{i,j}^{k+1}$, *is non-negative under the time-step restriction*

$$\Delta t \leq \left(\frac{2}{\vartheta + 2}\right)\frac{\Delta x \Delta y}{\Delta x + \Delta y} \tag{A.10}$$

where Δt, Δx, *and* Δy *are the time step and the spatial mesh size respectively, and* $\vartheta \in [1, 2]$ *is the parameter in* (A.7) *and* (A.8) *used for second-order spatial-cell reconstructions.*

Proof. Since $(R\mathbf{u})_0 = 0$, one can assume without loss of generality, that $\sigma = 0$. The result then follows from a simple extension of Theorem 2.5 in [15], with a change in

the constant based on the dimension of the solver. □

Remark 7. *The result above can be easily extended to three dimensions, with CFL*

$$\Delta t \leq \left(\frac{2}{\vartheta+2}\right) \frac{\Delta x \Delta y \Delta z}{\Delta x \Delta y + \Delta y \Delta z + \Delta z \Delta x}\,, \tag{A.11}$$

where Δz is the mesh in the z direction.

Remark 8. *The result above can be easily extended to the FP_N equations (2.14). Indeed, since $(L\mathbf{u})_0 = 0$, one can assume for the purposes of the proof that $\sigma_{\mathrm{F}} = 0$. Hence the filter matrix F plays no role.*

Appendix B: Global Convergence Proof for Algorithm CR-MPC

In this appendix, we prove the global convergence property for Algorithm CR-MPC.

We first note some immediate results to be used in the proof. First, from (4.9) and (4.12), the approximate MPC search direction defined in (4.14) solves

$$\begin{bmatrix} H & -A_Q^T & 0 \\ A_Q & 0 & -I \\ 0 & S_Q & \Lambda_Q \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\lambda}_Q \\ \Delta\mathbf{s}_Q \end{bmatrix} = \begin{bmatrix} -(H\mathbf{x}+\mathbf{c}) + A_Q^T\boldsymbol{\lambda}_Q \\ \mathbf{0} \\ -S_Q\boldsymbol{\lambda}_Q + \gamma\sigma\mu_{(Q)}\mathbf{1} - \gamma\Delta S_Q^{\mathrm{a}}\Delta\boldsymbol{\lambda}_Q^{\mathrm{a}} \end{bmatrix}, \tag{B.1}$$

and the associated normal system is

$$\begin{aligned} M_{(Q)}\Delta\mathbf{x} &= -(H\mathbf{x}+\mathbf{c}) + A_Q^T S_Q^{-1}(\gamma\sigma\mu_{(Q)}\mathbf{1} - \gamma\Delta S_Q^{\mathrm{a}}\Delta\boldsymbol{\lambda}_Q^{\mathrm{a}}), \\ \Delta\mathbf{s}_Q &= A_Q\Delta\mathbf{x}, \\ \Delta\boldsymbol{\lambda}_Q &= -\boldsymbol{\lambda}_Q + S_Q^{-1}(-\Lambda_Q\Delta\mathbf{s}_Q + \gamma\sigma\mu_{(Q)}\mathbf{1} - \gamma\Delta S_Q^{\mathrm{a}}\Delta\boldsymbol{\lambda}_Q^{\mathrm{a}})\,. \end{aligned} \tag{B.2}$$

Next, for all $i \in \mathcal{I}$, let us define

$$\tilde{\lambda}_i := \begin{cases} \lambda_i + \Delta\lambda_i & i \in Q, \\ 0 & i \notin Q, \end{cases} \quad \text{and} \quad \tilde{\lambda}_i^{\mathrm{a}} := \begin{cases} \lambda_i + \Delta\lambda_i^{\mathrm{a}} & i \in Q, \\ 0 & i \notin Q. \end{cases} \tag{B.3}$$

Then, from the last equation of (B.2) and (B.3), we have

$$\tilde{\boldsymbol{\lambda}}_Q = S_Q^{-1}(-\Lambda_Q \Delta \mathbf{s}_Q + \gamma\sigma\mu_{(Q)}\mathbf{1} - \gamma\Delta S_Q^{\mathrm{a}}\Delta\boldsymbol{\lambda}_Q^{\mathrm{a}})\,. \tag{B.4}$$

Similarly, from the last equation of (4.10) and (B.3), we have

$$\tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}} = -S_Q^{-1}\Lambda_Q \Delta \mathbf{s}_Q^{\mathrm{a}}\,. \tag{B.5}$$

It follows from (B.5) and the second equation of (4.10) that

$$(\Delta\mathbf{x}^{\mathrm{a}})^T A_Q^T \tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}} = -(\Delta\mathbf{x}^{\mathrm{a}})^T A_Q^T S_Q^{-1}\Lambda_Q A_Q \Delta\mathbf{x}^{\mathrm{a}} \leq 0\,. \tag{B.6}$$

Note that, assuming $\boldsymbol{\lambda} > \mathbf{0}$ and $\mathbf{s} > \mathbf{0}$, the inequality in (B.6) holds as an equality if and only if $A_Q \Delta\mathbf{x}^{\mathrm{a}} = \mathbf{0}$.

Definition 5. $J(A, \mathbf{s}, \boldsymbol{\lambda}) := \begin{bmatrix} H & -A^T & 0 \\ A & 0 & -I \\ 0 & S & \Lambda \end{bmatrix}$.

Lemma 6 (Corresponding to Lemma B.1 in [74]). *For $\mathbf{s}, \boldsymbol{\lambda} \geq 0$, $J(A, \mathbf{s}, \boldsymbol{\lambda})$ is non-singular if and only if*

(i) $\forall i \in \mathcal{I} : s_i + \lambda_i > 0\,,$

(ii) Rows of $A_{\{i:s_i=0\}}$ are linearly independent,

(iii) $\begin{bmatrix} H & A_{\{i:\lambda_i \neq 0\}}^T \end{bmatrix}$ *has full row rank.*

Proof. See Lemma B.1 in [74]. □

Lemma 7. *Suppose* $\boldsymbol{\lambda} > \mathbf{0}$ *and* $\mathbf{s} > \mathbf{0}$. *Then*

(i) If $\Delta\mathbf{x}^a \neq \mathbf{0}$,

$$f(\mathbf{x} + \alpha\Delta\mathbf{x}^a) < f(\mathbf{x}) \,, \quad \forall\alpha \in (0, 2) \,, \tag{B.7}$$

and

$$\frac{\partial}{\partial\alpha} f(\mathbf{x} + \alpha\Delta\mathbf{x}^a) < 0 \,, \quad \forall\alpha < 1 \,. \tag{B.8}$$

(ii) Given $\zeta \leq 1$. *Suppose for some* $\hat{\theta} \in [0, 1]$,

$$f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^a + \hat{\theta}\Delta\mathbf{x}^c) \geq \zeta(f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^a)) \,, \tag{B.9}$$

then

$$f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^a + \theta\Delta\mathbf{x}^c) \geq \zeta(f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^a)) \,, \quad \forall\theta \in [0, \hat{\theta}] \,. \tag{B.10}$$

(iii) Given $\zeta \geq 0$. *Define* $\Delta\mathbf{x} := \Delta\mathbf{x}^a + \gamma\Delta\mathbf{x}^c$ *as in* (4.6), *with* γ *defined in* (4.7). *Suppose* $f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}) \geq \zeta(f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^a))$, *then*

$$f(\mathbf{x}) - f(\mathbf{x} + \alpha\Delta\mathbf{x}) \geq \frac{\zeta}{2}(f(\mathbf{x}) - f(\mathbf{x} + \alpha\Delta\mathbf{x}^a)), \quad \forall\alpha \in [0, 1]. \tag{B.11}$$

Proof. Since f is a quadratic function, it can be expressed exactly with the 2nd order Taylor expansion

$$\begin{aligned} f(\mathbf{x} + \alpha\Delta\mathbf{x}^{\mathrm{a}}) &= f(\mathbf{x}) + \alpha(\Delta\mathbf{x}^{\mathrm{a}})^T(H\mathbf{x} + \mathbf{c}) + \frac{1}{2}\alpha^2(\Delta\mathbf{x}^{\mathrm{a}})^T H\Delta\mathbf{x}^{\mathrm{a}} \\ &= f(\mathbf{x}) + \alpha(\Delta\mathbf{x}^{\mathrm{a}})^T(-H\Delta\mathbf{x}^{\mathrm{a}} + A_Q^T\tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}}) + \frac{1}{2}\alpha^2(\Delta\mathbf{x}^{\mathrm{a}})^T H\Delta\mathbf{x}^{\mathrm{a}} \quad \text{(by (4.9))} \\ &= f(\mathbf{x}) - \alpha\left(1 - \frac{1}{2}\alpha\right)(\Delta\mathbf{x}^{\mathrm{a}})^T H\Delta\mathbf{x}^{\mathrm{a}} + \alpha(\Delta\mathbf{x}^{\mathrm{a}})^T A_Q^T\tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}} \,. \end{aligned} \tag{B.12}$$

We know $(\Delta\mathbf{x}^{\mathrm{a}})^T H \Delta\mathbf{x}^{\mathrm{a}} \geq 0$ and $(\Delta\mathbf{x}^{\mathrm{a}})^T A_Q^T \tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}} \leq 0$ (by (B.6)). By Assumption 1, Condition 1 in Assumption 5, $(\Delta\mathbf{x}^{\mathrm{a}})^T H \Delta\mathbf{x}^{\mathrm{a}} = (\Delta\mathbf{x}^{\mathrm{a}})^T A_Q^T \tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}} = 0$ if and only if $\Delta\mathbf{x}^{\mathrm{a}} = \mathbf{0}$. Suppose $\Delta\mathbf{x}^{\mathrm{a}} \neq \mathbf{0}$, (B.12) gives

$$f(\mathbf{x} + \alpha\Delta\mathbf{x}^{\mathrm{a}}) - f(\mathbf{x}) < 0\,, \quad \forall \alpha \in (0, 2)\,. \tag{B.13}$$

Next, from (B.12), we have

$$\frac{\partial}{\partial\alpha} f(\mathbf{x} + \alpha\Delta\mathbf{x}^{\mathrm{a}}) = (\alpha - 1)\,(\Delta\mathbf{x}^{\mathrm{a}})^T H \Delta\mathbf{x}^{\mathrm{a}} + (\Delta\mathbf{x}^{\mathrm{a}})^T A_Q^T \tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}}\,. \tag{B.14}$$

Then we conclude from similar argument that when $\Delta\mathbf{x}^{\mathrm{a}} \neq \mathbf{0}$,

$$\frac{\partial}{\partial\alpha} f(\mathbf{x} + \alpha\Delta\mathbf{x}^{\mathrm{a}}) < 0\,, \quad \forall \alpha < 1\,. \tag{B.15}$$

From (B.13) and (B.15), Claim (i) holds.

To prove Claim (ii), we first define

$$\phi(\theta) := \zeta(f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^{\mathrm{a}})) - (f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^{\mathrm{a}} + \theta\Delta\mathbf{x}^{\mathrm{c}}))\,. \tag{B.16}$$

Note that f is quadratic with respect to θ and so is ϕ. Hence, ϕ is convex as $\partial^2\phi/\partial\theta^2 = (\Delta\mathbf{x}^{\mathrm{c}})^T H \Delta\mathbf{x}^{\mathrm{c}} \geq 0$. From the assumption (B.9), $\phi(\hat{\theta}) \leq 0$. From Claim (i), $f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^{\mathrm{a}}) \geq 0$. Since $\zeta \leq 1$, we have $\phi(0) = (\zeta - 1)(f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^{\mathrm{a}})) \leq 0$. Hence, the convexity of ϕ gives $\phi(\theta) \leq 0$ for all $\theta \in [0, \hat{\theta}]$, which leads to (B.10). Claim (ii) is now proved.

For proving Claim (iii), we expand $f(\mathbf{x} + \alpha\Delta\mathbf{x})$ with 2nd order Taylor expansion, which yields

$$f(\mathbf{x} + \alpha\Delta\mathbf{x}) = f(\mathbf{x}) + \alpha\Delta\mathbf{x}^T(H\mathbf{x} + \mathbf{c}) + \frac{1}{2}\alpha^2\Delta\mathbf{x}^T H \Delta\mathbf{x}\,, \tag{B.17}$$

and (B.12) gives

$$f(\mathbf{x}+\alpha\Delta\mathbf{x}^{\mathrm{a}}) = f(\mathbf{x}) - \alpha\left(1-\frac{1}{2}\alpha\right)(\Delta\mathbf{x}^{\mathrm{a}})^T H\Delta\mathbf{x}^{\mathrm{a}} + \alpha(\Delta\mathbf{x}^{\mathrm{a}})^T A_Q^T\tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}}\,. \tag{B.18}$$

Hence, by (B.17) and (B.18) with $\alpha = 1$, the assumption

$$f(\mathbf{x}) - f(\mathbf{x}+\Delta\mathbf{x}) \geq \zeta(f(\mathbf{x}) - f(\mathbf{x}+\Delta\mathbf{x}^{\mathrm{a}})) \tag{B.19}$$

can be written as

$$-\Delta\mathbf{x}^T(H\mathbf{x}+\mathbf{c}) - \frac{1}{2}\Delta\mathbf{x}^T H\Delta\mathbf{x} \geq \zeta\left(\frac{1}{2}(\Delta\mathbf{x}^{\mathrm{a}})^T H\Delta\mathbf{x}^{\mathrm{a}} - (\Delta\mathbf{x}^{\mathrm{a}})^T A_Q^T\tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}}\right)\,. \tag{B.20}$$

Recall that $H \geq 0$ and $(\Delta\mathbf{x}^{\mathrm{a}})^T A_Q^T\tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}} \leq 0$ (by (B.6)). Then, for $\alpha \in [0,1]$, we have

$$\begin{aligned}
f(\mathbf{x}) - f(\mathbf{x}+\alpha\Delta\mathbf{x}) &= -\alpha\Delta\mathbf{x}^T(H\mathbf{x}+\mathbf{c}) - \frac{1}{2}\alpha^2\Delta\mathbf{x}^T H\Delta\mathbf{x} \quad \text{(by (B.17))} \\
&\geq \alpha\left(-\Delta\mathbf{x}^T(H\mathbf{x}+\mathbf{c}) - \frac{1}{2}\Delta\mathbf{x}^T H\Delta\mathbf{x}\right) \\
&\geq \zeta\alpha\left(\frac{1}{2}(\Delta\mathbf{x}^{\mathrm{a}})^T H\Delta\mathbf{x}^{\mathrm{a}} - (\Delta\mathbf{x}^{\mathrm{a}})^T A_Q^T\tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}}\right) \quad \text{(by (B.20))} \\
&\geq \zeta\left(\frac{1}{2}\alpha(1-\frac{1}{2}\alpha)(\Delta\mathbf{x}^{\mathrm{a}})^T H\Delta\mathbf{x}^{\mathrm{a}} - \frac{1}{2}\alpha(\Delta\mathbf{x}^{\mathrm{a}})^T A_Q^T\tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}}\right) \\
&= \frac{\zeta}{2}(f(\mathbf{x}) - f(\mathbf{x}+\alpha\Delta\mathbf{x}^{\mathrm{a}}))\,. \quad \text{(by (B.12))}
\end{aligned} \tag{B.21}$$

Claim (iii) holds.

□

Lemma 8. *If* $\boldsymbol{\lambda} > \mathbf{0}$ *and* $\mathbf{s} > \mathbf{0}$*, then* $\Delta\mathbf{x} = \mathbf{0}$ *if and only if* $\Delta\mathbf{x}^a = \mathbf{0}$.

Proof. Let $\Delta\mathbf{x}^{\mathrm{a}} = \mathbf{0}$, from (4.7), we have $\gamma = 0$ and $\Delta\mathbf{x} = \Delta\mathbf{x}^{\mathrm{a}} + \gamma\Delta\mathbf{x}^{\mathrm{c}} = \mathbf{0}$. Conversely, suppose $\Delta\mathbf{x}^{\mathrm{a}} \neq \mathbf{0}$. From Claim (i) in Lemma 7, we have $f(\mathbf{x}) - f(\mathbf{x}+\Delta\mathbf{x}^{\mathrm{a}}) > 0$. Also, (B.9) holds for $\hat{\theta} = \gamma_1$ with γ_1 defined in (4.8). Since $\zeta < 1$, it

follows from Claim (ii) in Lemma 7 that

$$f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^{\mathrm{a}} + \theta\Delta\mathbf{x}^{\mathrm{c}}) > 0\,, \quad \forall\theta \in [0, \gamma_1]\,. \tag{B.22}$$

From (4.7), $\gamma \in [0, \gamma_1]$, thus (B.22) holds for $\theta = \gamma$, which implies $\Delta\mathbf{x} := \Delta\mathbf{x}^{\mathrm{a}} + \gamma\Delta\mathbf{x}^{\mathrm{c}} \neq \mathbf{0}$.

□

Proposition 5 (Corresponding to Proposition B.4 in [74]). *Suppose* $\boldsymbol{\lambda} > \mathbf{0}$ *and* $\mathbf{s} > \mathbf{0}$. *Let* $\Delta\mathbf{x}$ *be the search direction in* (4.6) *with algorithm parameter* $\zeta \in (0, 1)$ *in* (4.8). *Then*

$$f(\mathbf{x}) - f(\mathbf{x} + \alpha\Delta\mathbf{x}) \geq \frac{\zeta}{2}(f(\mathbf{x}) - f(\mathbf{x} + \alpha\Delta\mathbf{x}^{a}))\,, \quad \forall\alpha \in [0, 1]\,. \tag{B.23}$$

If $\Delta\mathbf{x} \neq \mathbf{0}$, *then* $\Delta\mathbf{x}$ *is a descent direction, and*

$$f(\mathbf{x} + \alpha\Delta\mathbf{x}) < f(\mathbf{x})\,, \quad \forall\alpha \in (0, 1]\,. \tag{B.24}$$

Proof. From (4.8), (B.9) holds for $\hat{\theta} = \gamma_1$. Since $\gamma \leq \gamma_1$ (see (4.7)), it follows from Claim (ii) in Lemma 7 that

$$(f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}) :=)\, f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^{\mathrm{a}} + \gamma\Delta\mathbf{x}^{\mathrm{c}}) \geq \zeta(f(\mathbf{x}) - f(\mathbf{x} + \Delta\mathbf{x}^{\mathrm{a}}))\,. \tag{B.25}$$

Then (B.23) follows from Claim (iii) in Lemma 7.

Suppose $\Delta\mathbf{x} \neq \mathbf{0}$, Lemma 8 gives $\Delta\mathbf{x}^{\mathrm{a}} \neq \mathbf{0}$. Then, applying Claim (i) in Lemma 7, we have $f(\mathbf{x}) - f(\mathbf{x} + \alpha\Delta\mathbf{x}^{\mathrm{a}}) > 0$ for $\alpha \in (0, 2)$. Hence (B.24) is a direct consequence of (B.23). □

Proposition 6 (Corresponding to Proposition B.2 in [74]). *Given* $\mathbf{s}^k > \mathbf{0}$ *and* $\boldsymbol{\lambda}^k > \mathbf{0}$*, if the iteration of Algorithm CR-MPC does not stop at Step 1, the points generated by the iteration satisfy:*

(i) $\Delta\mathbf{x}^k \neq \mathbf{0}$ *if and only if* $H\mathbf{x}^k + \mathbf{c} \neq \mathbf{0}$,

(ii) $\alpha_p^k > 0$,

(iii) $\mathbf{s}^{k+1} = A\mathbf{x}^{k+1} - \mathbf{b} > \mathbf{0}$ *and* $\mathbf{x}^{k+1} \in \mathcal{F}_P^o$,

(iv) $\boldsymbol{\lambda}^{k+1} > \mathbf{0}$.

Proof. From (4.10) and the positive definiteness of $M_{(Q)}$ (by Assumption 1, Condition 1 in Assumption 5, and Lemma 4), we have $\Delta\mathbf{x}^{\mathrm{a},k} \neq \mathbf{0}$ if and only if $H\mathbf{x}^k + \mathbf{c} \neq \mathbf{0}$. Claim (i) then follows from Lemma 8. Claims (ii) and (iii) are true due to (4.18), (4.19), and $\Delta\mathbf{s}^k = A\Delta\mathbf{x}^k$. For Claim (iv), first note that since $\mathbf{x}^k$ does not solve (P), $H\mathbf{x}^k + \mathbf{c} \neq \mathbf{0}$. From Claim (i), $\Delta\mathbf{x}^k \neq \mathbf{0}$. It follows from Lemma 8 that $\Delta\mathbf{x}^{\mathrm{a},k} \neq \mathbf{0}$. Claim (iv) is then a consequence of (4.20) and (4.21) since $\chi := \|\Delta\mathbf{x}^{\mathrm{a},k}\|^2 + \|[\tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a},k}]_-\|^2 > 0$. □

Proposition 6 shows that given a strictly feasible initial point $(\mathbf{x}^0, \mathbf{s}^0, \boldsymbol{\lambda}^0)$, Algorithm CR-MPC generates a well-defined sequence of points that are strictly feasible. In addition, Claim (i) in Proposition 6 implies that, the sequence produced by Algorithm CR-MPC is either infinite, or terminated at some iteration k such that the primal point $\mathbf{x}^k$ satisfies $H\mathbf{x}^k + \mathbf{c} = \mathbf{0}$, i.e., $\mathbf{x}^k$ is the solution to the unconstrained version of the primal CQP (P). In the latter case, since $\mathbf{x}^k$ is strictly primal feasible, it also solves (P), and $\boldsymbol{\lambda} = \mathbf{0}$ is the associated multiplier vector. Hence, in the

remainder of this appendix (unless otherwise indicated), we only focus on the case that Algorithm CR-MPC generates an infinite sequence of points.

Lemma 9. *Given an infinite index set K, $\{\Delta\mathbf{x}^k\} \to \mathbf{0}$ as $k \to \infty$, $k \in K$ if and only if $\{\Delta\mathbf{x}^{a,k}\} \to \mathbf{0}$ as $k \to \infty$, $k \in K$.*

Proof. From (4.14) and (4.7), given algorithm parameter $\tau \in [0, 1)$, we have $\|\Delta\mathbf{x}^k - \Delta\mathbf{x}^{\mathrm{a},k}\| = \|\gamma\Delta\mathbf{x}^{\mathrm{c},k}\| \leq \tau\|\Delta\mathbf{x}^{\mathrm{a},k}\|$, which implies $(1-\tau)\|\Delta\mathbf{x}^{\mathrm{a},k}\| \leq \|\Delta\mathbf{x}^k\| \leq (1+\tau)\|\Delta\mathbf{x}^{\mathrm{a},k}\|$ for all $k \in K$. Hence, $\{\Delta\mathbf{x}^k\} \to \mathbf{0}$ on K implies $\{\Delta\mathbf{x}^{\mathrm{a},k}\} \to \mathbf{0}$ on K, and vice versa. □

Lemma 10 (Corresponding to Corollary B.5 in [74])**.** *The sequence $\{\mathbf{x}^k\}$ is bounded.*

Proof. See Corollary B.5 in [74]. □

Lemma 11 (Corresponding to Lemma B.7 in [74])**.** *Suppose $\{\mathbf{x}^k\}$ converges to some limit point $\mathbf{x}^*$ on an infinite index set K. If $\{\Delta\mathbf{x}^{a,k}\}$ converges to zero on K, then (i) $\mathbf{x}^*$ is stationary, (ii) $\{\tilde{\boldsymbol{\lambda}}^{a,k}\}$ converges on K to $\boldsymbol{\lambda}^*$, the unique multiplier associated with $\mathbf{x}^*$, and (iii) $\{\tilde{\boldsymbol{\lambda}}^k\}$ also converges to $\boldsymbol{\lambda}^*$ on K.*

Proof. Suppose $\{\mathbf{x}^k\} \to \mathbf{x}^*$ on K and $\{\Delta\mathbf{x}^{\mathrm{a},k}\} \to \mathbf{0}$ as $k \to \infty$, $k \in K$. Since $\mathbf{s}^k := A\mathbf{x}^k - \mathbf{b} > \mathbf{0}$ for all $k \in K$, we have $\{\mathbf{s}^k\} \to \mathbf{s}^*$ on K, where $\mathbf{s}^* := A\mathbf{x}^* - \mathbf{b} \geq \mathbf{0}$. Define $\delta := \frac{1}{2}\min_{i \notin \mathcal{A}(\mathbf{x}^*)} s_i^* > 0$. Then, it follows from $\{\mathbf{x}^k\} \to \mathbf{x}^*$ on K that, there exists $k' > 0$ such that, for all $k > k'$, $k \in K$,

$$s_i^k > \delta, \quad \forall i \notin \mathcal{A}(\mathbf{x}^*)\,. \tag{B.26}$$

From (B.3) and (4.10), we have, for all $k \in K$,

$$\tilde{\lambda}_i^{\mathrm{a},k} = 0\,,\ \forall i \notin Q^k \quad \text{and} \quad \tilde{\lambda}_i^{\mathrm{a},k} = -(s_i^k)^{-1}\lambda_i^k \Delta s_i^{\mathrm{a},k}\,,\ \forall i \in Q^k\,. \tag{B.27}$$

For all k and all $i \in \mathcal{I}$, we have $\lambda_i^k < \lambda^{\max}$ from (4.20) and (4.21). Hence, from (B.26) and (B.27), we conclude that

$$|\tilde{\lambda}_i^{\mathrm{a},k}| \le c|\Delta s_i^{\mathrm{a},k}|\,, \quad \forall i \notin \mathcal{A}(\mathbf{x}^*),\ \forall k > k',\ k \in K\,, \tag{B.28}$$

where the constant $c := \delta^{-1}\lambda^{\max}$. Since $\Delta\mathbf{s}^{\mathrm{a},k} := A\Delta\mathbf{x}^{\mathrm{a},k}$, the assumption $\{\Delta\mathbf{x}^{\mathrm{a},k}\} \to \mathbf{0}$ on K implies $\{\Delta\mathbf{s}^{\mathrm{a},k}\} \to \mathbf{0}$ on K. Hence, it follows from (B.28) that, for all $i \notin \mathcal{A}(\mathbf{x}^*)$,

$$\{\tilde{\lambda}_i^{\mathrm{a},k}\} \to 0\,, \quad \text{as} \quad k \to \infty,\ k \in K\,. \tag{B.29}$$

To show the convergence of $\{\tilde{\lambda}_i^{\mathrm{a},k}\}$ for $i \in \mathcal{A}(\mathbf{x}^*)$, at iteration k, we write the first equation of (4.9) as

$$H\mathbf{x}^k + \mathbf{c} - A_{Q^k}^T \tilde{\boldsymbol{\lambda}}_{Q^k}^{\mathrm{a},k} = -H\Delta\mathbf{x}^{\mathrm{a},k}\,. \tag{B.30}$$

Note that the set Q^k can be split into two sets $Q^k \cap \mathcal{A}(\mathbf{x}^*)$ and $Q^k \setminus \mathcal{A}(\mathbf{x}^*)$. (B.30) can then be written as

$$H\mathbf{x}^k + \mathbf{c} - A_{Q^k\cap\mathcal{A}(\mathbf{x}^*)}^T \tilde{\boldsymbol{\lambda}}_{Q^k\cap\mathcal{A}(\mathbf{x}^*)}^{\mathrm{a},k} = -H\Delta\mathbf{x}^{\mathrm{a},k} + A_{Q^k\setminus\mathcal{A}(\mathbf{x}^*)}^T \tilde{\boldsymbol{\lambda}}_{Q^k\setminus\mathcal{A}(\mathbf{x}^*)}^{\mathrm{a},k}\,. \tag{B.31}$$

From (B.27), $\tilde{\boldsymbol{\lambda}}_{\mathcal{A}(\mathbf{x}^*)\setminus Q^k}^{\mathrm{a},k} = \mathbf{0}$. Thus, subtracting $A_{\mathcal{A}(\mathbf{x}^*)\setminus Q^k}^T \tilde{\boldsymbol{\lambda}}_{\mathcal{A}(\mathbf{x}^*)\setminus Q^k}^{\mathrm{a},k} = \mathbf{0}$ on both sides of (B.31) yields

$$H\mathbf{x}^k + \mathbf{c} - A_{\mathcal{A}(\mathbf{x}^*)}^T \tilde{\boldsymbol{\lambda}}_{\mathcal{A}(\mathbf{x}^*)}^{\mathrm{a},k} = -H\Delta\mathbf{x}^{\mathrm{a},k} + A_{Q^k\setminus\mathcal{A}(\mathbf{x}^*)}^T \tilde{\boldsymbol{\lambda}}_{Q^k\setminus\mathcal{A}(\mathbf{x}^*)}^{\mathrm{a},k}\,. \tag{B.32}$$

From the assumption, $\{\Delta\mathbf{x}^{\mathrm{a},k}\} \to \mathbf{0}$ on K, and (B.29), we then have

$$H\mathbf{x}^k + \mathbf{c} - A_{\mathcal{A}(\mathbf{x}^*)}^T \tilde{\boldsymbol{\lambda}}_{\mathcal{A}(\mathbf{x}^*)}^{\mathrm{a},k} \to \mathbf{0}\,, \text{ as } k \to \infty\,,\ k \in K\,. \tag{B.33}$$

By Assumption 4, the rows of $A_{\mathcal{A}(\mathbf{x}^*)}$ are linearly independent. Hence, from (B.29) and (B.33), $\{\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k}\}$ converges to a unique $\boldsymbol{\lambda}^*$ on K.

Next, we show that $\{\tilde{\lambda}_i^k\} \to 0$ on K for all $i \notin \mathcal{A}(\mathbf{x}^*)$. From (B.3) and (B.2), we have, for all k,

$$\tilde{\lambda}_i^k = 0\,,\ \forall i \notin Q^k\,, \tag{B.34}$$

and

$$\tilde{\lambda}_i^k = (s_i^k)^{-1}(-\lambda_i^k \Delta s_i^k + \gamma^k \sigma^k \mu_{(Q^k)}^{(k)} - \gamma^k \Delta s_i^{\mathrm{a},k} \Delta \lambda_i^{\mathrm{a},k})\,, \quad \forall i \in Q^k\,. \tag{B.35}$$

From (4.7), $|\gamma^k \sigma^k \mu_{(Q^k)}^{(k)}| \le \tau \|\Delta \mathbf{x}^{\mathrm{a},k}\|$ for all k. For $i \in Q^k \setminus \mathcal{A}(\mathbf{x}^*)$, the boundedness of $\{\lambda_i^k\}$ and $\{\tilde{\lambda}_i^{\mathrm{a},k}\}$ (by (B.29)) implies the boundedness of $\{\Delta\lambda_i^{\mathrm{a},k}\}$ (see (B.3)), i.e., there exists some constant $\Delta\lambda^{\max} > 0$ such that $|\Delta\lambda_i^{\mathrm{a},k}| < \Delta\lambda^{\max}$ for all $k \in K$. Further, $\gamma^k \in [0,1]$ for all k by (4.7) and (4.8). Then it follows from (B.34), (B.35), and (B.26) that, for all $i \notin \mathcal{A}(\mathbf{x}^*)$,

$$|\tilde{\lambda}_i^k| \le c_1 |\Delta s_i^k| + c_2 \|\Delta \mathbf{x}^{\mathrm{a},k}\| + c_3 |\Delta s_i^{\mathrm{a},k}|\,, \quad \forall k > k',\, k \in K\,, \tag{B.36}$$

where the constants $c_1 = \delta^{-1}\lambda^{\max}$, $c_2 = \delta^{-1}\tau$, and $c_3 = \delta^{-1}\Delta\lambda^{\max}$. From Lemma 9 and the assumption $\{\Delta \mathbf{x}^{\mathrm{a},k}\} \to \mathbf{0}$ on K, we have

$$\{\Delta \mathbf{x}^k\} \to \mathbf{0}\,,\ \text{as } k \to \infty,\, k \in K\,, \tag{B.37}$$

which implies $\{\Delta \mathbf{s}^k\} := \{A\Delta \mathbf{x}^k\} \to \mathbf{0}$ on K. Hence, it follows from (B.36) that, for all $i \notin \mathcal{A}(\mathbf{x}^*)$,

$$\{\tilde{\lambda}_i^k\} \to 0\,, \quad \text{as} \quad k \to \infty,\, k \in K\,. \tag{B.38}$$

To show the convergence of $\{\tilde{\lambda}_i^k\}$ for $i \in \mathcal{A}(\mathbf{x}^*)$, at iteration k, the first equation of

(B.1) gives

$$H\mathbf{x}^k + \mathbf{c} - A_{Q^k}^T \tilde{\boldsymbol{\lambda}}_{Q^k}^k = -H\Delta\mathbf{x}^k \,. \tag{B.39}$$

Again, by splitting Q^k into $Q^k \cap \mathcal{A}(\mathbf{x}^*)$ and $Q^k \setminus \mathcal{A}(\mathbf{x}^*)$, we have

$$H\mathbf{x}^k + \mathbf{c} - A_{Q^k\cap\mathcal{A}(\mathbf{x}^*)}^T \tilde{\boldsymbol{\lambda}}_{Q^k\cap\mathcal{A}(\mathbf{x}^*)}^k = -H\Delta\mathbf{x}^{\mathrm{a},k} + A_{Q^k\setminus\mathcal{A}(\mathbf{x}^*)}^T \tilde{\boldsymbol{\lambda}}_{Q^k\setminus\mathcal{A}(\mathbf{x}^*)}^k \,. \tag{B.40}$$

From (B.34), $\tilde{\boldsymbol{\lambda}}_{\mathcal{A}(\mathbf{x}^*)\setminus Q^k}^k = \mathbf{0}$. Thus, subtracting $A_{\mathcal{A}(\mathbf{x}^*)\setminus Q^k}^T \tilde{\boldsymbol{\lambda}}_{\mathcal{A}(\mathbf{x}^*)\setminus Q^k}^k = \mathbf{0}$ on both sides of (B.40) yields

$$H\mathbf{x}^k + \mathbf{c} - A_{\mathcal{A}(\mathbf{x}^*)}^T \tilde{\boldsymbol{\lambda}}_{\mathcal{A}(\mathbf{x}^*)}^k = -H\Delta\mathbf{x}^k + A_{Q^k\setminus\mathcal{A}(\mathbf{x}^*)}^T \tilde{\boldsymbol{\lambda}}_{Q^k\setminus\mathcal{A}(\mathbf{x}^*)}^k \,. \tag{B.41}$$

It then follows from (B.37) and (B.38) that

$$H\mathbf{x}^k + \mathbf{c} - A_{\mathcal{A}(\mathbf{x}^*)}^T \tilde{\boldsymbol{\lambda}}_{\mathcal{A}(\mathbf{x}^*)}^k \to \mathbf{0} \,, \text{ as } k \to \infty \,, \, k \in K \,. \tag{B.42}$$

By Assumption 4, the rows of $A_{\mathcal{A}(\mathbf{x}^*)}$ are linearly independent. Hence, from (B.38) and (B.42), $\{\tilde{\boldsymbol{\lambda}}^k\}$ also converges to the same unique $\boldsymbol{\lambda}^*$ on K (see (B.29) and (B.33)). In addition, by taking limits in (B.29) and (B.33) (or (B.38) and (B.42)), we obtain

$$\begin{aligned} H\mathbf{x}^* + \mathbf{c} - A^T\boldsymbol{\lambda}^* &= \mathbf{0} \,, \\ S^*\boldsymbol{\lambda}^* &= \mathbf{0} \,, \end{aligned} \tag{B.43}$$

which implies $\mathbf{x}^*$ is stationary and $\boldsymbol{\lambda}^*$ is the unique associated multiplier. □

Lemma 12 (Corresponding to Lemma B.8 in [74]). *Let K be an infinite index set such that*

$$\inf\{\|\Delta\mathbf{x}^{a,k-1}\|^2 + \|[\tilde{\boldsymbol{\lambda}}_{Q^k}^{a,k-1}]_-\|^2 : k \in K\} > 0 \,. \tag{B.44}$$

Then $\{\Delta\mathbf{x}^k\} \to \mathbf{0}$ as $k \to \infty$, $k \in K$.

Proof. By contradiction. Suppose $\{\Delta\mathbf{x}^k\} \not\to \mathbf{0}$ as $k \to \infty$, $k \in K$. Then, from Lemma 9, $\{\Delta\mathbf{x}^{\mathrm{a},k}\} \not\to \mathbf{0}$ as $k \to \infty$, $k \in K$. On the other hand, by (B.44), $\inf\{\chi^{k-1} : k \in K\} := \inf\{\|\Delta\mathbf{x}^{\mathrm{a},k-1}\|^2 + \|[\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k-1}_{Q^k}]_-\|^2 : k \in K\} > 0$. Thus, from (4.20) and (4.21), $\{\lambda_i^k\}$ is bounded away from zero on K for all $i \in \mathcal{I}$. Since $\{\mathbf{x}^k\}$ is bounded (see Lemma 10), and $\{\boldsymbol{\lambda}^k\}$ is bounded (by construction), there exist $\mathbf{x}^*$, $\boldsymbol{\lambda}^* > \mathbf{0}$, an index set $Q^* \subseteq \mathcal{I}$, and some infinite index set $K^* \subseteq K$ such that

$$\inf_{k \in K^*} \|\Delta\mathbf{x}^{\mathrm{a},k}\| > 0\,, \tag{B.45}$$

$$\{\mathbf{x}^k\} \to \mathbf{x}^* \text{ as } k \to \infty\,,\ k \in K^*\,, \tag{B.46}$$

$$\{\boldsymbol{\lambda}^k\} \to \boldsymbol{\lambda}^* > \mathbf{0} \text{ as } k \to \infty\,,\ k \in K^*\,, \tag{B.47}$$

$$Q^k = Q^*\,,\ \forall k \in K^*\,. \tag{B.48}$$

Since $\mathbf{s}^k := A\mathbf{x}^k - \mathbf{b} > \mathbf{0}$ and $\{\mathbf{x}^k\} \to \mathbf{x}^*$ on K^*, we have

$$\{\mathbf{s}^k\} \to \mathbf{s}^* := \{A\mathbf{x}^* - \mathbf{b}\} \geq \mathbf{0} \text{ as } k \to \infty\,,\ k \in K^*\,. \tag{B.49}$$

By Lemma 6, $J(A_{Q^*}, \mathbf{s}^*_{Q^*}, \boldsymbol{\lambda}^*_{Q^*})$ is nonsingular given $\boldsymbol{\lambda}^* > \mathbf{0}$, $\mathbf{s}^* \geq \mathbf{0}$, Assumption 4 and Condition 1 in Assumption 5. On the other hand, by continuity, $J(A_{Q^*}, \mathbf{s}^k_{Q^*}, \boldsymbol{\lambda}^k_{Q^*})$ tends to $J(A_{Q^*}, \mathbf{s}^*_{Q^*}, \boldsymbol{\lambda}^*_{Q^*})$ as $k \to \infty$, $k \in K^*$, so that, from (4.9) and (B.1), there exists $\Delta\mathbf{x}^{\mathrm{a},*}$, $\Delta\mathbf{x}^*$, $\tilde{\boldsymbol{\lambda}}^{\mathrm{a},*}_{Q^*}$, $\tilde{\boldsymbol{\lambda}}^*_{Q^*}$ such that

$$\{\Delta\mathbf{x}^{\mathrm{a},k}\} \to \Delta\mathbf{x}^{\mathrm{a},*} \text{ as } k \to \infty\,,\ k \in K^*\,, \tag{B.50}$$

$$\{\Delta \mathbf{x}^k\} \to \Delta \mathbf{x}^* \text{ as } k \to \infty\,,\ k \in K^*\,, \tag{B.51}$$

$$\{\Delta \mathbf{s}^k\} \to \Delta \mathbf{s}^* := A\Delta \mathbf{x}^* \text{ as } k \to \infty\,,\ k \in K^*\,, \tag{B.52}$$

$$\{\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k}_{Q^*}\} := \{\boldsymbol{\lambda}^k_{Q^*} + \Delta\boldsymbol{\lambda}^{\mathrm{a},k}_{Q^*}\} \to \tilde{\boldsymbol{\lambda}}^{\mathrm{a},*}_{Q^*} \text{ as } k \to \infty\,,\ k \in K^*\,, \tag{B.53}$$

$$\{\tilde{\boldsymbol{\lambda}}^{k}_{Q^*}\} := \{\boldsymbol{\lambda}^k_{Q^*} + \Delta\boldsymbol{\lambda}^{k}_{Q^*}\} \to \tilde{\boldsymbol{\lambda}}^{*}_{Q^*} \text{ as } k \to \infty\,,\ k \in K^*\,, \tag{B.54}$$

where the last two equations also imply that $\{\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k}_{Q^*}\}$ and $\{\tilde{\boldsymbol{\lambda}}^{k}_{Q^*}\}$ are bounded on K^*. Furthermore, in view of (B.45), $\Delta\mathbf{x}^{\mathrm{a},*} \neq \mathbf{0}$. Thus, $\mathbf{x}^* \notin \mathcal{F}^*_P$, and we have $\mathcal{A}(\mathbf{x}^*) \subseteq Q^*$ from Condition 2 in Assumption 5. With these facts, we next show that $f(\mathbf{x}^*) \to -\infty$ as $k \to \infty$ on K^*, which contradicts Lemma 10.

First, let us define

$$\hat{\boldsymbol{\lambda}}^{k} := -(S^k)^{-1}\Lambda^k \Delta\mathbf{s}^k\,, \quad \forall k\,, \tag{B.55}$$

so that, for all $i \in \mathcal{I}$ and all $k \in K^*$, $\frac{\lambda^k_i}{\hat{\lambda}^k_i} = -\frac{s^k_i}{\Delta s^k_i}$, and $\hat{\lambda}^k_i > 0$ if and only if $\Delta s^k_i < 0$. Then, since $\mathbf{s}^k > \mathbf{0}$ and $\boldsymbol{\lambda}^k > \mathbf{0}$ for all $k \in K^*$, the primal step size α^k_{p} defined in (4.18) can be rewritten as

$$\bar{\alpha}^k_{\mathrm{p}} := \begin{cases} \infty & \text{if } \hat{\boldsymbol{\lambda}}^{k} \leq \mathbf{0}\,, \\ \min_i \left\{ \frac{\lambda^k_i}{\hat{\lambda}^k_i} : \hat{\lambda}^k_i > 0,\ i \in \mathcal{I} \right\} & \text{otherwise.} \end{cases} \tag{B.56}$$

$$\alpha^k_{\mathrm{p}} := \min\left\{1,\, \max\{\varkappa\bar{\alpha}_{\mathrm{p}},\, \bar{\alpha}_{\mathrm{p}} - \|\Delta\mathbf{x}^k\|\}\right\}\,.$$

We then show that $\{\hat{\boldsymbol{\lambda}}^k\}$ is bounded above componentwise on K^*. Note that, for $i \in \mathcal{I} \setminus Q^*$, $k \in K^*$, $\{s_i^k\}$ is bounded away from zero since $\mathcal{A}(\mathbf{x}^*) \subseteq Q^*$. Hence, it follows from (B.47),(B.49), and (B.52) that $\{\hat{\boldsymbol{\lambda}}^k_{I\setminus Q^*}\}$ is convergent, thus bounded, on K^*. On the other hand, subtracting (B.55) from (B.4) yields that, for all $k \in K^*$,

$$\hat{\boldsymbol{\lambda}}^k_{Q^*} = \tilde{\boldsymbol{\lambda}}^k_{Q^*} - \gamma^k \sigma^k \mu^k_{(Q^*)} (S^k_{Q^*})^{-1}\mathbf{1} + \gamma^k (S^k_{Q^*})^{-1} \Delta S^{\mathrm{a},k}_{Q^*} \Delta \boldsymbol{\lambda}^{\mathrm{a},k}_{Q^*} . \tag{B.57}$$

From (B.54), $\{\tilde{\boldsymbol{\lambda}}^k_{Q^*}\}$ is bounded on K^*. By definitions, $\gamma^k \geq 0$, $\sigma^k \geq 0$, $\mu^k_{(Q^*)} \geq 0$ and $s^k_i > 0$, for all k and all $i \in \mathcal{I}$. Hence, we have $\gamma^k \sigma^k \mu^k_{(Q^*)} (S^k_{Q^*})^{-1}\mathbf{1} \geq \mathbf{0}$ for all k. Also, from (B.47) and (B.53), $\{\Delta \boldsymbol{\lambda}^{\mathrm{a},k}_{Q^*}\}$ is bounded on K^*. Let $\tilde{\Lambda}^{\mathrm{a},k}_{Q^*} = \mathrm{diag}(\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k}_{Q^*})$, for any $k \in K^*$ and $Q \in \mathcal{I}$, the last equation in (4.10) gives $(S^k_Q)^{-1} \Delta S^{\mathrm{a},k}_Q = (\Lambda^k_Q)^{-1} \tilde{\Lambda}^{\mathrm{a},k}_Q$, so that we have

$$\gamma^k (S^k_{Q^*})^{-1} \Delta S^{\mathrm{a},k}_{Q^*} \Delta \boldsymbol{\lambda}^{\mathrm{a},k}_{Q^*} = \gamma^k (\Lambda^k_{Q^*})^{-1} \tilde{\Lambda}^{\mathrm{a},k}_{Q^*} \Delta \boldsymbol{\lambda}^{\mathrm{a},k}_{Q^*} , \quad \forall k \in K^* . \tag{B.58}$$

Hence the sequence $\{\gamma^k (S^k_{Q^*})^{-1} \Delta S^{\mathrm{a},k}_{Q^*} \Delta \boldsymbol{\lambda}^{\mathrm{a},k}_{Q^*}\}$ is bounded on K^* since $\{\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k}_{Q^*}\}$ (by (B.53)) and $\{\Delta \boldsymbol{\lambda}^{\mathrm{a},k}_{Q^*}\}$ are both bounded on K^* and $\{\boldsymbol{\lambda}^k_{Q^*}\}$ is bounded away from zero on K^* (by (B.44) and (4.20)). Thus, each term on the right-hand side of (B.57) is either bounded or bounded above on K^*. Therefore, $\{\hat{\boldsymbol{\lambda}}^k_{Q^*}\}$ is bounded above on K^*. We then conclude $\{\hat{\boldsymbol{\lambda}}^k\}$ is bounded above on K^*.

For all $i \in \mathcal{I}$, we have shown that, on K^*, $\{\hat{\lambda}^k_i\}$ is bounded above and $\{\lambda^k_i\}$ tends to a positive limit λ^*_i (from (B.47)). It then follows from (B.56) that $\bar{\alpha}^k_{\mathrm{p}}$ is also bounded away from zero on K^*, and there exists $\underline{\alpha} > 0$ such that $\alpha^k_{\mathrm{p}} > \underline{\alpha}$, for all $k \in K^*$.

For all $k \in K^*$, since $\Delta \mathbf{x}^{\mathrm{a},k} \neq \mathbf{0}$ (by (B.45)) and $\alpha^k_{\mathrm{p}} \in (\underline{\alpha}, 1]$, Claim (i) in

Lemma 7 implies

$$f(\mathbf{x}^k + \alpha_{\mathrm{p}}^k \Delta \mathbf{x}^{\mathrm{a},k}) < f(\mathbf{x}^k + \underline{\alpha} \Delta \mathbf{x}^{\mathrm{a},k}) \,. \tag{B.59}$$

Expanding $f(\mathbf{x}^k + \underline{\alpha} \Delta \mathbf{x}^{\mathrm{a},k})$ with 2nd order Taylor expansion (see (B.12)) gives

$$f(\mathbf{x}^k + \underline{\alpha} \Delta \mathbf{x}^{\mathrm{a},k}) = f(\mathbf{x}^k) - \underline{\alpha}(1 - \frac{1}{2}\underline{\alpha})(\Delta \mathbf{x}^{\mathrm{a},k})^T H \Delta \mathbf{x}^{\mathrm{a},k} + \underline{\alpha}(\tilde{\boldsymbol{\lambda}}_{Q^*}^{\mathrm{a},k})^T A_{Q^*} \Delta \mathbf{x}^{\mathrm{a},k} \,. \tag{B.60}$$

Taking limits on the last two terms on the right-hand-side of (B.60) by (B.50) and (B.53) yields $-\underline{\alpha}(1 - \frac{1}{2}\underline{\alpha})(\Delta \mathbf{x}^{\mathrm{a},*})^T H \Delta \mathbf{x}^{\mathrm{a},*} + \underline{\alpha}(\tilde{\boldsymbol{\lambda}}_{Q^*}^{\mathrm{a},*})^T A_{Q^*} \Delta \mathbf{x}^{\mathrm{a},*}$, which is strictly negative due to $H \geq 0$, (B.6), Assumption 1, and Condition 1 in Assumption 5. Hence, from (B.59) and (B.60), there exist $k' > 0$ and $\delta > 0$ such that

$$f(\mathbf{x}^k + \alpha_{\mathrm{p}}^k \Delta \mathbf{x}^{\mathrm{a},k}) < f(\mathbf{x}^k) - \delta \,, \quad \forall k > k' \,, \, k \in K^* \,. \tag{B.61}$$

On the other hand, it follows from Proposition 5 that

$$f(\mathbf{x}^k) - f(\mathbf{x}^k + \alpha_{\mathrm{p}}^k \Delta \mathbf{x}^k) \geq \frac{\zeta}{2}(f(\mathbf{x}^k) - f(\mathbf{x}^k + \alpha_{\mathrm{p}}^k \Delta \mathbf{x}^{\mathrm{a},k})) \,, \tag{B.62}$$

where $\zeta \in (0, 1)$ is an algorithm parameter. Hence, for $k > k'$, $k \in K^*$,

$$f(\mathbf{x}^{k+1}) = f(\mathbf{x}^k + \alpha_{\mathrm{p}}^k \Delta \mathbf{x}^k) < f(\mathbf{x}^k) - \frac{\zeta}{2}\delta \,. \tag{B.63}$$

By Proposition 5, $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$ for all k. Hence, we conclude from (B.63) that $f(\mathbf{x}^k) \to -\infty$ as $k \to \infty$, which contradicts the boundedness of $\{\mathbf{x}^k\}$ given in Lemma 10. □

Lemma 13 (Corresponding to Lemma B.9 in [74])**.** *Suppose $\{\mathbf{x}^k\}$ is bounded away from $\mathcal{F}_P^*$ on some infinite index set K. Then $\{\Delta \mathbf{x}^k\} \to \mathbf{0}$ as $k \to \infty$, $k \in K$.*

Proof. By contradiction. Suppose $\{\Delta \mathbf{x}^k\} \not\to \mathbf{0}$ as $k \to \infty$, $k \in K$. It follows from Lemma 12 that, there exists an infinite index set $K^* \subseteq K$ such that $\{\Delta \mathbf{x}^{\mathrm{a},k-1}\} \to \mathbf{0}$

and $\{[\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k-1}_{Q^k}]_-\} \to \mathbf{0}$ as $k \to \infty$, $k \in K^*$. It is known that $\{\mathbf{x}^k\}$ is bounded and bounded away from $\mathcal{F}_P^*$ on K from Lemma 10 and the assumption, respectively. Hence, without loss of generality, we can assume that there exists an index set $Q^* \subseteq \mathcal{I}$, and some point $\mathbf{x}^* \notin \mathcal{F}_P^*$ such that

$$\begin{aligned} &\{\mathbf{x}^k\} \to \mathbf{x}^*\,, \text{ as } k \to \infty,\ k \in K^*\,, \\ &\{\Delta\mathbf{x}^{\mathrm{a},k-1}\} \to \mathbf{0}\,, \text{ as } k \to \infty,\ k \in K^*\,, \\ &\{[\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k-1}_{Q^k}]_-\} \to \mathbf{0}\,, \text{ as } k \to \infty,\ k \in K^*\,, \\ &Q^k = Q^*\,, \forall k \in K^*\,. \end{aligned} \tag{B.64}$$

From (4.19), (4.14) and (4.7),

$$\|\mathbf{x}^k - \mathbf{x}^{k-1}\| = \|\alpha_{\mathrm{p}}^{k-1}\Delta\mathbf{x}^{k-1}\| \le \|\Delta\mathbf{x}^{k-1}\| \le (1+\tau)\|\Delta\mathbf{x}^{\mathrm{a},k-1}\|\,, \tag{B.65}$$

which implies $\{\mathbf{x}^{k-1}\} \to \mathbf{x}^*$ as $k \to \infty$, $k \in K^*$. It then follows from Lemma 11 that $\mathbf{x}^*$ is stationary and $\{\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k-1}\}$ converges to $\boldsymbol{\lambda}^*$, the associated multiplier to $\mathbf{x}^*$, as $k \to \infty$, $k \in K^*$. From (B.64), $\lambda_i^* \ge 0$ for all $i \in Q^*$. Since $\mathbf{x}^*$ is stationary, we have $S^*\boldsymbol{\lambda}^* = 0$, where $S^* = \mathrm{diag}(\mathbf{s}^*)$ and $\mathbf{s}^* := A\mathbf{x}^* - \mathbf{b}$. For $i \notin Q^*$, by Condition 2 in Assumption 5, $i \notin \mathcal{A}(\mathbf{x}^*)$ thus $s_i^* > 0$. Hence, $\lambda_i^* = 0$ for all $i \notin Q^*$. We proved $\boldsymbol{\lambda}^* \ge \mathbf{0}$, which, together with the stationarity of $\mathbf{x}^*$, implies that $\mathbf{x}^* \in \mathcal{F}_P^*$ and contradicts to the assumption. □

Proposition 7 (Corresponding to Proposition B.10 in [74])**.** $\{\mathbf{x}^k\}$ *approaches the set of stationary points of* (P)*, i.e., for any* $\epsilon > 0$*, there exists* k' *such that, for all* $k > k'$*, there is some stationary point* $\hat{\mathbf{x}}^k$ *that is* ϵ*-close to* $\mathbf{x}^k$*.*

Proof. By contradiction. Suppose not. From Lemma 10, $\{\mathbf{x}^k\}$ is bounded. Hence, $\{\mathbf{x}^k\}$ converges to some non-stationary point $\mathbf{x}^*$ on some infinite index set K. By

Lemma 11, $\{\Delta \mathbf{x}^{\mathrm{a},k}\}$ does not converge to zero on K. Lemma 8 tells that $\{\Delta \mathbf{x}^k\}$ does not converge to zero on K as well. Thus, by Lemma 13, for every $\epsilon > 0$, there exists some $k' \in K$ such that $\mathrm{dist}(\mathbf{x}^{k'}, \mathcal{F}_P^*) < \epsilon$, which implies that $\mathbf{x}^* \in \mathcal{F}_P^*$, and contradicts to the non-stationary assumption. □

Lemma 14 (Corresponding to Lemma B.14 in [74])**.** *Suppose* $\{\mathbf{x}^k\}$ *is bounded away from* $\mathcal{F}_P^*$*. Let* $\mathbf{x}^*$ *and* $\mathbf{x}^{*\prime}$ *are limit points of* $\{\mathbf{x}^k\}$*. Let* $\boldsymbol{\lambda}$ *and* $\boldsymbol{\lambda}'$ *be the associated multipliers to* $\mathbf{x}^*$ *and* $\mathbf{x}^{*\prime}$*. Then* $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$*.*

Proof. See Lemma B.14 in [74]. □

Now we are ready to prove Theorem 3.

Proof of Theorem 3. By contradiction. Suppose $\{\mathbf{x}^k\}$ does not converge to $\mathcal{F}_P^*$. Then, since $\{\mathbf{x}^k\}$ is bounded, it has at least one limit point $\hat{\mathbf{x}}$ that is not in $\mathcal{F}_P^*$. From Lemma 10 and Proposition 5, $\{f(\mathbf{x}^k)\}$ is a bounded, monotonically decreasing sequence. Hence, $f(\hat{\mathbf{x}}) = \inf_k f(\mathbf{x}^k)$ and $\{\mathbf{x}^k\}$ is bounded away from $\mathcal{F}_P^*$. Lemma 13 then gives $\{\Delta \mathbf{x}^k\} \to \mathbf{0}$ as $k \to \infty$, and $\{\Delta \mathbf{x}^{\mathrm{a},k}\}$ also converges to zero by Lemma 8. Thus, there exists $\mathbf{x}^*$ such that $\{\mathbf{x}^k\} \to \mathbf{x}^*$ as $k \to \infty$, and $\mathbf{x}^* \notin \mathcal{F}_P^*$ by assumption. From Lemma 14, there exists a common multiplier $\boldsymbol{\lambda}^*$ associated to all limit points of $\{\mathbf{x}^k\}$. It follows from Lemma 11 that $\mathbf{x}^*$ is stationary, $\{\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k}\} \to \boldsymbol{\lambda}^*$, and $\{\tilde{\boldsymbol{\lambda}}^{k}\} \to \boldsymbol{\lambda}^*$. Since $\mathbf{x}^*$ is stationary and $\mathbf{x}^* \notin \mathcal{F}_P^*$, there exists $i_0 \in \mathcal{I}$ such that $\lambda_{i_0}^* < 0$. Thus there exists $\hat{k} > 0$ such that

$$\tilde{\lambda}_{i_0}^{\mathrm{a},k} < 0\,, \text{ and } \tilde{\lambda}_{i_0}^{k} < 0\,, \quad \forall k > \hat{k}\,. \tag{B.66}$$

Also, since $\lambda_{i_0}^* < 0$, the stationarity of $\mathbf{x}^*$ implies $\{s_{i_0}^k\} \to s_{i_0}^* = 0$, i.e., $i_0 \in \mathcal{A}(\mathbf{x}^*)$, so that $s_{i_0}^* \lambda_{i_0}^* = 0$.

On the other hand, it is known from Condition 2 in Assumption 5 that there exists $k' > \hat{k}$ such that $i_0 \in \mathcal{A}(\mathbf{x}^*) \subseteq Q^k$ for all $k > k'$. Then (B.5) gives

$$\Delta s_{i_0}^{\mathrm{a},k} = -(\lambda_{i_0}^k)^{-1} s_{i_0}^k \tilde{\lambda}_{i_0}^{\mathrm{a},k} , \quad \forall k > k' , \tag{B.67}$$

where $s_{i_0}^k > 0$, $\lambda_{i_0}^k > 0$ by construction of Algorithm CR-MPC. Thus, from (B.66), we obtain $\Delta s_{i_0}^{\mathrm{a},k} > 0$ for all $k > k'$. For all $k > k'$, the last equation of (B.2) gives

$$\Delta s_{i_0}^k = (\lambda_{i_0}^k)^{-1}(-s_{i_0}^k \tilde{\lambda}_{i_0}^k + \gamma^k \sigma^k \mu_{(Q^k)}^k - \gamma^k \Delta s_{i_0}^{\mathrm{a},k} \Delta \lambda_{i_0}^{\mathrm{a},k}) , \tag{B.68}$$

where $\gamma^k \geq 0$, $\sigma^k \geq 0$, and $\mu_{(Q^k)}^k \geq 0$ by construction of Algorithm CR-MPC. Also, for $k > k'$, $\Delta \lambda_{i_0}^{\mathrm{a},k} < 0$ since $\lambda_{i_0}^k > 0$ and $\tilde{\lambda}_{i_0}^{\mathrm{a},k} < 0$. It is then easily verified that all terms in (B.68) are non-negative and the first term is positive, thus $\Delta s_{i_0}^k > 0$ for all $k > k'$. Moreover, for all $k > k'$, we have $s_{i_0}^{k+1} = s_{i_0}^k + \alpha_{\mathrm{p}}^k \Delta s_{i_0}^k > s_{i_0}^k > 0$, where $\alpha_{\mathrm{p}}^k > 0$ since $\mathbf{s}^k > \mathbf{0}$. We then conclude that $\{s_{i_0}^k\} \to s_{i_0}^* > 0$ and $s_{i_0}^* \lambda_{i_0}^* < 0$, which contradicts to the stationarity of $\mathbf{x}^*$. □

Proof of Proposition 3. From Theorem 3, we know $\{\mathbf{x}^k\}$ converges to $\mathcal{F}_P^*$. Thus, there exists an infinite index set K such that $\{\mathbf{x}^k\} \to \mathbf{x}^* \in \mathcal{F}_P^*$ on K. Suppose $\{\Delta \mathbf{x}^k\} \to \mathbf{0}$ on K, then it follows from Lemma 9 that $\{\Delta \mathbf{x}^{\mathrm{a},k}\} \to \mathbf{0}$ on K. By Lemma 11, we have that $\{\tilde{\boldsymbol{\lambda}}^k\} \to \boldsymbol{\lambda}^*$ on K, where $\boldsymbol{\lambda}^*$ is the unique multiplier associated with $\mathbf{x}^*$. Hence, in (4.15), $\{E^k\}$ converges to zero as $k \to \infty$, $k \in K$, and the claim holds.

On the other hand, suppose $\{\Delta \mathbf{x}^k\} \not\to \mathbf{0}$ on K. From Lemma 12, there exists an infinite index set $K' \subseteq K$ such that $\{\Delta \mathbf{x}^{\mathrm{a},k-1}\} \to \mathbf{0}$ on K'. On the other hand, from (4.19), (4.14) and (4.7), we have

$$\|\mathbf{x}^k - \mathbf{x}^{k-1}\| = \|\alpha_{\mathrm{p}}^{k-1} \Delta \mathbf{x}^{k-1}\| \leq \|\Delta \mathbf{x}^{k-1}\| \leq (1+\tau)\|\Delta \mathbf{x}^{\mathrm{a},k-1}\| . \tag{B.69}$$

Thus, $\{\mathbf{x}^{k-1}\}$ also converges to $\mathbf{x}^*$ on K'. Define infinite index set $K'' := \{k : k+1 \in K'\}$. Then, on K'', we have $\{\mathbf{x}^k\} \to \mathbf{x}^*$ and $\{\Delta\mathbf{x}^{\mathrm{a},k}\} \to \mathbf{0}$. Applying Lemma 11 on K'' yields $\{\tilde{\boldsymbol{\lambda}}^k\} \to \boldsymbol{\lambda}^*$ on K''. Hence, $\{E^k\}$ in (4.15) again converges to zero as $k \to \infty$, $k \in K''$, and the claim holds. □

Suppose the primal optimal set $\mathcal{F}_P^*$ is a singleton, i.e., Assumption 6 holds. The following lemma then gives the global convergence property of the dual iterates $\{\tilde{\boldsymbol{\lambda}}^k\}$ generated by Algorithm CR-MPC.

Lemma 15 (Corresponding to Lemma B.17 in [74])**.** *Under Assumptions 1–6, we have*

(i) $\{\Delta\mathbf{x}^{a,k}\} \to \mathbf{0}$ *and* $\{\Delta\mathbf{x}^k\} \to \mathbf{0}$,

(ii) $\{\tilde{\boldsymbol{\lambda}}^{a,k}\} \to \boldsymbol{\lambda}^*$ *and* $\{\tilde{\boldsymbol{\lambda}}^k\} \to \boldsymbol{\lambda}^*$.

Proof. Since $\{\mathbf{x}^k\} \to \mathbf{x}^*$, we have $\{\Delta\mathbf{x}^k\} \to \mathbf{0}$, and Claim (i) immediately follows from Lemma 9. Claim (ii) is then given by Lemma 11. □

Proof of Proposition 4. Proposition 4 is a direct consequence from Theorem 3 and Lemma 15. □

Appendix C: Proof of Local Convergence Rate for Algorithm CR-MPC

In this appendix, we will show that, under Assumptions 1–7, the primal-dual iterate $[\mathbf{x}^{k^T}, \boldsymbol{\lambda}^{k^T}]^T$ generated by Algorithm CR-MPC converges to the solution at a

locally q-quadratic rate. Hence, we assume Assumptions 1–7 hold in the following analysis.

Some useful notations in the results in this section are defined in the following definitions.

Definition 6. $J_a(A, \mathbf{s}, \boldsymbol{\lambda}) := \begin{bmatrix} H & -A^T \\ \Lambda A & S \end{bmatrix}$.

Definition 7. $\tilde{\mathbf{s}} := \mathbf{s} + \Delta\mathbf{s}$, *and* $\tilde{\mathbf{s}}^a := \mathbf{s} + \Delta\mathbf{s}^a$.

Lemma 16 (Corresponding to Lemma B.15. in [74])**.** $J_a(A, \mathbf{s}, \boldsymbol{\lambda})$ *is nonsingular if and only if* $J(A, \mathbf{s}, \boldsymbol{\lambda})$ *is nonsingular.*

Proof. See Lemma B.15. in [74]. □

In the rest of the proof, we use $\mathbf{x}^*$ to denote the optimal solution of (P) (the existence and uniqueness of $\mathbf{x}^*$ is guaranteed by Assumption 6), $\boldsymbol{\lambda}^*$ to denote the Lagrange multiplier associated to $\mathbf{x}^*$, and define $\mathbf{s}^* := A\mathbf{x}^* - \mathbf{b}$. Note that from Theorem 3, we have $\{\mathbf{x}^k\} \to \mathbf{x}^*$.

Lemma 17 (Corresponding to Lemma B.16. in [74])**.** *For any* Q *such that* $\mathcal{A}(\mathbf{x}^*) \subseteq Q$, $J_a(A_Q, \mathbf{s}^*_Q, \boldsymbol{\lambda}^*_Q)$ *and* $J(A_Q, \mathbf{s}^*_Q, \boldsymbol{\lambda}^*_Q)$ *are nonsingular.*

Proof. See Lemma B.16. in [74]. □

Lemma 18. *Under our assumptions, there exists* $k' > 0$ *such that,* $\mathcal{A}(\mathbf{x}^*) \subseteq Q^k$ *for all* $k > k'$.

Proof. For all $Q \subseteq \mathcal{I}$, define $K(Q) := \{k \in \mathbb{N} : Q^k = Q\}$ and $\mathcal{Q} := \{Q \subseteq \mathcal{I} : |K(Q)| = \infty\}$. Since $|\mathcal{Q}|$ is finite ($|\mathcal{I}|$ is finite), it suffices to show that, for all $Q \in \mathcal{Q}$, $\mathcal{A}(\mathbf{x}^*) \subseteq Q$ holds.

For all $Q \in \mathcal{Q}$, Claim (ii) in Lemma 15 gives that $\{\tilde{\boldsymbol{\lambda}}^k_{\mathcal{I}\setminus Q}\} \to \boldsymbol{\lambda}^*_{\mathcal{I}\setminus Q}$ on $K(Q)$. By (B.3), $\tilde{\boldsymbol{\lambda}}^k_{\mathcal{I}\setminus Q} = \mathbf{0}$ for all $k \in K(Q)$. Thus, $\boldsymbol{\lambda}^*_{\mathcal{I}\setminus Q} = \mathbf{0}$. It then follows from Assumption 7 that $\mathbf{s}^*_{\mathcal{I}\setminus Q} > \mathbf{0}$. Hence, by Definition 4, we conclude that $\mathcal{A}(\mathbf{x}^*) \subseteq Q$, for all $Q \in \mathcal{Q}$. □

Lemma 19 (Corresponding to Lemma A.2 in the supplementary materials of [37])**.** *Under our assumptions. Let* $\mathbf{x} \in \mathcal{F}_P^o$, $\boldsymbol{\lambda} > \mathbf{0}$, $\mathbf{s} := A\mathbf{x} - \mathbf{b}$. *If* $\tilde{\boldsymbol{\lambda}}$ *and* $\tilde{\mathbf{s}}$ *produced by Algorithm CR-MPC are such that* $\tilde{\lambda}_i > 0$ *for all* $i \in \mathcal{A}(\mathbf{x}^*)$ *and* $\tilde{s}_i > 0$ *for all* $i \in \mathcal{I} \setminus \mathcal{A}(\mathbf{x}^*)$*, the step sizes generated by Algorithm CR-MPC then satisfy*

$$\bar{\alpha}_p \geq \min\left\{1, \min_{i\in\mathcal{A}(\mathbf{x}^*)}\left\{\frac{\lambda_i}{|\tilde{\lambda}^a_i|}\right\}, \min_{i\in\mathcal{A}(\mathbf{x}^*)}\left\{\frac{\lambda_i}{|\tilde{\lambda}_i|} - \frac{\Delta\lambda^a_i}{|\tilde{\lambda}_i|}\right\}\right\}, \tag{C.1}$$

and

$$\bar{\alpha}_d \geq \min\left\{1, \min_{i\in Q\setminus\mathcal{A}(\mathbf{x}^*)}\left\{\frac{s_i}{|\tilde{s}^a_i|}\right\}, \min_{i\in Q\setminus\mathcal{A}(\mathbf{x}^*)}\left\{\frac{s_i}{|\tilde{s}_i|} - \frac{\Delta s^a_i}{|\tilde{s}_i|}\right\}\right\}, \tag{C.2}$$

where Q, $\Delta\boldsymbol{\lambda}^a_Q$, $\Delta\mathbf{s}^a$, $\tilde{\boldsymbol{\lambda}}^a$, $\tilde{\mathbf{s}}^a$, $\tilde{\boldsymbol{\lambda}}$, *and* $\tilde{\mathbf{s}}$ *are all generated by Algorithm CR-MPC.*

Proof. Let us first consider (C.1). If $\bar{\alpha}_{\mathrm{p}} \geq 1$, (C.1) holds trivially. From (4.18), we know that if $\bar{\alpha}_{\mathrm{p}} < 1$, there exists some index i_0 such that

$$\Delta s_{i_0} < 0 \quad \text{and} \quad \bar{\alpha}_{\mathrm{p}} = \frac{s_{i_0}}{-\Delta s_{i_0}} < 1\,. \tag{C.3}$$

If $i_0 \in \mathcal{I} \setminus \mathcal{A}(\mathbf{x}^*)$, then by assumption, $s_{i_0} > 0$ (since $\mathbf{x} \in \mathcal{F}_P^o$) and $\tilde{s}_{i_0} = s_{i_0} + \Delta s_{i_0} > 0$, contradicting (C.3). Thus we have $i_0 \in \mathcal{A}(\mathbf{x}^*)$. Now we consider two cases: $|\Delta s^{\mathrm{a}}_{i_0}| \geq |\Delta s_{i_0}|$ and $|\Delta s^{\mathrm{a}}_{i_0}| < |\Delta s_{i_0}|$. If $|\Delta s^{\mathrm{a}}_{i_0}| \geq |\Delta s_{i_0}|$, then it follows from (C.3) and (B.5) that

$$\bar{\alpha}_{\mathrm{p}} = \frac{s_{i_0}}{|\Delta s_{i_0}|} \geq \frac{s_{i_0}}{|\Delta s^{\mathrm{a}}_{i_0}|} = \frac{\lambda_{i_0}}{|\tilde{\lambda}^{\mathrm{a}}_{i_0}|}\,, \tag{C.4}$$

which verifies (C.1). On the other hand, suppose $|\Delta s^{\mathrm{a}}_{i_0}| < |\Delta s_{i_0}|$, since $\tilde{\lambda}_i > 0$ for $i \in \mathcal{A}(\mathbf{x}^*)$, taking the absolute value of the i_0 components in the last equation of (B.2) yields

$$\tilde{\lambda}_{i_0} s_{i_0} \geq \lambda_{i_0} |\Delta s_{i_0}| + \gamma\sigma\mu_{(Q)} - \gamma |\Delta s^{\mathrm{a}}_{i_0}||\Delta \lambda^{\mathrm{a}}_{i_0}| \,, \tag{C.5}$$

where $\gamma \geq 0$, $\sigma \geq 0$, and $\mu_{(Q)} \geq 0$ are generated by Algorithm CR-MPC. By (C.5) and (C.3), the primal step size is then bounded by

$$\begin{aligned} \bar{\alpha}_{\mathrm{p}} = \frac{s_{i_0}}{|\Delta s_{i_0}|} &\geq \frac{\lambda_{i_0}}{\tilde{\lambda}_{i_0}} + \frac{\gamma\sigma\mu_{(Q)}}{\tilde{\lambda}_{i_0}|\Delta s_{i_0}|} - \frac{\gamma|\Delta s^{\mathrm{a}}_{i_0}||\Delta\lambda^{\mathrm{a}}_{i_0}|}{\tilde{\lambda}_{i_0}|\Delta s_{i_0}|} \\ &\geq \frac{\lambda_{i_0}}{\tilde{\lambda}_{i_0}} - \gamma \frac{|\Delta s^{\mathrm{a}}_{i_0}|}{|\Delta s_{i_0}|} \frac{|\Delta\lambda^{\mathrm{a}}_{i_0}|}{|\tilde{\lambda}_{i_0}|} \geq \frac{\lambda_{i_0}}{\tilde{\lambda}_{i_0}} - \frac{|\Delta\lambda^{\mathrm{a}}_{i_0}|}{|\tilde{\lambda}_{i_0}|} \,, \end{aligned} \tag{C.6}$$

where the last inequality holds since $\gamma \leq 1$, and $|\Delta s^{\mathrm{a}}_{i_0}| < |\Delta s_{i_0}|$. Hence, (C.1) holds.

Following a very similar argument that flips the roles of $\mathbf{s}$ and $\boldsymbol{\lambda}$, one can prove that (C.2) also holds. □

Lemma 20 (Corresponding to Lemma B.17. in [74])**.** *Under our assumptions, if* $\lambda^*_i < \lambda_{\max}$ *for all* $i \in \mathcal{I}$*, then*

(i) $\{\hat{\boldsymbol{\lambda}}^k\} \to \boldsymbol{\lambda}^*$ *and* $\{\boldsymbol{\lambda}^k\} \to \boldsymbol{\lambda}^*$.

(ii) $\{\Delta\boldsymbol{\lambda}^{a,k}\} \to \mathbf{0}$ *and* $\{\Delta\boldsymbol{\lambda}^k\} \to \mathbf{0}$.

Proof. First consider Claim (i). It follows from Lemma 15 that $\{\tilde{\boldsymbol{\lambda}}^k\} \to \boldsymbol{\lambda}^*$. Hence, we prove Claim (i) by showing $\|\tilde{\boldsymbol{\lambda}}^k - \hat{\boldsymbol{\lambda}}^k\| \to 0$ and $\|\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k\| \to 0$ as $k \to \infty$. Note that, suppose $\{\hat{\boldsymbol{\lambda}}^k\} \to \boldsymbol{\lambda}^*$, then $\|\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k\| \to 0$ implies $\{\boldsymbol{\lambda}^{k+1}\} \to \boldsymbol{\lambda}^*$, thus $\{\boldsymbol{\lambda}^k\} \to \boldsymbol{\lambda}^*$.

Based on (4.20) and (4.21), we first split $\|\tilde{\boldsymbol{\lambda}}^k - \hat{\boldsymbol{\lambda}}^k\|$ into $\|\tilde{\boldsymbol{\lambda}}^k_{Q^k} - \hat{\boldsymbol{\lambda}}^k_{Q^k}\|$ and $\|\tilde{\boldsymbol{\lambda}}^k_{\mathcal{I}\setminus Q^k} - \hat{\boldsymbol{\lambda}}^k_{\mathcal{I}\setminus Q^k}\|$, and show the convergence for each term. For all $Q \subseteq \mathcal{I}$, define

index set $K(Q) := \{k \in \mathbb{N} : Q^k = Q\}$. Since there are only finitely many choices of Q ($\mathcal{I}$ is finite), it suffices to show that, for all $Q \in \mathcal{Q}$, $\|\tilde{\boldsymbol{\lambda}}^k_{Q^k} - \hat{\boldsymbol{\lambda}}^k_{Q^k}\|$ and $\|\tilde{\boldsymbol{\lambda}}^k_{\mathcal{I}\setminus Q^k} - \hat{\boldsymbol{\lambda}}^k_{\mathcal{I}\setminus Q^k}\|$ converge to zero on $K(Q)$, where $\mathcal{Q} := \{Q \subseteq \mathcal{I} : |K(Q)| = \infty,\, Q \neq \emptyset\}$.

Now, let $Q \in \mathcal{Q}$ and $k \in K(Q)$, then from (4.20), (4.21), and (B.3), we have

$$\|\tilde{\boldsymbol{\lambda}}^k_{Q^k} - \hat{\boldsymbol{\lambda}}^k_{Q^k}\| = \|\tilde{\boldsymbol{\lambda}}^k_{Q} - \hat{\boldsymbol{\lambda}}^k_{Q}\| = (1 - \alpha^k_{\mathrm{d}})\|\Delta\boldsymbol{\lambda}^k_Q\|\,, \tag{C.7}$$

and

$$\|\tilde{\boldsymbol{\lambda}}^k_{\mathcal{I}\setminus Q^k} - \hat{\boldsymbol{\lambda}}^k_{\mathcal{I}\setminus Q^k}\| = \|\tilde{\boldsymbol{\lambda}}^k_{\mathcal{I}\setminus Q} - \hat{\boldsymbol{\lambda}}^k_{\mathcal{I}\setminus Q}\| = \mu^{k+1}_{(Q)}\|(S^{k+1}_{\mathcal{I}\setminus Q})^{-1}\mathbf{1}\|\,. \tag{C.8}$$

Since the boundedness of $\{\boldsymbol{\lambda}^k\}$ (by construction) and $\{\tilde{\boldsymbol{\lambda}}^k\}$ (by Claim (ii) in Lemma 15) implies the boundedness of $\{\Delta\boldsymbol{\lambda}^k_Q\}$, we only need $\{\alpha^k_{\mathrm{d}}\} \to 1$ to guarantee the right-hand-side of (C.7) converges to zero on $K(Q)$. Now, Theorem 3 and Lemma 15 give $\{\mathbf{x}^k\} \to \mathbf{x}^*$, $\{\Delta\mathbf{x}^{\mathrm{a},k}\} \to \mathbf{0}$, $\{\Delta\mathbf{x}^k\} \to \mathbf{0}$, and $\{\tilde{\boldsymbol{\lambda}}^k\} \to \boldsymbol{\lambda}^*$. Hence, by definitions, we have $\{\mathbf{s}^k\} \to \mathbf{s}^*$, $\{\Delta\mathbf{s}^{\mathrm{a},k}\} \to \mathbf{0}$, and $\{\Delta\mathbf{s}^k\} \to \mathbf{0}$, which also imply that $\{\tilde{\mathbf{s}}^{\mathrm{a},k}\} \to \mathbf{s}^*$ and $\{\tilde{\mathbf{s}}^k\} \to \mathbf{s}^*$. Moreover, Assumption 7 gives $\lambda^*_i > 0$ for all $i \in \mathcal{A}(\mathbf{x}^*)$ and $s^*_i > 0$ for all $i \in \mathcal{I} \setminus \mathcal{A}(\mathbf{x}^*)$. Hence, for sufficiently large k, the assumptions of Lemma 19 are satisfied, i.e., $\tilde{\lambda}^k_i > 0$ for all $i \in \mathcal{A}(\mathbf{x}^*)$ and $\tilde{s}^k_i > 0$ for all $i \in \mathcal{I}\setminus\mathcal{A}(\mathbf{x}^*)$. Lemma 19 gives $\{\bar{\alpha}^k_{\mathrm{d}}\} \to 1$ as $k \to \infty$, since all terms on the right-hand-side of (C.2) converge to one. Thus, from (4.18) and the fact that $\{\Delta\mathbf{x}^k\} \to \mathbf{0}$, we have $\{\alpha^k_{\mathrm{d}}\} \to 1$ as $k \to \infty$. We then conclude that, for all $Q \in \mathcal{Q}$, $\|\tilde{\boldsymbol{\lambda}}^k_{Q^k} - \hat{\boldsymbol{\lambda}}^k_{Q^k}\| \to 0$, in other words, $\{\hat{\boldsymbol{\lambda}}^k_{Q^k}\} \to \boldsymbol{\lambda}^*_Q$, on $K(Q)$.

On the other hand, since $\{\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k}\} \to \boldsymbol{\lambda}^* \geq 0$ and $\{\Delta\mathbf{x}^{\mathrm{a},k}\} \to \mathbf{0}$, $\chi^k := \|\Delta\mathbf{x}^{\mathrm{a},k}\|^2 + \|[\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k}_{Q^k}]_-\|^2$ converges to zero as $k \to \infty$. It then follows from (4.20), (4.21), $\{\hat{\boldsymbol{\lambda}}^k_{Q^k}\} \to \boldsymbol{\lambda}^*_Q$ on $K(Q)$, and the assumption $\boldsymbol{\lambda}^* < \lambda_{\max}\mathbf{1}$ that, for all $Q \in \mathcal{Q}$, $\|\boldsymbol{\lambda}^{k+1}_{Q^k} - \hat{\boldsymbol{\lambda}}^k_{Q^k}\| \to 0$

on $K(Q)$, which implies $\{\boldsymbol{\lambda}_{Q^k}^{k+1}\} \to \boldsymbol{\lambda}_Q^*$ on $K(Q)$.

For (C.8), since $\mathbf{s}^k > \mathbf{0}$ for all k and $s_i^* > 0$ for all $i \in \mathcal{I} \setminus \mathcal{A}(\mathbf{x}^*)$, $\|(S_{\mathcal{I}\setminus Q}^{k+1})^{-1}\mathbf{1}\|$ is then bounded. Also, for all $Q \in \mathcal{Q}$, on $K(Q)$ we have $\{\mathbf{s}_{Q^k}^{k+1}\} = \{\mathbf{s}_Q^{k+1}\} \to \mathbf{s}_Q^*$ and $\{\boldsymbol{\lambda}_{Q^k}^{k+1}\} = \{\boldsymbol{\lambda}_Q^{k+1}\} \to \boldsymbol{\lambda}_Q^*$. Thus, by the definition of $\mu_{(Q)}^{k+1}$ and Assumption 7, we have

$$\left\{\mu_{(Q)}^{k+1}\right\} := \left\{\frac{(\mathbf{s}_Q^{k+1})^T(\boldsymbol{\lambda}_Q^{k+1})}{|Q|}\right\} \to \left\{\frac{(\mathbf{s}_Q^*)^T(\boldsymbol{\lambda}_Q^*)}{|Q|}\right\} = 0 \ \text{ on } K(Q)\,, \tag{C.9}$$

which leads to $\|\tilde{\boldsymbol{\lambda}}_{\mathcal{I}\setminus Q^k}^k - \hat{\boldsymbol{\lambda}}_{\mathcal{I}\setminus Q^k}^k\| \to 0$ on $K(Q)$, for all $Q \in \mathcal{Q}$. Hence, we have shown that $\|\tilde{\boldsymbol{\lambda}}^k - \hat{\boldsymbol{\lambda}}^k\| \to 0$ and thus $\{\hat{\boldsymbol{\lambda}}^k\} \to \boldsymbol{\lambda}^*$. It then again follows from (4.20), (4.21), $\{\hat{\boldsymbol{\lambda}}^k\} \to \boldsymbol{\lambda}^*$, $\{\chi^k\} \to 0$, and the assumption $\boldsymbol{\lambda}^* < \lambda_{\max}\mathbf{1}$ that, $\|\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k\| \to 0$, which implies $\{\boldsymbol{\lambda}^{k+1}\} \to \boldsymbol{\lambda}^*$. Claim (i) is now verified.

Next, let us consider Claim (ii). From Claim (i), we have $\{\boldsymbol{\lambda}^{k+1}\} \to \boldsymbol{\lambda}^*$, which is equivalent to $\{\boldsymbol{\lambda}^k\} \to \boldsymbol{\lambda}^*$. Also, Claim (ii) of Lemma 15 gives $\{\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k}\} \to \boldsymbol{\lambda}^*$ and $\{\tilde{\boldsymbol{\lambda}}^k\} \to \boldsymbol{\lambda}^*$. Claim (ii) is then a direct consequence of the fact that $\{\boldsymbol{\lambda}^k\}$, $\{\tilde{\boldsymbol{\lambda}}^{\mathrm{a},k}\}$, and $\{\tilde{\boldsymbol{\lambda}}^k\}$ all converge to $\boldsymbol{\lambda}^*$. □

For convenience, we use $\mathbf{z}$ to denote the vector contains the primal and dual variables, i.e., $\mathbf{z} := [\mathbf{x}^T, \boldsymbol{\lambda}^T]^T$. The strictly feasible set of (P) and (D) is then defined as

$$E^o := \{\mathbf{z} : \mathbf{x} \in \mathcal{F}_P^o,\ \boldsymbol{\lambda} > \mathbf{0}\}\,. \tag{C.10}$$

In the rest of the proof for local convergence rate, we only consider points in $E^o \cap B(\mathbf{z}^*, \rho)$, the set of strictly feasible points in a ball $B(\mathbf{z}^*, \rho) := \{\mathbf{z} \in \mathbb{R}^{n+m} : \|\mathbf{z}-\mathbf{z}^*\| \le \rho\}$, where $\mathbf{z}^* := [\mathbf{x}^{*T}, \boldsymbol{\lambda}^{*T}]^T$. We also define

$$\mathcal{Q}^* := \{Q \subseteq \mathcal{I} : \mathcal{A}(\mathbf{x}^*) \subseteq Q\}\,. \tag{C.11}$$

Note that, since $\{\mathbf{z}^k\}$ converges to $\mathbf{z}^*$ (by Theorem 3 and Lemma 20), Lemma 18 implies that, if ρ is sufficiently small, the working set Q is always in $\mathcal{Q}^*$ for any $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho)$.

Lemma 21 (Corresponding to Lemma A.5 in the supplementary materials of [37])**.** *Under our assumptions. Let* $\beta \in (0,1)$ *and* $\epsilon^* := \min\{1, \min_{i\in\mathcal{I}}(\lambda_i^* + s_i^*)\}$*. Then there exist* $\rho^* > 0$ *and* $R > 0$*, such that, for all* $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho^*)$ *and all* $Q \in \mathcal{Q}^*$*, the followings hold:*

(i) $\|J_a(A_Q, \boldsymbol{\lambda}_Q, \mathbf{s}_Q)^{-1}\| \le R$,

(ii) $\max\{\|\Delta\mathbf{z}_Q^a\|, \|\Delta\mathbf{z}_Q\|, \|\Delta\mathbf{s}_Q^a\|, \|\Delta\mathbf{s}_Q\|\} < \epsilon^*/4$,

(iii) $\min\{\lambda_i, \tilde{\lambda}_i^a, \tilde{\lambda}_i\} > \epsilon^*/2,\ \forall i \in \mathcal{A}(\mathbf{x}^*)$,

$\max\{\lambda_i, \tilde{\lambda}_i^a, \tilde{\lambda}_i\} < \epsilon^*/2,\ \forall i \in \mathcal{I} \setminus \mathcal{A}(\mathbf{x}^*)$,

$\max\{s_i, \tilde{s}_i^a, \tilde{s}_i\} < \epsilon^*/2,\ \forall i \in \mathcal{A}(\mathbf{x}^*)$,

$\min\{s_i, \tilde{s}_i^a, \tilde{s}_i\} > \epsilon^*/2,\ \forall i \in \mathcal{I} \setminus \mathcal{A}(\mathbf{x}^*)$.

(iv) If $\bar{\alpha}_p$ *and* $\bar{\alpha}_d$ *are finite,*

$\beta\bar{\alpha}_p < \bar{\alpha}_p - \|\Delta\mathbf{x}\|$,

$\beta\bar{\alpha}_d < \bar{\alpha}_d - \|\Delta\mathbf{x}\|$.

Proof. Since $\mathcal{I}$ is finite, $\mathcal{Q}^*$ is also finite. Therefore, it is suffice to show that, for every $Q \in \mathcal{Q}^*$, there exist $\rho_Q > 0$ and $R_Q > 0$ such that the claims hold for all $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho_Q)$. In such case, the lemma is proved with $\rho^* := \min_{Q\in\mathcal{Q}^*} \rho_Q$ and $R^* := \min_{Q\in\mathcal{Q}^*} R_Q$. Thus, we now consider only a fixed $Q \in \mathcal{Q}^*$, and find the corresponding $\rho_Q > 0$ and $R_Q > 0$ such that all claims hold.

Claim (i) follows from Lemma 17 and the continuity of $J_a(A_Q, \boldsymbol{\lambda}_Q, \mathbf{s}_Q)$. Claim (ii) follows from Claim (i), Lemma 16, and the continuity of the right-hand-sides of (4.9) and (B.1), which are zero at the solution. Claim (iii) is true due to the strict complementary slackness given by Assumption 7, the definition of ϵ^*, and Claim (ii). For Claim (iv), first note that the assumptions of Lemma 19 is satisfied due to Claim (iii), then positive lower bounds on $\bar{\alpha}_{\mathrm{p}}$ and $\bar{\alpha}_{\mathrm{d}}$ can be obtained by applying Lemma 19. Following the same argument in the proof of Claim (ii), it is clear that, given sufficiently small ρ_Q, Claim (iv) holds. □

Proposition 8 (Corresponding to Proposition B.18. of [74]). *Let $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ be twice continuously differentiable and let $\mathbf{t}^* \in \mathbb{R}^n$ such that $\Phi(\mathbf{t}^*) = 0$. Suppose there exists $\rho > 0$ such that $\frac{\partial \Phi}{\partial \mathbf{t}}(\mathbf{t})$ is nonsingular for all $\mathbf{t} \in B(\mathbf{t}^*, \rho)$. Define $\Delta^N \mathbf{t}$ as the Newton increment at $\mathbf{t}$, i.e., $-\left(\frac{\partial \Phi}{\partial \mathbf{t}}(\mathbf{t})\right)^{-1} \Phi(\mathbf{t})$. Then, given any $c > 0$, for all $\mathbf{t} \in B(\mathbf{t}^*, \rho)$, if $\mathbf{t}^+ \in \mathbb{R}^n$ satisfies*

$$\min\{|t_i^+ - t_i^*|, |t_i^+ - (t_i + \Delta^N t_i)|\} \leq c\|\Delta^N \mathbf{t}\|^2, \quad \forall i \in \mathcal{I}\,, \tag{C.12}$$

there exists $\hat{c} > 0$ such that

$$\|\mathbf{t}^+ - \mathbf{t}^*\| \leq \hat{c}\|\mathbf{t} - \mathbf{t}^*\|^2\,. \tag{C.13}$$

Proof. See Proposition B.18. of [74]. □

Corollary 3 (Corresponding to Corollary A.7 in the supplementary materials of [37]). *Let Φ, $\mathbf{t}^*$, ρ, and $\Delta^N \mathbf{t}$ be as in Proposition 8. Then, given any $c > 0$, for all*

$\mathbf{t} \in B(\mathbf{t}^*, \rho)$, *if* $\mathbf{t}^+ \in \mathbb{R}^n$ *satisfies*

$$\min\{|t_i^+ - t_i^*|, |t_i^+ - (t_i + \Delta^N t_i)|\} \leq c \max\{\|\Delta^N \mathbf{t}\|^2, \|\mathbf{t} - \mathbf{t}^*\|^2\}, \quad \forall i \in \mathcal{I}\,, \tag{C.14}$$

there exists $c^* > 0$ *such that*

$$\|\mathbf{t}^+ - \mathbf{t}^*\| \leq c^* \|\mathbf{t} - \mathbf{t}^*\|^2\,. \tag{C.15}$$

Proof. See Corollary A.7 in the supplementary materials of [37]. □

In the following proof, we define Φ as

$$\Phi(\mathbf{z}) := \begin{bmatrix} H\mathbf{x} - A^T\boldsymbol{\lambda} + \mathbf{c} \\ \\ \Lambda(A\mathbf{x} - \mathbf{b}) \end{bmatrix}. \tag{C.16}$$

Hence, the first three conditions of the KKT system (4.1) are equivalent to $\Phi(\mathbf{z}) = \mathbf{0}$, and let $\mathbf{s} := A\mathbf{x} - \mathbf{b}$, $J_a(A, \mathbf{s}, \boldsymbol{\lambda})$ is the Jacobian of $\Phi(\mathbf{z})$, i.e.,

$$J_a(A, \mathbf{s}, \boldsymbol{\lambda})\Delta^N \mathbf{z} = -\Phi(\mathbf{z})\,. \tag{C.17}$$

Note that $J_a(A, \mathbf{s}, \boldsymbol{\lambda})$ is nonsingular near $\mathbf{z}*$ (Lemma 21 (i)), and the unreduced affine-scaling direction given in (4.2) is exactly the Newton direction for the solution of $\Phi(\mathbf{z}) = \mathbf{0}$.

To make use of Proposition 8, we next verify that the iterate generated by Algorithm CR-MPC satisfies the condition in Proposition 8.

Let $\mathbf{z}_Q := [\mathbf{x}^T, \boldsymbol{\lambda}_Q^T]^T$, then the step taken on the Q components along the search direction generated by the CR-MPC algorithm is analogously given by

$$\hat{\mathbf{z}}_Q^+ := [\mathbf{x}^{+T}, \hat{\boldsymbol{\lambda}}_Q^T]^T\,. \tag{C.18}$$

We then compare $\hat{\mathbf{z}}_Q^+$ to the Q components of the Newton/affine-scaling step, i.e., $\mathbf{z}_Q + \Delta^N \mathbf{z}_Q$.

First, for $Q \in \mathcal{Q}^*$ define

$$\mathscr{A} := \begin{bmatrix} \alpha_{\mathrm{p}} I_n & 0 \\ 0 & \alpha_{\mathrm{d}} I_{|Q|} \end{bmatrix}, \tag{C.19}$$

and let

$$\alpha = \min\{\alpha_{\mathrm{p}}, \alpha_{\mathrm{d}}\}\,. \tag{C.20}$$

The difference between the CR-MPC and the Newton/affine-scaling steps can be written as

$$\begin{aligned} &\|\hat{\mathbf{z}}_Q^+ - (\mathbf{z}_Q + \Delta^N \mathbf{z}_Q)\| \\ &\leq \|\hat{\mathbf{z}}_Q^+ - (\mathbf{z}_Q + \Delta \mathbf{z}_Q)\| + \|\Delta \mathbf{z}_Q - \Delta \mathbf{z}_Q^{\mathrm{a}}\| + \|\Delta \mathbf{z}_Q^{\mathrm{a}} - \Delta^N \mathbf{z}_Q\| \\ &= \|(I - \mathscr{A})\Delta \mathbf{z}_Q\| + \gamma \|\Delta \mathbf{z}_Q^{\mathrm{c}}\| + \|\Delta \mathbf{z}_Q^{\mathrm{a}} - \Delta^N \mathbf{z}_Q\| \\ &\leq (1-\alpha)\|\Delta \mathbf{z}_Q\| + \|\Delta \mathbf{z}_Q^{\mathrm{c}}\| + \|\Delta \mathbf{z}_Q^{\mathrm{a}} - \Delta^N \mathbf{z}_Q\|\,. \end{aligned} \tag{C.21}$$

The following lemmas provide bounds on the three terms in the last line of (C.21). Note that the ρ^* used in the following lemmas comes from Lemma 21.

Lemma 22 (Corresponding to Lemma B.19. of [74])**.** *For all $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho^*)$, and for all $Q \in \mathcal{Q}^*$, there exists a constant $c_1 > 0$ such that*

$$\|\Delta \mathbf{z}_Q^a - \Delta^N \mathbf{z}_Q\| \leq c_1 \|\mathbf{z} - \mathbf{z}^*\| \|\Delta^N \mathbf{z}_Q\|\,. \tag{C.22}$$

Proof. See Lemma B.19 of [74]. □

Lemma 23 (Corresponding to Lemma A.9 in the supplementary materials of [37])**.** *Under our assumptions, there exists a constant $c_2 > 0$ such that for all $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho^*)$, and for all $Q \in \mathcal{Q}^*$,*

$$\|\Delta \mathbf{z}_Q^c\| \le c_2 \|\Delta \mathbf{z}_Q^a\|^2 . \tag{C.23}$$

Proof. Let $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho^*)$ and $Q \in \mathcal{Q}^*$. Using (4.5) and Lemma 21 (i) yields

$$\begin{aligned} \|\Delta \mathbf{z}_Q^{\mathrm{c}}\| &\le \|J_a(A_Q, \mathbf{s}_Q, \boldsymbol{\lambda}_Q)\| \left\| \begin{pmatrix} 0 \\ \sigma \mu_{(Q)} \mathbf{1} - \Delta S_Q^{\mathrm{a}} \Delta \boldsymbol{\lambda}_Q^{\mathrm{a}} \end{pmatrix} \right\| \\ &\le R(\sqrt{m}\sigma \mu_{(Q)} + \|\Delta S_Q^{\mathrm{a}} \Delta \boldsymbol{\lambda}_Q^{\mathrm{a}}\|) . \end{aligned} \tag{C.24}$$

Note that the second term can be bounded by

$$\|\Delta S_Q^{\mathrm{a}} \Delta \boldsymbol{\lambda}_Q^{\mathrm{a}}\| \le \|\Delta s_Q^{\mathrm{a}}\| \|\Delta \boldsymbol{\lambda}_Q^{\mathrm{a}}\| \le \|A_Q\| \|\Delta \mathbf{x}^{\mathrm{a}}\| \|\Delta \boldsymbol{\lambda}_Q^{\mathrm{a}}\| \le \|A\| \|\Delta \mathbf{z}_Q^{\mathrm{a}}\|^2 , \tag{C.25}$$

and that $\mu_{(Q)} := \mathbf{s}_Q^T \boldsymbol{\lambda}_Q / |Q|$ is bounded on $E^o \cap B(\mathbf{z}^*, \rho^*)$. Hence it suffices to show that there exists some constant d independent to $\mathbf{z}$ and Q such that

$$\sigma \le d \|\Delta \mathbf{z}_Q^{\mathrm{a}}\|^2 . \tag{C.26}$$

In Step 6 of Algorithm CR-MPC, $\sigma := (1 - \alpha^{\mathrm{a}})^3$, with $\alpha^{\mathrm{a}} \in [0, 1]$ defined in (4.16). For $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho^*)$ and $Q \in \mathcal{Q}^*$, Lemma 21 (ii) and (iii) gives that

$$\frac{s_i}{\Delta s_i^a} > \frac{\epsilon^*/2}{\epsilon^*/4} = 2, \quad \forall i \in \mathcal{I} \setminus \mathcal{A}(\mathbf{x}^*), \quad \text{and} \quad \frac{\lambda_i}{\Delta \lambda^a} > \frac{\epsilon^*/2}{\epsilon^*/4} = 2, \quad \forall i \in \mathcal{A}(\mathbf{x}^*). \tag{C.27}$$

It then follows from (4.16) that, if $\alpha^{\mathrm{a}} < 1$ $(1 - \alpha^{\mathrm{a}} \neq 0)$, then

$$\alpha^{\mathrm{a}} = \frac{s_i}{-\Delta s_i^{\mathrm{a}}}, \text{ for some } i \in \mathcal{A}(\mathbf{x}^*), \text{ or } \alpha^{\mathrm{a}} = \frac{\lambda_i}{-\Delta \lambda_i^{\mathrm{a}}}, \text{ for some } i \in \mathcal{I} \setminus \mathcal{A}(\mathbf{x}^*) . \tag{C.28}$$

In the former case, using the last equation of (4.13) and Lemma 21 (iii), we have, for some $i \in \mathcal{A}(\mathbf{x}^*)$,

$$1 - \alpha^{\mathrm{a}} = 1 - \frac{s_i}{-\Delta s_i^{\mathrm{a}}} = 1 - \frac{\lambda_i}{\tilde{\lambda}_i^{\mathrm{a}}} = \frac{\Delta \lambda_i^{\mathrm{a}}}{\tilde{\lambda}_i^{\mathrm{a}}} \leq \frac{2}{\epsilon^*} \|\Delta \mathbf{z}_Q^{\mathrm{a}}\| \,. \tag{C.29}$$

Similarly, in the latter case, we have, for some $i \in \mathcal{I} \setminus \mathcal{A}(\mathbf{x}^*)$,

$$1 - \alpha^{\mathrm{a}} = 1 - \frac{\lambda_i}{-\Delta \lambda_i^{\mathrm{a}}} = 1 - \frac{s_i}{\tilde{s}_i^{\mathrm{a}}} = \frac{\Delta s_i^{\mathrm{a}}}{\tilde{s}_i^{\mathrm{a}}} \leq \frac{2}{\epsilon^*} \|A\| \|\Delta \mathbf{z}_Q^{\mathrm{a}}\| \,. \tag{C.30}$$

Hence, (C.26) holds with $d = (2/\epsilon^*)^3 \max\{\|A\|^3, 1\}$. The proof is then complete. □

Note that an upper bound on magnitude of the MPC search direction $\Delta \mathbf{z}_Q$ can be obtained by using Lemma 23 and Lemma 21 (ii), specifically,

$$\|\Delta \mathbf{z}_Q\| \leq \|\Delta \mathbf{z}_Q^{\mathrm{a}}\| + \|\Delta \mathbf{z}_Q^{\mathrm{c}}\| \leq \|\Delta \mathbf{z}_Q^{\mathrm{a}}\| + c_2 \|\Delta \mathbf{z}_Q^{\mathrm{a}}\|^2 \leq \left(1 + c_2 \frac{\epsilon^*}{4}\right) \|\Delta \mathbf{z}_Q^{\mathrm{a}}\| \,. \tag{C.31}$$

This bound will be used in the proof of Lemma 24 and 25.

Lemma 24 (Corresponding to Lemma A.10 in the supplementary materials of [37])**.** *Under our assumptions, there exists a constant $c_3 > 0$ such that for all $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho^*)$, and for all $Q \in \mathcal{Q}^*$,*

$$|1 - \alpha| \leq c_3 \|\Delta \mathbf{z}_Q^a\| \,. \tag{C.32}$$

Proof. From the definition of α in (C.20), it suffices to show that, for $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho^*)$ and $Q \in \mathcal{Q}^*$, there exist some d_1 and d_2 independent of $\mathbf{z}$ and Q such that

$$|1 - \alpha_{\mathrm{p}}| \leq d_1 \|\Delta \mathbf{z}_Q^{\mathrm{a}}\| \quad \text{and} \quad |1 - \alpha_{\mathrm{d}}| \leq d_2 \|\Delta \mathbf{z}_Q^{\mathrm{a}}\| \,. \tag{C.33}$$

Lemma 21 (iii) implies that the assumptions for Lemma 19 are satisfied for $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho^*)$ and $Q \in \mathcal{Q}^*$. Thus, from (C.1) and (C.2), we have

$$\begin{aligned} 1 - \bar{\alpha}_{\mathrm{p}} &\leq \max\left\{0, \max_{i \in \mathcal{A}(\mathbf{x}^*)} \left\{1 - \frac{\lambda_i}{\tilde{\lambda}_i^{\mathrm{a}}}\right\}, \max_{i \in \mathcal{A}(\mathbf{x}^*)} \left\{1 - \frac{\lambda_i}{\tilde{\lambda}_i} + \frac{\Delta\lambda_i^{\mathrm{a}}}{\tilde{\lambda}_i}\right\}\right\} \\ &\leq \max\left\{0, \max_{i \in \mathcal{A}(\mathbf{x}^*)} \left\{\frac{|\Delta\lambda_i^{\mathrm{a}}|}{\tilde{\lambda}_i^{\mathrm{a}}}\right\}, \max_{i \in \mathcal{A}(\mathbf{x}^*)} \left\{\frac{|\Delta\lambda_i|}{\tilde{\lambda}_i} + \frac{|\Delta\lambda_i^{\mathrm{a}}|}{\tilde{\lambda}_i}\right\}\right\}, \end{aligned} \tag{C.34}$$

and

$$\begin{aligned} 1 - \bar{\alpha}_{\mathrm{d}} &\leq \max\left\{0, \max_{i \in Q \setminus \mathcal{A}(\mathbf{x}^*)} \left\{1 - \frac{s_i}{\tilde{s}_i^{\mathrm{a}}}\right\}, \max_{i \in Q \setminus \mathcal{A}(\mathbf{x}^*)} \left\{1 - \frac{s_i}{\tilde{s}_i} + \frac{\Delta s_i^{\mathrm{a}}}{\tilde{s}_i}\right\}\right\} \\ &\leq \max\left\{0, \max_{i \in Q \setminus \mathcal{A}(\mathbf{x}^*)} \left\{\frac{|\Delta s_i^{\mathrm{a}}|}{\tilde{s}_i^{\mathrm{a}}}\right\}, \max_{i \in Q \setminus \mathcal{A}(\mathbf{x}^*)} \left\{\frac{|\Delta s_i|}{\tilde{s}_i} + \frac{|\Delta s_i^{\mathrm{a}}|}{\tilde{s}_i}\right\}\right\}. \end{aligned} \tag{C.35}$$

Applying Lemma 21 (iii) on (C.34) and (C.35) and using (C.31) yield

$$\begin{aligned} 1 - \bar{\alpha}_{\mathrm{p}} &\leq \frac{2}{\epsilon^*} \max_{i \in \mathcal{A}(\mathbf{x}^*)} \{|\Delta\lambda_i^{\mathrm{a}}| + |\Delta\lambda_i|\} \\ &\leq \frac{2}{\epsilon^*}(\|\Delta\boldsymbol{\lambda}_Q^{\mathrm{a}}\| + \|\Delta\boldsymbol{\lambda}_Q\|) \\ &\leq \frac{2}{\epsilon^*}\left(2 + c_2\frac{\epsilon^*}{4}\right)\|\Delta\mathbf{z}_Q^{\mathrm{a}}\|, \end{aligned} \tag{C.36}$$

and

$$\begin{aligned} 1 - \bar{\alpha}_{\mathrm{d}} &\leq \frac{2}{\epsilon^*} \max_{i \in Q \setminus \mathcal{A}(\mathbf{x}^*)} \{|\Delta s_i^{\mathrm{a}}| + |\Delta s_i|\} \\ &\leq \frac{2}{\epsilon^*}\|A\|(\|\Delta\mathbf{x}^{\mathrm{a}}\| + \|\Delta\mathbf{x}\|) \\ &\leq \frac{2}{\epsilon^*}\|A\|\left(2 + c_2\frac{\epsilon^*}{4}\right)\|\Delta\mathbf{z}_Q^{\mathrm{a}}\|. \end{aligned} \tag{C.37}$$

It then follows from (4.18), Lemma 21 (iv), and (C.31) that

$$1 - \alpha_{\mathrm{p}} = 1 - \bar{\alpha}_{\mathrm{p}} + \|\Delta\mathbf{x}\| \leq d_1\|\Delta\mathbf{z}_Q^{\mathrm{a}}\|, \tag{C.38}$$

and

$$1 - \alpha_{\mathrm{d}} = 1 - \bar{\alpha}_{\mathrm{d}} + \|\Delta\mathbf{x}\| \leq d_2\|\Delta\mathbf{z}_Q^{\mathrm{a}}\|, \tag{C.39}$$

with $d_1 := \frac{2}{\epsilon^*}\left(2 + c_2\frac{\epsilon^*}{4}\right) + \left(1 + c_2\frac{\epsilon^*}{4}\right)$ and $d_2 := \frac{2}{\epsilon^*}\|A\|\left(2 + c_2\frac{\epsilon^*}{4}\right) + \left(1 + c_2\frac{\epsilon^*}{4}\right)$. The proof is then complete with $c_3 := \max\{d_1, d_2\}$. □

Lemma 25 (Corresponding to Lemma A.11 in the supplementary materials of [37]). *Under our assumptions, there exists a constant $c_4 > 0$ such that for all $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho^*)$, and for all $Q \in \mathcal{Q}^*$,*

$$\|\hat{\mathbf{z}}_Q^+ - (\mathbf{z}_Q + \Delta^N \mathbf{z}_Q)\| \leq c_4 \max\{\|\Delta^N \mathbf{z}\|^2, \|\mathbf{z} - \mathbf{z}^*\|^2\} . \tag{C.40}$$

Proof. Let $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho^*)$ and $Q \in \mathcal{Q}^*$. It follows from (C.21), (C.31), Lemmas 22, 23, and 24 that

$$\begin{aligned}
&\|\hat{\mathbf{z}}_Q^+ - (\mathbf{z}_Q + \Delta^N \mathbf{z}_Q)\| \\
&\leq (1-\alpha)\|\Delta \mathbf{z}_Q\| + \|\Delta \mathbf{z}_Q^{\mathrm{c}}\| + \|\Delta \mathbf{z}_Q^{\mathrm{a}} - \Delta^N \mathbf{z}_Q\| \\
&\leq c_3 \|\Delta \mathbf{z}_Q^{\mathrm{a}}\| \|\Delta \mathbf{z}_Q\| + c_2 \|\Delta \mathbf{z}_Q^{\mathrm{a}}\|^2 + c_1 \|\mathbf{z} - \mathbf{z}^*\| \|\Delta^N \mathbf{z}_Q\| \\
&\leq \left(c_3 \left(1 + c_2 \frac{\epsilon^*}{4} \right) + c_2 \right) \|\Delta \mathbf{z}_Q^{\mathrm{a}}\|^2 + c_1 \|\mathbf{z} - \mathbf{z}^*\| \|\Delta^N \mathbf{z}_Q\| .
\end{aligned} \tag{C.41}$$

Also, by Lemma 22, we have

$$\begin{aligned}
\|\Delta \mathbf{z}_Q^{\mathrm{a}}\| &\leq \|\Delta^N \mathbf{z}_Q\| + \|\Delta \mathbf{z}_Q^{\mathrm{a}} - \Delta^N \mathbf{z}_Q\| \\
&\leq (1 + c_1 \|\mathbf{z} - \mathbf{z}^*\|) \|\Delta^N \mathbf{z}_Q\| \\
&\leq (1 + c_1 \rho^*) \|\Delta^N \mathbf{z}_Q\|
\end{aligned} \tag{C.42}$$

Hence, the claim is proved with $c_4 = (1 + c_1\rho^*) \left(c_3 \left(1 + c_2 \frac{\epsilon^*}{4} \right) + c_2 \right) + c_1$.

□

Now, we can prove Theorem 4.

Proof of Theorem 4. Let $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho^*)$ and Q be the working set selected given $\mathbf{z}$. Recall that Lemma 18 implies $Q \in \mathcal{Q}^*$ when $\mathbf{z} \in E^o \cap B(\mathbf{z}^*, \rho^*)$, with ρ^* given

by Lemma 21. Now, define $\rho := \rho^*$, $\mathbf{t} := \mathbf{z}$, and $\mathbf{t}^* := \mathbf{z}^*$. Then the desired q-quadratic convergence is a direct consequence of Corollary 3, if the condition (C.14) is satisfied. Hence, we will then show that there exists some constant $c > 0$ such that, for each $i \in \mathcal{I}$,

$$\min\{|z_i^+ - z_i^*|, |z_i^+ - (z_i + \Delta^N z_i)|\} \leq c \max\{\|\Delta^N \mathbf{z}\|^2, \|\mathbf{z} - \mathbf{z}^*\|^2\} . \tag{C.43}$$

By the definition of $\hat{\mathbf{z}}_Q^+$ in (C.18), Lemma 25 implies that (C.43) holds for the $\mathbf{x}^+$ components of $\mathbf{z}^+$. We next show that (C.43) holds for the $\boldsymbol{\lambda}^+$ components of $\mathbf{z}^+$.

First, for all $i \in \mathcal{A}(\mathbf{x}^*)$, we show that $\lambda_i^+ = \hat{\lambda}_i$, thus (C.43) holds for all λ_i^+ such that $i \in \mathcal{A}(\mathbf{x}^*)$ by Lemma 25. From the fact that $\boldsymbol{\lambda} > \mathbf{0}$ and Lemma 21 (ii), it follows that

$$\chi := \|\Delta \mathbf{x}^{\mathrm{a}}\|^2 + \|[\tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}}]_-\|^2 \leq \|\Delta \mathbf{x}^{\mathrm{a}}\|^2 + \|\Delta \boldsymbol{\lambda}_Q^{\mathrm{a}}\|^2 \leq 2\left(\frac{\epsilon^*}{4}\right)^2 \leq \frac{\epsilon^*}{2} . \tag{C.44}$$

Also, from Lemma 21 (iii) and the fact that $\hat{\boldsymbol{\lambda}}_Q$ is a convex combination of $\boldsymbol{\lambda}_Q$ and $\tilde{\boldsymbol{\lambda}}_Q$, we have, for all $i \in \mathcal{A}(\mathbf{x}^*)$,

$$\frac{\epsilon^*}{2} < \min\{\lambda_i, \tilde{\lambda}_i\} \leq \hat{\lambda}_i . \tag{C.45}$$

Hence, from (C.44), (C.45), and (4.20), we conclude that $\lambda_i^+ = \hat{\lambda}_i$ for all $i \in \mathcal{A}(\mathbf{x}^*)$.

Next, for $i \in Q \setminus \mathcal{A}(\mathbf{x}^*)$, we prove the following inequality

$$\|\boldsymbol{\lambda}^+_{Q \setminus \mathcal{A}(\mathbf{x}^*)}\| \leq d_1 \max\{\|\Delta^N \mathbf{z}\|^2, \|\mathbf{z} - \mathbf{z}^*\|^2\} , \tag{C.46}$$

where $d_1 > 0$ is a constant. For $i \in Q \setminus \mathcal{A}(\mathbf{x}^*)$, we know from (4.20) that, either $\lambda_i^+ = \hat{\lambda}_i$, or $\lambda_i^+ = \min\{\underline{\lambda}, \|\Delta \mathbf{x}^{\mathrm{a}}\|^2 + \|[\tilde{\boldsymbol{\lambda}}_Q^{\mathrm{a}}]_-\|^2\}$. In the former case, from the fact

that $\boldsymbol{\lambda}^*_{Q\setminus\mathcal{A}(\mathbf{x}^*)} = 0$ (by Assumption 7), we have

$$\begin{aligned} |\lambda_i^+| = |\hat{\lambda}_i| &= |\hat{\lambda}_i - \lambda_i^*| \\ &\leq |\hat{\lambda}_i - (\lambda_i + \Delta^N\lambda_i)| + |(\lambda_i + \Delta^N\lambda_i) - \lambda_i^*| \\ &\leq d_2 \max\{\|\Delta^N\mathbf{z}\|^2, \|\mathbf{z}-\mathbf{z}^*\|^2\} + d_3\|\mathbf{z}-\mathbf{z}^*\|^2\,, \end{aligned} \tag{C.47}$$

where the last inequality follows from Lemma 25 and the quadratic rate of the Newton step given in Proposition 8. In the latter case, since $\boldsymbol{\lambda} > \mathbf{0}$, we obtain the following inequality

$$\begin{aligned} |\lambda_i^+| &\leq \|\Delta\mathbf{x}^{\mathrm{a}}\|^2 + \|[\tilde{\boldsymbol{\lambda}}^{\mathrm{a}}_Q]_-\|^2 \\ &\leq \|\Delta\mathbf{x}^{\mathrm{a}}\|^2 + \|\Delta\boldsymbol{\lambda}^{\mathrm{a}}_Q\|^2 = \|\Delta\mathbf{z}^{\mathrm{a}}_Q\|^2 \\ &\leq (1 + c_1\rho^*)\|\Delta^N\mathbf{z}_Q\|, \end{aligned} \tag{C.48}$$

where the equality is directly from the definition of $\Delta\mathbf{z}^{\mathrm{a}}$, and the last inequality follows from (C.42). Hence, we have established (C.46), which, together with the fact that $\boldsymbol{\lambda}^*_{Q\setminus\mathcal{A}(\mathbf{x}^*)} = 0$ (by Assumption 7), implies that (C.43) holds for all $i \in Q\setminus\mathcal{A}(\mathbf{x}^*)$.

Finally, let us consider the case that $i \in \mathcal{I}\setminus Q$. From Assumption 7 we have $\boldsymbol{\lambda}^*_{\mathcal{I}\setminus Q} = \mathbf{0}$, thus it follows from (4.21) that

$$|\lambda_i^+ - \lambda_i^*| = |\lambda_i^+| \leq \mu^+_{(Q)}/s_i^+ \tag{C.49}$$

By definition, $s_i^+ := s_i + \alpha_{\mathrm{p}}\Delta s_i$ is a convex combination of s_i and $\tilde{s}_i$. Thus, Lemma 21 (iii) gives that $s_i^+ \geq \min\{s_i, \tilde{s}_i\} > \epsilon^*/2$. Then using the definition of $\mu^+_{(Q)}$ and the fact that $|Q| \geq 1$, (C.49) then becomes

$$|\lambda_i^+ - \lambda_i^*| \leq \frac{2}{\epsilon^*}\left((\mathbf{s}^+_{\mathcal{A}(\mathbf{x}^*)})^T(\boldsymbol{\lambda}^+_{\mathcal{A}(\mathbf{x}^*)}) + (\mathbf{s}^+_{Q\setminus\mathcal{A}(\mathbf{x}^*)})^T(\boldsymbol{\lambda}^+_{Q\setminus\mathcal{A}(\mathbf{x}^*)})\right)\,. \tag{C.50}$$

Since $\mathbf{z}$ is in $B(\mathbf{z}^*, \rho^*)$, $\boldsymbol{\lambda}^+_{\mathcal{A}(\mathbf{x}^*)}$ and $\mathbf{s}^+_{Q\setminus\mathcal{A}(\mathbf{x}^*)}$ are bounded by Lemma 21 (ii). Also, by definition, $\mathbf{s}^*_{\mathcal{A}(\mathbf{x}^*)} = \mathbf{0}$. Then from (C.50), there exist some constant d_4 and d_5 such that

$$|\lambda_i^+ - \lambda_i^*| \leq d_4 \|\mathbf{s}^+_{\mathcal{A}(\mathbf{x}^*)} - \mathbf{s}^*_{\mathcal{A}(\mathbf{x}^*)}\| + d_5 \|\boldsymbol{\lambda}^+_{Q\setminus\mathcal{A}(\mathbf{x}^*)}\| \,. \tag{C.51}$$

Note that the second term in (C.51) is bounded by (C.46), and we are left to prove that the first term in (C.51) is properly bounded. By definition, the first term in (C.51) is bounded by

$$\|\mathbf{s}^+_{\mathcal{A}(\mathbf{x}^*)} - \mathbf{s}^*_{\mathcal{A}(\mathbf{x}^*)}\| \leq \|A\mathbf{x}^+ - A\mathbf{x}^*\| \leq \|A\| \|\hat{\mathbf{z}}^+_Q - \mathbf{z}^*_Q\| \,. \tag{C.52}$$

Let $d_6 := \|A\|$, (C.52) can be further bounded by

$$\begin{aligned} \|\mathbf{s}^+_{\mathcal{A}(\mathbf{x}^*)} - \mathbf{s}^*_{\mathcal{A}(\mathbf{x}^*)}\| &\leq d_6 \|\hat{\mathbf{z}}^+_Q - \mathbf{z}^*_Q\| \\ &\leq d_6 \|\hat{\mathbf{z}}^+_Q - (\mathbf{z}_Q + \Delta^N \mathbf{z}_Q)\| + d_6 \|(\mathbf{z}_Q + \Delta^N \mathbf{z}_Q) - \mathbf{z}^*_Q\| \\ &\leq d_7 \max\{\|\Delta^N \mathbf{z}\|^2, \|\mathbf{z} - \mathbf{z}^*\|^2\} + d_8 \|\mathbf{z} - \mathbf{z}^*\|^2 \,, \end{aligned} \tag{C.53}$$

where the second inequality uses triangle inequality, and the third inequality follows from Lemma 25 and the quadratic rate of the Newton step given in Proposition 8. Hence, we established (C.43) for all $i \in \mathcal{I}$, and the q-quadratic convergence is obtained. □

Bibliography

[1] C. Cercignani. *The Boltzmann Equation and Its Applications*, volume 67 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1988.

[2] C. Cercignani, R. Illner, and M. Pulvirenti. *The Mathematical Theory of Dilute Gases*, volume 106 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1994.

[3] E. E. Lewis and Jr. W. F. Miller. *Computational Methods in Neutron Transport*. John Wiley and Sons, New York, 1984.

[4] G. C. Pomraning. *Radiation Hydrodynamics*. Pergamon Press, New York, 1973.

[5] R. Dautray and J. L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology, Volume 6: Evolution Problems II*. Spinger-Verlag, Berlin, 2000.

[6] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser. *Semiconductor Equations*. Springer-Verlag, New York, 1990.

[7] T. A. Brunner and J. P. Holloway. One-dimensional Riemann solvers and the maximum entropy closure. *J. Quant. Spectrosc. Ra.*, 69(5):543 – 566, 2001.

[8] C. D. Hauck. High-order entropy-based closures for linear transport in slab geometry. *Comm. Math. Sci.*, 9, 2011.

[9] C. D. Hauck and R. G. McClarren. Positive P_N closures. *SIAM J. Sci. Comput.*, 32(5):2603–2626, 2010.

[10] C. K. Garrett and C. D. Hauck. A comparison of moment closures for linear kinetic transport equations: the line source benchmark. *Transport Theor. Stat.*, 42:203 – 235, 2013.

[11] R. G. McClarren and C. D. Hauck. Robust and accurate filtered spherical harmonics expansions for radiative transfer. *J. Comput. Phys.*, 229(16):5597 – 5614, 2010.

[12] K. Case and P. Zweifel. *Linear Transport Theory.* Addison-Wesley, Reading, MA, 1967.

[13] G. N. Minerbo. Maximum entropy Eddington factors. *J. Quant. Spectrosc. Ra.*, 20:541–545, 1978.

[14] B. Dubroca and J.-L. Fuegas. Étude théorique et numérique d'une hiérarchie de modèles aus moments pour le transfert radiatif. *C.R. Acad. Sci. Paris*, I. 329:915–920, 1999.

[15] G. W. Alldredge, C. D. Hauck, and A. L. Tits. High-order entropy-based closures for linear transport in slab geometry II: A computational study of the optimization problem. *SIAM J. Sci. Comput.*, 34(4):B361–B391, 2012.

[16] G. W. Alldredge, C. D. Hauck, D. P. O'Leary, and A. L. Tits. Adaptive change of basis in entropy-based moment closures for linear kinetic equations. *J. Comput. Phys.*, 258:489–508, 2014.

[17] C. K. Garrett, C. Hauck, and J. Hill. Optimization and large scale computation of an entropy-based moment closure. *J. Comput. Phys.*, 302:573 – 590, 2015.

[18] D. Radice, E. Abdikamalov, L. Rezzolla, and C. D. Ott. A new spherical harmonics scheme for multi-dimensional radiation transport I: Static matter configurations. *J. Comput. Phys.*, 242(0):648 – 669, 2013.

[19] R. G. McClarren, J. P. Holloway, and T. A. Brunner. On solutions to the P_N equations for thermal radiative transfer. *J. Comput. Phys.*, 227(5):2864–2885, 2008.

[20] G. L. Olson. Second-order time evolution of P_N equations for radiation transport. *J. Comput. Phys.*, 228(8):3072–3083, May 2009.

[21] X. D. Liu and S. Osher. Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes I. *SIAM J. Numer. Anal.*, 33(2):pp. 760–779, 1996.

[22] X. Zhang and C. W. Shu. On maximum-principle-satisfying high order schemes for scalar conservation laws. *J. Comput. Phys.*, 229(9):3091 – 3120, 2010.

[23] X. Zhang and C. W. Shu. Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. *Proc. Roy. Soc. London A: Math., Phys. and Eng. Sci.*, 467(2134):2752–2776, 2011.

[24] B. D. Ganapol. Homogeneous infinite media time-dependent analytic benchmarks for X-TM transport methods development. Technical report, Los Alamos National Laboratory, March 1999.

[25] S. Jin. Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review. *Lecture Notes for Summer School on "Methods and Models of Kinetic Theory" (M&MKT), Porto Ercole (Grosseto, Italy)*, 2010.

[26] E. W. Larsen and J. B. Keller. Asymptotic solution of neutron transport problems for small mean free paths. *J. Math. Phys.*, 15:75–81, January 1974.

[27] C. Bardos, F. Golse, and D. Levermore. Fluid dynamic limits of kinetic equations. I. Formal derivations. *J. Stat. Phys.*, 63(1-2):323–344, 1991.

[28] C. Berthon and R. Turpault. Asymptotic preserving HLL schemes. *Numer. Meth. Part. D. E.*, 27(6):1396–1422, 2011.

[29] P. Lafitte and G. Samaey. Asymptotic-preserving projective integration schemes for kinetic equations in the diffusion limit. *SIAM J. Sci. Comput.*, 34(2):A579–A602, 2012.

[30] L. Mieussens. On the asymptotic preserving property of the unified gas kinetic scheme for the diffusion limit of linear kinetic models. *J. Comput. Phys.*, 253:138–156, November 2013.

[31] S. Jin. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM J. Sci. Comput.*, 21(2):441–454, 1999.

[32] S. Jin, L. Pareschi, and G. Toscani. Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations. *SIAM J. Numer. Anal.*, 35(6):2405–2439, 1998.

[33] S. Jin, L. Pareschi, and G. Toscani. Uniformly accurate diffusive relaxation schemes for multiscale transport equations. *SIAM J. Numer. Anal.*, 38(3):913–936, 2000.

[34] K. Küpper, M. Frank, and S. Jin. An asymptotic preserving two-dimensional staggered grid method for multiscale transport equations. *SIAM J. Numer. Anal.*, 54(1):440–461, 2016.

[35] M. Lemou and L. Mieussens. A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit. *SIAM J Sci. Comput.*, 31(1):334–368, 2008.

[36] S. J. Wright. *Primal-Dual Interior-Point Methods*. SIAM, 1997.

[37] L.B. Winternitz, S.O. Nicholls, A.L. Tits, and D.P. O'Leary. A constraint-reduced variant of Mehrotra's predictor-corrector algorithm. *Comput. Optim. Appl.*, 51(1):1001 – 1036, 2012.

[38] A.L. Tits, P.A. Absil, and W.P. Woessner. Constraint reduction for linear programs with many inequality constraints. *SIAM J. Optimiz.*, 17(1):119 – 146, 2006.

[39] L. B. Winternitz, A. L. Tits, and P.-A. Absil. Addressing rank degeneracy in constraint-reduced interior-point methods for linear optimization. *J. Optimiz. Theory App.*, 160(1):127–157, 2014.

[40] J.H. Jung, D.P. O'Leary, and A.L. Tits. Adaptive constraint reduction for convex quadratic programming. *Comput. Optim. Appl.*, 51(1):125 – 157, 2012.

[41] J. H. Jung, D. P. O'Leary, and A. L. Tits. Adaptive constraint reduction for training support vector machines. *Electron. T. Numer. Ana.*, 31:156–177, 2008.

[42] S. Park and D. P. O'Leary. A polynomial time constraint-reduced algorithm for semidefinite optimization problems. *J. Optimiz. Theory App.*, 166(2):558–571, 2015.

[43] S. Park. A constraint-reduced algorithm for semidefinite optimization problems with superlinear convergence. *J. Optimiz. Theory App.*, pages 1–16, 2016.

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

[45] D. Gottlieb, S. Gottlieb, and J. Hesthaven. *Spectral Methods for Time-Dependent Problems*. Cambridge University Press, New York, 2007.

[46] B. Guo. *Spectral Methods and Their Applications*. World Scientific, Singapore, 1998.

[47] M. Frank, C. Hauck, and K. Küpper. Convergence of filtered spherical harmonic equations for radiation transport. *Commun. Math. Sci.*, 14(5):1443–1465, 2016.

[48] K. Atkinson. Numerical integration on the sphere. *J. Austral. Math. Soc. Ser. B*, 23:332–347, 1982.

[49] W. Walters. Use of the Chebyshev-Legendre quadrature set in discrete-ordinate codes. Technical Report LA-UR-87-3621, Los Alamos National Laboratory, 1987.

[50] V.I. Lebedev. Quadratures on a sphere. *Comput. Math. Math. Phys.*, 16:10–24, 1976.

[51] V.I. Lebedev. A quadrature formula for the sphere of 59th algebraic order of accuracy. *Russian Acad. Sci. Dokl. Math.*, 50:283–286, 1995.

[52] V.I. Lebedev and A.L. Skorokhodov. Quadrature formulas of orders 41, 47, and 53 for the sphere. *Russian Acad. Sci. Dokl. Math.*, 45:587–592, 1992.

[53] V.I. Lebedev. Spherical quadrature formulas exact to orders 25–29. *Sib. Math. J.*, 18(1):99–107, 1977.

[54] V.I. Lebedev and D.N. Laikov. A quadrature formula for the sphere of the 131st algebraic order of accuracy. In *Doklady. Mathematics*, volume 59, pages 477–481. MAIK Nauka/Interperiodica, 1999.

[55] E. Di Nezza, G. Palatucci, and E. Valdinoci. Hitchhiker's guide to the fractional Sobolev spaces. *Bull. Sci. Math.*, 136(5):521 – 573, 2012.

[56] F. Dai and Y. Xu. *Approximation Theory and Harmonic Analysis on Spheres and Balls*. Springer Monographs in Mathematics. Springer New York, 2013.

[57] F. Dai and Y. Xu. Polynomial approximation in Sobolev spaces on the unit sphere and the unit ball. *J. of Approx. Theory*, 163(10):1400 – 1418, 2011.

[58] A. Quarteroni. Some results of Bernstein and Jackson type for polynomial approximation in L^p-spaces. *Jpn. J. Appl. Math.*, 1(1):173–181, 1984.

[59] R. A. DeVore and G. G. Lorentz. *Contructive Approximation*. Springer-Verlag, 1993.

[60] P. Garrett. Harmonic analysis on spheres, II. 2011. Available at http://www.math.umn.edu/~garrett/m/mfms/notes_c/spheres_II.pdf.

[61] C. Canuto and A. Quarteroni. Approximation results for orthogonal polynomials in Sobolev spaces. *Math. Comp.*, 38(157):pp. 67–86, 1982.

[62] T. A. Brunner. Forms of approximate radiation transport. Technical Report SAND2002-1778, Sandia National Laboratories, 2002.

[63] G. J. Habetler and B. J. Matkowsky. Uniform asymptotic expansions in transport theory with small mean free paths, and the diffusion approximation. *J. Math. Phys.*, 16:846–854, April 1975.

[64] H. Egger and M. Schlottbom. A mixed variational framework for the radiative transfer equation. *Math. Mod. Meth. Appl. S.*, 22(03):1150014, 2012.

[65] W. F. Miller. An analysis of the finite differenced, even-parity, discrete ordinates equations in slab geometry. *Nucl. Sci. Eng.*, 108(3):247–266, 1991.

[66] B. Seibold and M. Frank. Starmap—a second order staggered grid method for spherical harmonics moment equations of radiative transfer. *ACM T. Math. Software*, 41(1):4, 2014.

[67] P. Degond and M. Tang. All speed scheme for the low mach number limit of the isentropic Euler equations. *Commun. Comput. Phys.*, 10(01):1–31, 2011.

[68] C. Parés. Numerical methods for nonconservative hyperbolic systems: a theoretical framework. *SIAM J. Numer. Anal.*, 44(1):300–321, 2006.

[69] G. Dal Maso, P. G. Lefloch, and F. Murat. Definition and weak stability of nonconservative products. *J. Math. Pures Appl.*, 74(6):483–548, 1995.

[70] R. Abgrall and S. Karni. A comment on the computation of non-conservative products. *J. Comput. Phys.*, 229(8):2759 – 2763, 2010.

[71] H. Nessyahu and E. Tadmor. Non-oscillatory central differencing for hyperbolic conservation laws. *J. Comput. Phys.*, 87(2):408 – 463, 1990.

[72] B. Van Leer. Towards the ultimate conservative difference scheme. IV. a new approach to numerical convection. *J. Comput. Phys.*, 23(3):276 – 299, 1977.

[73] S. Gottlieb, C. W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Rev.*, 43(1):pp. 89–112, 2001.

[74] J. H. Jung. *Adaptive Constraint Reduction for Convex Quadratic Programming and Training Support Vector Machines*. PhD thesis, University of Maryland, 2008.