12-31-2019

Convex relaxations of a continuum aggregation model, and their efficient numerical solution

Mahdi Bandegi
New Jersey Institute of Technology

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

CONVEX RELAXATIONS OF A CONTINUUM AGGREGATION MODEL, AND THEIR EFFICIENT NUMERICAL SOLUTION

by
Mahdi Bandegi

In this dissertation, the global minimization of a large deviations rate function (the Helmholtz free energy functional) for the Boltzmann distribution is discussed. The Helmholtz functional arises in large systems of interacting particles — which are widely used as models in computational chemistry and molecular dynamics. Global minimizers of the rate function (Helmholtz functional) characterize the asymptotics of the partition function and thereby determine many important physical properties such as self-assembly, or phase transitions. Finding and verifying local minima to the Helmholtz free energy functional is relatively straightforward. However, finding and verifying global minima is much more difficult since the Helmholtz energy is nonconvex and nonlocal. Instead of minimizing the original nonconvex functional, the approach in this dissertation is to find minimizers to a convex lower bound functional. The so-called relaxed problem consists of a linear variational problem with an infinite number of Fourier constraints, leading to a variety of computational challenges. A fast solver (for the relaxed problem) based on matrix-free interior-point algorithms is developed by exploiting the Fourier structure in the problem in conjunction with a new preconditioner.

CONVEX RELAXATIONS OF A CONTINUUM AGGREGATION MODEL, AND THEIR EFFICIENT NUMERICAL SOLUTION

by
Mahdi Bandegi

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Mathematical Sciences

Department of Mathematical Sciences
Department of Mathematics and Computer Science, Rutgers-Newark

December 2019

APPROVAL PAGE

CONVEX RELAXATIONS OF A CONTINUUM AGGREGATION MODEL, AND THEIR EFFICIENT NUMERICAL SOLUTION

Mahdi Bandegi

David G. Shirokoff, Dissertation Advisor — Date
Assistant Professor of Mathematics, New Jersey Institute of Technology

---

Cyrill B. Muratov, Committee Member — Date
Professor of Mathematics, New Jersey Institute of Technology

Brittany Froese Hamfeldt, Committee Member — Date
Assistant Professor of Mathematics, New Jersey Institute of Technology

---

Andrew J. Bernoff, Committee Member — Date
Professor of Mathematics, Harvey Mudd College

Travis L. Askham, Committee Member — Date
Assistant Professor of Mathematics, New Jersey Institute of Technology

BIOGRAPHICAL SKETCH

Author: Mahdi Bandegi

Degree: Doctor of Philosophy

Date: December 2019

Undergraduate and Graduate Education:

- Doctor of Philosophy in Mathematical Sciences,
  New Jersey Institute of Technology, Newark, NJ, 2019
- Master of Science in General Mathematics
  Western Kentucky University, Bowling Green, KY, 2014
- Master of Science in Structural Engineering
  Ferdowsi University of Mashhad, Iran, 2010
- Bachelors of Science in Civil Engineering
  Ferdowsi University of Mashhad, Iran, 2008

Major: Mathematical Sciences

Presentations and Publications:

M. Bandegi and D. Shirokoff, Approximate global minimizers to pairwise interaction problems via convex relaxation, *SIAM Journal on Applied Dynamical Systems*, 17(1):417–456, 2018.

M. Bandegi*, D. Shirokoff, "Efficient Solvers for Some Conic Variational Problems in pattern formation," *Poster Presentation, Dana Knox Student Showcase, New Jersey Institute of Technology*, Newark, NJ, April 2019.

M. Bandegi*, D. Shirokoff, "Convex Relaxations for Variational Problems Arising from Self-Assembly," *Poster Presentation, Mid-Atlantic Numerical Analysis Day, Temple University*, Philadelphia, PA, November 2018.

M. Bandegi*, D. Shirokoff, "Convex Relaxations for Variational Problems Arising from Self-Assembly," *Poster Presentation, Princeton Optimization Day, Princeton University*, Princeton, NJ, September 2018.

M. Bandegi*, D. Shirokoff, "Efficient Solvers for Some Conic Variational Problems," *Poster Presentation, FACM, New Jersey Institute of Technology*, Newark, NJ, August 2018.

M. Bandegi*, D. Shirokoff, "Efficient Solvers for Some Conic Variational Problems," *Poster Presentation, SIAM annual meeting*, Portland, OR, July 2018.

M. Bandegi*, D. Shirokoff, "Conic programming of a variational inequality motivated from self-assembly," *Poster Presentation, Dana Knox Student Showcase, New Jersey Institute of Technology*, Newark, NJ, April 2018.

M. Bandegi*, D. Shirokoff, "Approximate Global Minimizers for Pairwise Interaction Problem," *Poster Presentation, FACM, New Jersey Institute of Technology*, Newark, NJ, June 2016.

* Presenting author.

To the Memory of My Father.

ACKNOWLEDGMENT

I would like to express my deepest appreciation to my dissertation advisor, Dr. David Shirokoff, without whom writing this thesis would not be possible. Dave has not only being a mentor with immense patience but also a great friend.

I am grateful to have Dr. Cyrill Muratov, Dr. Brittany Hamfeldt, Dr. Andrew Bernoff and Dr. Travis Askham in my committee, for their comments and help through writing this dissertation. I am also thankful to the Department of Mathematical Sciences for giving me the opportunity to pursue my Ph.D. degree at New Jersey Institute of Technology.

I would like to acknowledge all my friends at NJIT, especially Andrew, Malik, Matt, Pejman, Atefeh, Soheil and all other fellow Ph.D. students in the math department with whom I shared lots of good memories. I will keep your friendship forever.

Finally, I have to thank my family, my mother for her boundless love and encouragement, and my siblings, Sanaz and Mehrdad, who helped me a lot through my life. It has been a quiet struggle to be far from my family but I am so lucky to have their support by myself.

TABLE OF CONTENTS

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

This dissertation develops theory and numerical methods for computing global minimizers (ground states) to variants of the Helmholtz free energy functional.

In Chapter 2, we introduce the Helmholtz free energy functional, and present suitable conditions under which minimizers to the Helmholtz free energy characterize the long-time behavior for large systems of interacting particles undergoing Brownian motion (which are models used in molecular dynamics). Specifically, the Helmholtz energy arises as a large deviations rate function for the Boltzmann distribution (of discrete particle models). Minimizers of the Helmholtz energy provide information on the asymptotics of the Boltzmann distribution, and can characterize phenomena such as phase transitions and self-assembly.

In Chapter 3, we formulate sufficient conditions for global minimizers to the Helmholtz energy functional. The sufficient conditions are formulated as a lower bound obtained through a convex relaxation of the original nonconvex energy. The conditions take the form of a linear variational problem, and have the additional advantage of, in some cases, being exact. Recently, many works in the optimization community formulate similar convex relaxation approaches using methods such as sums-of-squares programming and semi-definite programming, however, these approaches have been primarily for finite dimensional problems.

Chapter 4 discusses some properties of solutions to the standard form of linear programming.

Chapter 5 focuses on developing numerical techniques to solve the sufficient conditions developed in Chapter 3. Numerical discretizations of the sufficient conditions take the form of linear programming problems with (a large number of)

Fourier mode constraints. To enable the solution of the large linear programming problems, we adopt a matrix-free, primal-dual interior-point algorithm. A central issue in the matrix-free[1] approach is ill-conditioning due to the interior-point method. We introduce a simple preconditioner to alleviate the ill-conditioning. We then compare the performance of the preconditioner, to other approaches in the literature, for a few of our variational problems. We observe that the preconditioner outperforms other approaches (by requiring fewer matrix-vector products and floating point operations).

[1]Matrix-free in a sense that they do not need to build and store matrices — only matrix vector products are needed.

CHAPTER 2

BACKGROUND ON PAIRWISE INTERACTION MODELS

In this chapter we introduce the Helmholtz functional and motivate the computation of ground states (global minimizers). We first introduce discrete pairwise interaction models and then discuss how the Helmholtz free energy arises as a large deviations rate function to the corresponding Boltzmann distribution. We conclude the section with examples from molecular dynamics where the discrete particle solutions are known to sample the Boltzmann distribution. When the large deviations result holds, minimizers to the Helmholtz rate function provide information on the long-time molecular/stochastic dynamics (such as phase transitions and self-assembly) of large interacting particle systems.

2.1 Pairwise Interaction Energy in Many Particle Systems

The total energy for a system of n identical particles in spatial dimension d, with positions $\mathbf{x}_i \in \mathbb{R}^d$ can be written as [44]

$$E(\mathbf{x}_1, \dots, \mathbf{x}_n) = \frac{1}{2} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} w(\mathbf{x}_i - \mathbf{x}_j) + n \sum_{i=1}^{n} u(\mathbf{x}_i). \tag{2.1}$$

Here $w(\mathbf{r})$ is referred to as the *interaction energy* between two pairs of particles. Meanwhile $u(x)$ is an *external potential* felt by all particles. The incorporation of the factor of n in front of the external potential (2.1) is done so that the summations in (2.1) involving $w(x)$ and $u(x)$ have similar (order of magnitude) contributions to the energy as $n \to \infty$.

We can write the continuum model (at zero temperature), analogous to the discrete energy (2.1), using the following energy functional [3]

$$\mathcal{E}(\rho) = \frac{1}{2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \rho(\mathbf{x}) w(\mathbf{x} - \mathbf{y}) \rho(\mathbf{y}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} + \int_{\mathbb{R}^d} u(\mathbf{x}) \rho(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{2.2}$$

Note that $\rho(\mathbf{x})\mathrm{d}\mathbf{x}$ in equation (2.2) is the fraction of particles in the region $\mathrm{d}\mathbf{x}$, and will be considered as a probability measure. Without a loss of generality, the total mass m of $\rho(\mathbf{x})\mathrm{d}\mathbf{x}$ is taken to be 1:

$$m := \int_{\mathbb{R}^d} \rho(\mathbf{x})\mathrm{d}\mathbf{x} = 1. \tag{2.3}$$

In Equation (2.2) the double integral weights the energy of $\rho(\mathbf{x})\mathrm{d}\mathbf{x}$ particles and $\rho(\mathbf{y})\mathrm{d}\mathbf{y}$ particles by the interaction cost $w(\mathbf{x}-\mathbf{y})$.

2.2 The Helmholtz Functional as a Large Deviations Rate Function

In this section we address the motivation to study the Helmholtz free energy functional and its global minimizers. The relation between the discrete form of the total energy (2.1) and the energy functional (2.2) can be shown using the *Large Deviation Principles* (LDP) [27, 52], furthermore, applying *Mean-field* results, the importance of finding the global minimizers of (2.2) in approximating the *Gibbs-Boltzmann distribution* will be shown [52].

The Gibbs-Boltzmann distribution gives the probability of the system being in state n as

$$\mathbb{P}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) = Z^{-1}\exp\big(-\beta E(\mathbf{x}_1, \ldots, \mathbf{x}_n)\big), \tag{2.4}$$

where $Z = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \exp\big(-\beta E(\mathbf{x}_1, \ldots, \mathbf{x}_n)\big)\mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_n$, and $\beta = (k_b T)^{-1}$ is the inverse temperature (with T being the temperature, and k_b being the Boltzmann constant).

The large deviations principle provides the asymptotics of $\mathbb{P}_n$ in the limit as $n \gg 1$. Assumptions that guarantee a large deviations principle, adopted from [16] are:

A1. $w(\mathbf{x})$ is continuous; except possibly at $\mathbf{0}$ where $w(\mathbf{0}) = +\infty$.

A2. $u(\mathbf{x})$ is continuous; and $u(\mathbf{x}) > c\|\mathbf{x}\|$ at large $\mathbf{x}$ (with $c > 0$).

A3. $w(\mathbf{x}) + u(\mathbf{x})$ is bounded from below.

A4. $\mathcal{E}(\rho)$ is weakly continuous (at points $\rho(\mathbf{x})$ where $\mathcal{E}(\rho) < \infty$).

For $\mathbb{P}$ under the assumptions (A1)-(A4), and fixed $G \subseteq (\mathbb{R}^d)^n$ and symmetric, where $n \gg 1$, we have the following large deviations principle [16]

$$Z^{-1} \int_G e^{-\beta E(\mathbf{x}_1,\ldots,\mathbf{x}_n)} \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_n \;\approx\; \begin{array}{ll} \text{maximize} & e^{-\beta n^2 \left(\mathcal{E}(\rho_n) - \mathcal{E}(\rho_0)\right)} \\ \text{subject to} & \rho_n \in G. \end{array} \tag{2.5}$$

Here, $\rho_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \delta(\mathbf{x} - \mathbf{x}_i)$ is the empirical measure where $\rho_n \in G$ means (with an abuse of notation) that $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in G$. The density $\rho_0(\mathbf{x})$ is the minimizer for (2.2), i.e., $\rho_0(\mathbf{x}) = \operatorname{argmin} \mathcal{E}(\rho)$ taken over the (larger) space of probability measures [1]. From the large deviation principle (2.5) we can see the rate function is the Helmholtz functional (2.2), which describes the importance of finding minimizers to the energy functional (2.2). Specifically, (2.5) shows that configurations of particles $\rho_n(\mathbf{x})$ that are not close to $\rho_0(\mathbf{x})$ are (exponentially) unlikely to occur. As stated in (2.5), the rate function is (2.2) and is applicable when β is held constant as $n \to \infty$ (corresponding to a low temperature model). For high temperatures, (achieved by letting $\beta \to 0$), additional terms, such as an entropy term $\beta^{-1} \int \rho \log \rho \; \mathrm{d}\mathbf{x}$, are added to the rate function $\mathcal{E}$ [16, 29].

2.3 Brownian Motion and Langevin Dynamics

In this section, we review a few aspects and examples of Brownian motion, and Langevin dynamics. The purpose is to demonstrate that the discrete energy (2.1) arises in a variety of stochastic models that sample the Boltzmann distribution. Hence, when the Boltzmann distribution admits a large deviations principle, the

[1] When $w(\mathbf{0})$ is infinite, $\mathcal{E}(\rho_n)$ is also infinite, so there is a technical modification of (2.5) to a renormalized energy [16, 52]

(continuum) Helmholtz energy (2.2) can be used to gain insight into the long-time behavior of Brownian motion.

Brownian motion refers to the random motion of particles immersed in a fluid (such as air or water) [45]. The random motion of microscopic particles was originally observed by Robert Brown in 1827 under microscopical observations on plant pollen of the plant. A mathematical theory modeling the density or probability of a particle undergoing Brownian motion (through the use of the diffusion equation) was first presented by Einstein in 1905 [28]. A subsequent set of differential equations used to model Brownian motion were then introduced by Langevin in 1908 [36]. The Langevin equation that describes the Brownian motion of a particle $\mathbf{X} \in \mathbb{R}^3$, with mass m, under the application of a force field $\mathbf{F} \in \mathbb{R}^3$ is:

$$m\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{X} = \mathbf{F}(\mathbf{X}) - \gamma\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{X} + \zeta(t). \tag{2.6}$$

Here $\gamma > 0$ is the friction constant and $\zeta(t)$ is the random force. As we can see from the right-hand side of (2.6), the Langevin equation contains both a frictional force term $\gamma\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{X}$ and stochastic term $\zeta(t)$. Note that $\zeta(t)$ is not a single unique function, i.e., $\zeta(t)$ and $\zeta(t')$ are independent whenever $t \neq t'$, which makes (2.6) a stochastic differential equation (SDE), and not an ordinary differential equation (ODE) [43,62]. In the case when frictional forces are much larger than inertial forces, i.e., $|\gamma\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{X}| \gg |m\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{X}|$ [39], the second order equation (2.6) is approximated by the overdamped Langevin equation. The overdamped Langevin equations for n particles in the presence of thermal noise with a conservative force then becomes:

$$d\mathbf{x}_j = \mathbf{F}_j\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}\eta, \qquad \mathbf{F}_j = -\nabla_j E(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n). \tag{2.7}$$

Here $\mathrm{d}\eta$ is a (Brownian) noise term, and E is the potential, which is often taken to be of the form (2.1)). In equation (2.7) we have taken $\mathbf{F}_j = -\nabla_j E(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ to be a conservative force governed by the potential $E(\mathbf{x}_1, \cdots, \mathbf{x}_n)$.

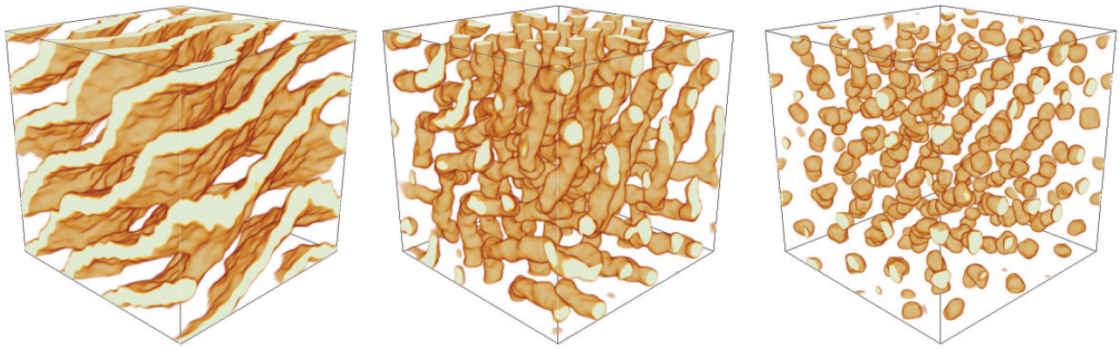

Figure 2.1 Structures observed in models (images from [51]) of nuclear matter using the energy (2.8): (l-r) particle densities are $0.05, 0.025, 0.01$ fm^{-3}, The shading show a density isosurface to highlight phase separation between the two species of particles.

We now collect several examples from the literature that use variations of (2.1).

Example 1. *(Nuclear matter)*

Astrophysicists have been long interested in studying the properties of neutron-rich matter[2]*. Neutron-rich matter can exhibit complex structures as the result of a competition between attractive nuclear forces and repulsive Coulomb (electromagnetic) forces [13].*

For example, Figure 2.1 shows three different nuclear pasta *formations for various densities of matter. With new advancements in computational power, different varieties of* nuclear pasta *have been recently identified [13].*

One interaction energy used to model nuclear matter, for a charge-neutral system of neutrons, protons, and electrons is given by the following

$$V_{total} = \frac{1}{2} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} V_{ij}. \tag{2.8}$$

In this model there are two species of particles, neutrons and protons. Particle i is represented by a position and a label of proton or neutron. The interaction energy,

[2] *Neutron-rich matter is a neutral system composed of a neutron enriched mixture of neutrons and protons embedded in a degenerate electron gas [1].*

Table 2.1 Nuclear interaction parameters [51] used in the model (2.8)

a (MeV)	b (MeV)	c (MeV)	Λ (fm^2)
110	−26	24	1.25

further can be written as $V_{ij} = V_{ij}^n + V_{ij}^c$, *where* V_{ij}^n *is a nuclear interparticle force and* V_{ij}^c *is a Coulomb force [35]:*

- *The nuclear component*

$$V_{ij}^n = ae^{-\frac{r_{ij}^2}{\Lambda}} + [b + c\tau_z(i)\tau_z(j)]e^{-\frac{r_{ij}^2}{2\Lambda}},$$

  where, r_{ij} *is the distance between two particles* i *and* j, $\tau_z = +1$ *or* (-1) *is the isospin projection of the proton (or neutron) particle. In addition,* a *is the strength of the short-range repulsion between nucleons,* b *and* c *are the strength of their intermediate-range attraction, and* Λ *is the length scale of the nuclear potential. Typical parameter values are shown in Table 2.1.*

- *The Coulomb component*

$$V_{ij}^c = \frac{\alpha}{r_{ij}} e^{-\frac{r_{ij}^2}{\lambda}} \tau_p(i)\tau_p(j),$$

  where α *is the fine structure constant,* λ *is the screening length which is fixed* $(\lambda = 10)$, *and* $\tau_p \equiv \dfrac{1+\tau_z}{2}$ *is the nucleon charge.*

Example 2. *(DLVO Theory)*

The DLVO theory[3] *uses two different forces, electrostatic repulsion and van der Waals attraction to explain the colloidal stability [57]. For example, the DLVO potential between two spheres of radius* R *at a distance* D *away from each other is [10]:*

$$W(D) = W(D)_A + W(D)_R, \tag{2.9}$$

[3] *The DLVO theory is named after Derjaguin, Landau, Verwey, and Overbeek [23].*

where $W(D)_A$ and $W(D)_R$ are van der Waals attractive energy, and repulsive energy due to electrostatic forces. Here, $W(D)_A$ and $W(D)_R$ are defined below

$$W(D)_A = -\frac{\pi^2 C\rho^2}{6D}\frac{R}{2}, \tag{2.10}$$

where C is a constant for the interaction energy, and ρ is the number density of the sphere.

$$W(D)_R = \frac{64\pi k_b T R \rho_\infty \gamma^2}{\kappa^2} e^{-\kappa D}. \tag{2.11}$$

Here, γ is the reduced surface potential

$$\gamma = \tanh(\frac{ze\psi_0}{4kT}),$$

where ψ_0 is the potential on the surface, and T is the temperature.
In addition, κ^{-1} is the characteristic thickness of the double layer (Debye length)

$$\kappa = \sqrt{\sum_i \frac{\rho_{\infty i} e^2 z_i^2}{\epsilon_r \epsilon_0 k_b T}},$$

where

- *$\rho_{\infty i}$ is the number density of the ion i in the bulk solution,*
- *z is the valency of the ion,*
- *ϵ_r is the relative static permittivity,*
- *ϵ_0 is the vacuum permittivity, and*
- *k_b is the Boltzmann constant.*

CHAPTER 3

CONVEX RELAXATION OF THE HELMHOLTZ FUNCTIONAL

This section discusses finding global minimizers to the Helmholtz functional on a periodic domain when there is no external potential. The presented approach relies on a convex relaxation of the pairwise Helmholtz functional, and results in sufficient conditions for global minimizers. The sufficient conditions are computationally tractable as convex problems, and when satisfied, guarantee that a probability density is a global minimizer. The sufficient conditions take the form of a linear optimization problem for the auto-correlation of the probability density with non-negative Fourier modes.

3.1 Convex Relaxations to Computationally Tractable Problems

We are interested in energy functionals that model systems with a large number of particles, and take the form

$$\mathcal{E}(\rho) := \frac{1}{2}\int_\Omega \int_\Omega \rho(\mathbf{y}) w(\mathbf{x}-\mathbf{y})\rho(\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}, \qquad \int_\Omega \rho(\mathbf{x})\mathrm{d}\mathbf{x} = 1. \tag{3.1}$$

Here, we restrict our focus to a periodic domain $\Omega = [0,1]^d$ with dimension $1 \leq d \leq 3$, $w(x)$ is the interaction energy, and $\rho(\mathbf{x})$ is the density function used to represent the distribution of particles.

Remark 1. *(Inversion symmetry of the interaction potential) For the energy functional in (3.1), without loss of generality one can take the interaction potential, $w(\mathbf{x})$, to be symmetric under inversion, i.e., $w(-\mathbf{x}) = w(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$. Here, $w(-\mathbf{x}) := w(-x_1, -x_2, \ldots, -x_n)$. The symmetry assumption on $w(\mathbf{x})$ can be justified by writing $w(\mathbf{x})$ in terms of its even and odd components, $w(\mathbf{x}) = w_E(\mathbf{x}) + w_O(\mathbf{x})$,*

where

$$w_E(\mathbf{x}) := \frac{1}{2}\big(w(\mathbf{x}) + w(-\mathbf{x})\big), \qquad w_O(\mathbf{x}) := \frac{1}{2}\big(w(\mathbf{x}) - w(-\mathbf{x})\big).$$

Note that the integral in (3.1) is zero for the function $w_O(x)$:

$$\int_\Omega \int_\Omega \rho(\mathbf{y}) w_O(\mathbf{x}) \rho(\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} = \int_\Omega \int_\Omega \rho(\mathbf{y}) \big(w(\mathbf{x}-\mathbf{y}) - w(\mathbf{y}-\mathbf{x})\big) \rho(\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} = 0.$$

Hence, if $w(\mathbf{x})$ is not symmetric under inversion, one can simply replace $w(\mathbf{x})$ with $w_E(\mathbf{x})$ since the integral with $w_O(\mathbf{x})$ vanishes.

The problem to find global minimizers to the pairwise energy (3.1) is:

$$\begin{aligned} &\text{minimize} && \frac{1}{2} \int_\Omega \int_\Omega \rho(\mathbf{y}) w(\mathbf{x}-\mathbf{y}) \rho(\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}, \\ &\text{over probability measures} && \rho(\mathbf{x}) \in \mathcal{C}_1 \text{ with } \int_\Omega \rho(\mathbf{x}) \mathrm{d}\mathbf{x} = 1. \end{aligned} \tag{P}$$

Here $\mathcal{C}_1$ is the following convex cone[1]

$$\mathcal{C}_1 := \Big\{ f \in \mathcal{C}^0(\Omega)' : \int_\Omega f(\mathbf{x}) u(\mathbf{x}) \mathrm{d}\mathbf{x} \geq 0 \text{ for all } u \in \mathcal{C}^0(\Omega) \text{ with } u(\mathbf{x}) \geq 0 \Big\}.$$

Remark 2. *(Solution to the problem (P)) We denote the global minimum of the problem (P) as $\mathcal{E}_0 := \mathcal{E}(\rho_0)$, achieved by some probability measure $\rho_0(\mathbf{x})\mathrm{d}\mathbf{x}$. Under the following assumptions 1, the minimizer (P) exists, however when Ω is not bounded the minimizer might not exist [12, 14, 19, 53].*

Assumption 1. *(Assumptions on the interaction energy $w(\mathbf{x})$)*

A1. $w(\mathbf{x})$ is continuous on Ω.

A2. $w(\mathbf{x})$ is periodic with period 1.

[1] In the definition of $\mathcal{C}_1$, $\mathcal{C}^0(\Omega)$ is the space of periodic continuous functions on Ω.

Definition 3.1.1. *(Definition of convexity) We say that a function (or functional)* $\mathcal{E}$ *is convex if for any* ρ_1, ρ_2 *and* $0 \leq \alpha \leq 1$*, then*

$$\mathcal{E}\big(\alpha\rho_1 + (\alpha - 1)\rho_2\big) \leq \alpha\mathcal{E}(\rho_1) + (1 - \alpha)\mathcal{E}(\rho_2).$$

A set $\mathcal{S}$ *is convex if for any* $x, y \in \mathcal{S}$ *and* $0 \leq \alpha \leq 1$ *then*

$$\alpha x + (\alpha - 1)y \in \mathcal{S}.$$

For convex problems, sufficient conditions for global minimizers can be formulated using the *Karush-Kuhn-Tucker* (KKT) conditions. For problem (P), the KKT conditions for a density $\rho^*(\mathbf{x})$ take the form [6, 15]

$$\Lambda(\mathbf{x}) := \int_\Omega w(\mathbf{x} - \mathbf{y})\rho^*_{\mathbf{y}}\mathrm{d}\mathbf{y}, \tag{3.2}$$

and satisfies

$$\Lambda(\mathbf{x}) = 2\mu, \quad \text{for all } \mathbf{x} \in \mathcal{S}_* := \mathrm{supp}(\rho^*)^2. \tag{3.3}$$

$$\Lambda(\mathbf{x}) \geq 2\mu, \quad \text{for all } \mathbf{x} \in \Omega. \tag{3.4}$$

Here, $\mu \in \mathbb{R}$ is a *Lagrange multiplier* constant. For general $w(\mathbf{x})$, the energy $\mathcal{E}$ is not convex. One of the primary difficulties when solving the problem (P) is a lack of sufficient conditions. Note that if ρ^* solves the KKT equations for problem (P), then ρ^* does not necessarily solve (P). In other words, the KKT conditions are not sufficient to guarantee global minimizers.

We do not work directly with the KKT conditions, but rather formulate global conditions that provide sufficient conditions for global minimizers. The sufficient conditions are formulated using a convex relaxation to the problem (P) and also result in a lower bound on the energy $\mathcal{E}_0$.

[2] $\mathrm{supp}(f)$ is the support, i.e., the set where $f(\mathbf{x})$ does not vanish.

To obtain the convex relaxation to the problem (P) (see [3]), we use a change of variables $\mathbf{s} = \mathbf{x} - \mathbf{y}$ in the integral of the energy (3.1)

$$\mathcal{E}(\rho) = \frac{1}{2}\int_\Omega \int_\Omega \rho(\mathbf{x})\rho(\mathbf{x}+\mathbf{s})w(\mathbf{s})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{s} = \frac{1}{2}\int_\Omega F(\mathbf{s})w(\mathbf{s})\mathrm{d}\mathbf{s}, \tag{3.5}$$

$$\text{where} \qquad F(\mathbf{s}) := \int_\Omega \rho(\mathbf{x})\rho(\mathbf{x}+\mathbf{s})\mathrm{d}\mathbf{x}.$$

Definition 3.1.2. *(Auto-correlation of $\rho(\mathbf{x})$) $F(\mathbf{s})$ defined in (3.5) is the auto-correlation of $\rho(\mathbf{x})$. We denote it with $F = \rho \circ \rho$, i.e.,*

$$\rho \circ \rho := F(\mathbf{s}) := \int_\Omega \rho(\mathbf{x})\rho(\mathbf{x}+\mathbf{s})\mathrm{d}\mathbf{x}. \tag{3.6}$$

Defining the set $\mathcal{A}$ as

$$\mathcal{A} := \Big\{F : F(\mathbf{s}) = \int_\Omega \rho(\mathbf{x})\rho(\mathbf{x}+\mathbf{s})\mathrm{d}\mathbf{x}, \text{ such that } \rho \in \mathcal{C}_1, \int_\Omega \rho(\mathbf{x})\mathrm{d}\mathbf{x} = 1\Big\},$$

the problem (P) can be written as

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\int_\Omega F(\mathbf{x})w(\mathbf{x})\mathrm{d}\mathbf{x}, \\ &\text{subject to} \quad F \in \mathcal{A}. \end{aligned} \tag{P$'$}$$

Note that the Problem (P$'$) is not a convex optimization problem despite the linear functional $\langle F, w\rangle$. This is because $\mathcal{A}$ is not a convex space (see Remark 3).

Remark 3. *(The set $\mathcal{A}$ is not convex) To show that $\mathcal{A}$ is not convex take $f_1(x) = 1+\cos(2\pi x)$ and $f_2(x) = 1+\cos(2\pi n x)$ on $\Omega = [0,1]$, where $n \gg 1$ is a large integer. The convex combination of*

$$\begin{aligned} \lambda(f_1 \circ f_1) + (1-\lambda)(f_2 \circ f_2) &= \lambda\big(1+\frac{1}{2}\cos(2\pi x)\big) + (1-\lambda)\big(1+\frac{1}{2}\cos(2n\pi x)\big) \\ &= 1 + \frac{1}{4}\cos(2\pi x) + \frac{1}{4}\cos(2n\pi x), \end{aligned}$$

when $\lambda = \frac{1}{2}$, must come from an auto-correlation of a function taking the form (with arbitrary phases φ_1, φ_2)

$$f_3(x) = 1 + \frac{1}{\sqrt{2}}\cos(2\pi x - \varphi_1) + \frac{1}{\sqrt{2}}\cos(2n\pi x - \varphi_2).$$

Choosing n large enough, the minimum value of $f_3(x)$, regardless of the values φ_1, φ_2, can be made arbitrary close to $1 - \sqrt{2} < 0$. Hence, for sufficiently large n, there is no non-negative probability $f_3(x)$ with auto-correlation $(\lambda(f_1 \circ f_1) + (1-\lambda)(f_2 \circ f_2))$.

Proposition 3.1.3. *(Properties of $\mathcal{A}$) Given any $F(\mathbf{x}) \in \mathcal{A}$, the following properties hold:*

P1. $F(\mathbf{x})$ is a probability.

P2. The Fourier transform of $F(\mathbf{x})$ is real and non-negative.

Proof. The proof for proposition 3.1.3 has two parts:

1. To prove (P1), it is sufficient to show that $F(\mathbf{x})$ is non-negative, and integrates to one.

   i. For any continuous, non-negative function $u(\mathbf{x}) \geq 0$, the integral $\langle F, u\rangle$ can be written as

   $$\langle F, u\rangle = \int_\Omega \int_\Omega \rho(\mathbf{x})\rho(\mathbf{y})u(\mathbf{x}-\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} = \langle \rho, U\rangle,$$

   where

   $$U(\mathbf{x}) := \int_\Omega \rho(\mathbf{y})u(\mathbf{x}-\mathbf{y})\mathrm{d}\mathbf{y},$$

   and $\rho(\mathbf{x}) \in \mathcal{C}_1$. The function $U(\mathbf{x}) \geq 0$ is non-negative, and also continuous since it is a convolution of $\rho(\mathbf{x})$ with a continuous function U. Hence, integrating $U(\mathbf{x})$ against $\rho(\mathbf{x})$ is also non-negative, implying: $\langle F, u\rangle = \langle \rho, U\rangle \geq 0$.

ii. Taking $u(\mathbf{x}) = 1$ in the definition for $U(\mathbf{x})$ implies that $U(\mathbf{x}) = 1$. It then follows that $\langle F, 1\rangle = \langle \rho, 1\rangle = 1$.

2. To prove (P2), it is enough to show that sine modes of $F(\mathbf{x})$ are all zero and cosine modes of $F(\mathbf{x})$ are all non-negative.

   i. Integrating $F(\mathbf{x})$ against any sine mode, $\sin(2\pi\mathbf{k}.\mathbf{x})$, yields:

$$\langle F, \sin(2\pi\mathbf{k}.\mathbf{x})\rangle = \int_\Omega \int_\Omega \rho(\mathbf{x})\rho(\mathbf{y})\big(\sin(2\pi\mathbf{k}.\mathbf{x})\cos(2\pi\mathbf{k}.\mathbf{y}) - \sin(2\pi\mathbf{k}.\mathbf{y})\cos(2\pi\mathbf{k}.\mathbf{x})\big)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} = 0.$$

   ii. Integrating $F(\mathbf{x})$ against any sine mode, $\cos(2\pi\mathbf{k}.\mathbf{x})$, yields:

$$\begin{aligned}\langle F, \cos(2\pi\mathbf{k}.\mathbf{x})\rangle &= \int_\Omega \int_\Omega \rho(\mathbf{x})\rho(\mathbf{y})\cos\big(2\pi\mathbf{k}.(\mathbf{x}-\mathbf{y})\big)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\ &= |\langle \rho, \cos(2\pi\mathbf{k}.\mathbf{x})\rangle|^2 + |\langle \rho, \sin(2\pi\mathbf{k}.\mathbf{x})\rangle|^2 \geq 0.\end{aligned}$$

□

Proposition 3.1.3 (P2) implies that $F \in \mathcal{C}_2$, where $\mathcal{C}_2$ is the following convex set:

$$\begin{aligned}\mathcal{C}_2 := \Big\{ f \in \mathcal{C}^0(\Omega)' : \text{for all continuous } u(\mathbf{x}) \geq 0, \text{ and } \mathbf{k} \in \mathbb{Z}^d \setminus 0, \\ \int_\Omega f(\mathbf{x})\cos(2\pi.\mathbf{x})\mathrm{d}\mathbf{x} \geq 0, \int_\Omega f(\mathbf{x})u(\mathbf{x})\mathrm{d}\mathbf{x} \geq 0, \\ \int_\Omega f(\mathbf{x})\sin(2\pi.\mathbf{x})\mathrm{d}\mathbf{x} = 0, \int_\Omega f(\mathbf{x})\mathrm{d}\mathbf{x} = 1\Big\}.\end{aligned} \tag{3.7}$$

Furthermore, (P1) implies that F is a probability, which is also a convex set. Since the intersection of two convex sets is convex, the set $\mathcal{C}_1 \cap \mathcal{C}_2$ is a convex space. This observation motivates relaxing the optimization of (P′) over the set $\mathcal{A}$, with the

following relaxation over the intersection of $\mathcal{C}_1$ and $\mathcal{C}_2$

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\int_\Omega w(\mathbf{x})F(\mathbf{x})\mathrm{d}\mathbf{x} \\ \text{subject to} \quad & \int_\Omega F(\mathbf{x})\varphi(\mathbf{x})\mathrm{d}\mathbf{x} \geq 0 \text{ for all } \varphi(\mathbf{x}) \in \mathcal{C}^0(\Omega)^{3} \text{ with } \varphi(\mathbf{x}) \geq 0, \\ & \int_\Omega F(\mathbf{x})\mathrm{d}\mathbf{x} = 1, \\ & \int_\Omega F(\mathbf{x})e^{-i2\pi\mathbf{k}\mathbf{x}}\mathrm{d}\mathbf{x} \geq 0 \ (\mathbf{k} \in \mathbb{Z}^d \setminus \mathbf{0}). \end{aligned} \tag{R}$$

Here, ≥ 0 means the Fourier transform is both real and non-negative.

Remark 4. *(Solution to the relaxed problem (R)) We denote the solution to the relaxed problem (problem (R)) as $F_R(\mathbf{x})$, and the corresponding energy to be $\mathcal{E}_R = \frac{1}{2}\langle F_R(\mathbf{x}), w\rangle$. In general, we observe that solutions to the problem (R) may be either continuous functions, i.e., $F_R(\mathbf{x}) \in C^0$; or may be non-classical functions, such as a combination of a finite number of Dirac point masses, i.e.,*

$$F_R(\mathbf{x}) = \frac{1}{|\chi|}\sum_{\mathbf{s}\in\chi}\delta(\mathbf{x}-\mathbf{s}), \tag{3.8}$$

where, $\chi \subset \Omega$, and $|\chi|$ is the number of Dirac points.

Since the optimization in (R) is over a set that is strictly larger than $\mathcal{A}$, one immediately has the inequality $\mathcal{E}_R \leq \mathcal{E}_0$.

Problem (R) also has a dual, which will be useful in formulating sufficient conditions for minimizers. In the following, we assume the existence of Lagrange multipliers to the constraint that $F(\mathbf{x}) \in \mathcal{C}_1$, and also $F(\mathbf{x}) \in \mathcal{C}_2$. With this

[3] $\mathcal{C}^0(\Omega)$ is the space of periodic continuous functions on Ω.

assumption, the dual (D) to the problem (R) takes the form

$$\begin{aligned} \text{maximize} \quad & \mathcal{E}_D \\ \text{subject to} \quad & w(\mathbf{x}) - 2\mathcal{E}_D = w^+(\mathbf{x}) + K(\mathbf{x}), \\ & w^+(\mathbf{x}) \geq 0, \\ & \int_\Omega K(\mathbf{x}) \cos(2\pi \mathbf{k}.\mathbf{x}) \mathrm{d}\mathbf{x} \geq 0. \end{aligned} \tag{D}$$

In problem (D) the function $w^+(\mathbf{x})$ plays the role of a Lagrange multiplier to the constraint that $F(\mathbf{x}) \in \mathcal{C}_1$, while $K(\mathbf{x})$ is the Lagrange multiplier to the constraint that $F(\mathbf{x}) \in \mathcal{C}_2$. The variable $\mathcal{E}_D$ actually plays two roles. It is the Lagrange multiplier to the constraint $\int_\Omega F \mathrm{d}\mathbf{x} = 1$, and is also the variable to optimize. Since $\mathcal{C}_1$ and $\mathcal{C}_2$ are convex cones, the functions $w^+(\mathbf{x})$ and $K(\mathbf{x})$ also reside in corresponding (dual) convex cones. Namely,

- The function $w^+(\mathbf{x})$ is assumed to be continuous, mirror symmetric and non-negative, i.e. $w^+(\mathbf{x}) \geq 0$. Together, these imply that for any $F(\mathbf{x}) \in \mathcal{C}_1$, we have

$$\int_\Omega F(\mathbf{x}) w^+(\mathbf{x}) \mathrm{d}\mathbf{x} \geq 0. \tag{3.9}$$

- $K(\mathbf{x})$ is a continuous, mirror symmetric, mean-zero function with real non-negative cosine coefficients[4], i.e.,

$$\begin{aligned} \hat{K}(\mathbf{k}) &:= \int_\Omega K(\mathbf{x}) \cos(2\pi \mathbf{k}.\mathbf{x}) \mathrm{d}\mathbf{x} \geq 0, \text{ for all } \mathbf{k} \in \mathbb{Z}^d, \text{ and } \hat{K}(0) = 0, \\ K(\mathbf{x}) &= \sum_{\mathbf{k} \in \mathbb{Z}^d} \hat{K}(\mathbf{k}) \cos(2\pi \mathbf{k}.\mathbf{x}). \end{aligned}$$

  Since $\hat{K}(\mathbf{k}) \geq 0$, for any $F(\mathbf{x}) \in \mathcal{C}_2$, we have

$$\int_\Omega F(\mathbf{x}) K(\mathbf{x}) \mathrm{d}\mathbf{x} = \sum_{\mathbf{k} \in \mathbb{Z}^d} \hat{K}(\mathbf{k}) \hat{F}(\mathbf{k}) \geq 0. \tag{3.10}$$

[4]We assume that $\sum_{\mathbf{k} \in \mathbb{Z}^d} \hat{K}(\mathbf{k}) < \infty$ to guarantee that the cosine series for $K(\mathbf{x})$ converges uniformly.

- Note that $\mathcal{E}_D$ is a lower bound for problem (R), which can be shown using (3.9) and (3.10) as follows

$$\begin{aligned}\mathcal{E}_R &= \frac{1}{2}\int_\Omega w(\mathbf{x})F_R(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\int_\Omega w^+(\mathbf{x})F_R(\mathbf{x})\mathrm{d}\mathbf{x} + \frac{1}{2}\int_\Omega K(\mathbf{x})F_R(\mathbf{x})\mathrm{d}\mathbf{x} + \mathcal{E}_D \\ &\geq \mathcal{E}_D.\end{aligned} \tag{3.11}$$

Remark 5. *(Dual cones in (D)) In finite dimensions, the cone $K \subset \mathbb{R}^n$ has the associate dual cone defined as*

$$K^* = \Big\{y \in \mathbb{R}^n : x^T y \geq 0 \text{ for all } x \in K\Big\}.$$

From (3.9) and (3.10), $w^+(\mathbf{x})$ and $K(\mathbf{x})$ are in cones that are dual to $\mathcal{C}_1$ and $\mathcal{C}_2$.

The solution to the dual problem (D) then gives an optimal decomposition of $w(\mathbf{x})$ as

$$w(\mathbf{x}) = w_R^+(\mathbf{x}) + K_R(\mathbf{x}) + 2\mathcal{E}_R. \tag{3.12}$$

3.2 Sufficient Conditions for Optimality

In this section we introduce sufficient conditions for a global minimizer of the problem (P), using the relaxation (R).

The sufficient conditions for a global minimum of the problem (P) are as follows:

- Suppose that $\rho^*(\mathbf{x})$ is a probability distribution with auto-correlation $F_R(\mathbf{x})$, i.e., $F_R(\mathbf{x}) = \rho^* \circ \rho^*$, that solves (R). Then, since the energy, $\mathcal{E}$, given by any probability distribution is by definition larger than the minimizer, $\mathcal{E}_0$, and also, problem (R) is a lower bound for (P), we have

$$\mathcal{E}(\rho_0) = \mathcal{E}_R \geq \mathcal{E}_0 \geq \mathcal{E}_R,$$

and therefore, $\mathcal{E}_R = \mathcal{E}_0$, which implies that $\rho^* = \rho_0$ is a global minimum to (P).

- There exists $\rho^*(\mathbf{x})$ that solves (D), i.e., $\mathcal{E}(\rho^*) = \mathcal{E}_D$. Then from (3.11), and the fact that any probability distribution is larger than the minimizer $\mathcal{E}_D = \mathcal{E}_R \leq \mathcal{E}_0 \leq \mathcal{E}_R$, which shows the ρ^* is optimal.

3.2.1 Complementarity Conditions

In this section we look at the complimentarity conditions which describes the relations between a solution ρ_0, and the dual variables $w_R^+(\mathbf{x})$ and $K_R(\mathbf{x})$.

Let's consider that we have the dual solution (3.12), and substitute it into the objective function $\langle F(\mathbf{x}), w(\mathbf{x})\rangle$ to obtain:

$$\mathcal{E}_R = \frac{1}{2}\int_\Omega w_R^+(\mathbf{x})F_R(\mathbf{x})\mathrm{d}\mathbf{x} + \frac{1}{2}\int_\Omega K_R(\mathbf{x})F_R(\mathbf{x})\mathrm{d}\mathbf{x} + \mathcal{E}_R. \tag{3.13}$$

Equations (3.9) and (3.10) imply that both integrals are in (3.13) are non-negative. Hence, for (3.13) to hold, the integrals must vanish:

$$\begin{aligned}\int_\Omega w_R^+(\mathbf{x})F_R(\mathbf{x})\mathrm{d}\mathbf{x} = 0,\\ \int_\Omega K_R(\mathbf{x})F_R(\mathbf{x})\mathrm{d}\mathbf{x} = 0.\end{aligned} \tag{3.14}$$

From (3.14) we can write a relationship for the support of $F_R(\mathbf{x})$ to $w_R^+(\mathbf{x})$ and the Fourier modes of $F_R(\mathbf{x})$ to $K_R(\mathbf{x})$. Specifically, the constraints in (3.14) infer the following complementary supports:

$$\begin{aligned} w_R^+(\mathbf{x})F_R(\mathbf{x}) &= 0, \qquad \text{for all } \mathbf{x} \in \Omega,\\ \hat{K}_R(\mathbf{k})\hat{F}_R(\mathbf{k}) &= 0, \qquad \text{for all } \mathbf{k} \in \mathbb{Z}^d, \end{aligned} \tag{3.15}$$

where $\hat{K}_R(\mathbf{k})$ and $\hat{F}_R(\mathbf{k})$ are the cosine coefficient $K_R(\mathbf{x})$ and $F_R(\mathbf{x})$.

Equations (3.15) shows that when the sufficient conditions are satisfied by some ρ_0, the complimentarity conditions prove insight into which spatial and Fourier components of $w_R^+(\mathbf{x})$ and $K_R(\mathbf{x})$ influence $\rho_0(\mathbf{x})$.

3.3 Example Solutions to Convex Relaxation

This subsection presents a few example solutions to the problem (R) and (D). The section focuses on a toy potential that is inspired by taking a Morse-type potential on a periodic domain

$$w_{PM}(x) = -GLe^{-\frac{1}{L}\sin(\pi|x|)} + e^{-\sin(\pi|x|)}, \qquad G, L > 0. \tag{3.16}$$

In analogy with the parameters often used in the Morse potential, G is the characteristic velocity induced by attraction of particles, while L is a characteristic length scale [5, 6, 47]. Note that the interaction energy is symmetric, i.e., $w_{PM}(x) = w_{PM}(-x)$, and bounded at the origin, i.e., $w_{PM}(0) < \infty$.

We present two key examples that highlight the different behavior in the solution $F_R(x)$ — that is, $F_R(x)$ may be a continuous function, or $F_R(x)$ may contain non-classical Dirac masses.

In the first example where $F_R(x)$ is continuous, we fix the parameter values in (3.16) to be $(G, L) = (0.9, 1.5)$. Figure 3.1 provides an example of the solution to (D), while Figure 3.2 shows the solution $F_R(x)$, and the decomposition $w_R^+(x)$. As we can see from Figure 3.2, $F_R(x)$, and $w_R^+(x)$ do not overlap, which is consistent with the results in §3.2.1.

The second example is for a solution where $F_R(x)$ is a collection of Dirac masses, such as the expression in (3.8). Figure (3.3) shows the solution $F_R(x) = \frac{1}{2}\delta(x) + \frac{1}{2}\delta(x - \frac{1}{2})$ which consists of two Dirac masses. The figure also demonstrates the complimentary conditions (3.14) and (3.15) regarding the relation between the support of the minimizer to (R), $F_R(x)$, and $w_R^+(x)$ and $K_R(x)$. Note that Fourier modes of $F_R(x)$ that shown by the blue dots in Figure 3.1 are approximately $\mathcal{O}(10^{-2})$, which is not zero. This is an artifact of the solution being computed using the *interior-point* method, and the values of the Fourier modes can be made smaller by tightening the tolerances in the interior-point method.

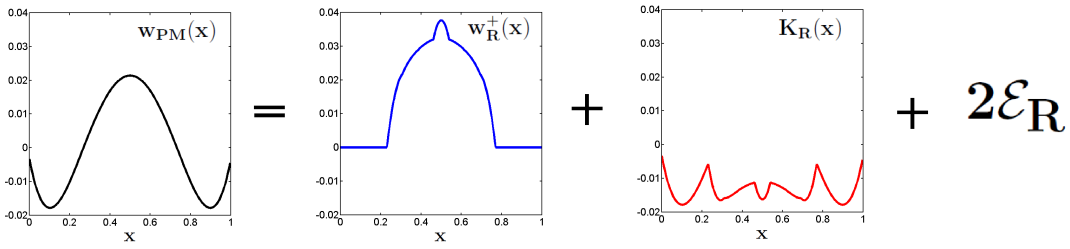

Figure 3.1 The figure shows the solution to (D) for the periodic potential $w_{PM}(x)$ defined in (3.16). The parameters are $(G, L) = (0.9, 1.5)$ resulting in a solution $F_R(x)$ (not shown) that is continuous. Computations are done with $n = 512$ grid points.

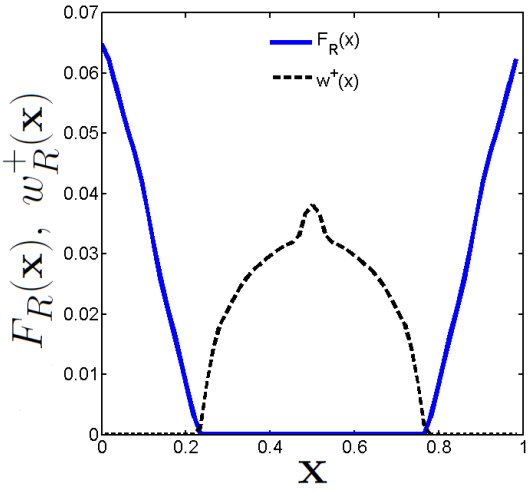

Figure 3.2 The figure is for parameters $(G, L) = (0.9, 1.5)$ which result in a non-classical solution $\hat{F}_R(x)$ (which is continuous). Computations were done with $n = 64$ grid points. The figure highlights the complementary support of $F_R(x)$ and $w_R^+(x)$ in real space.

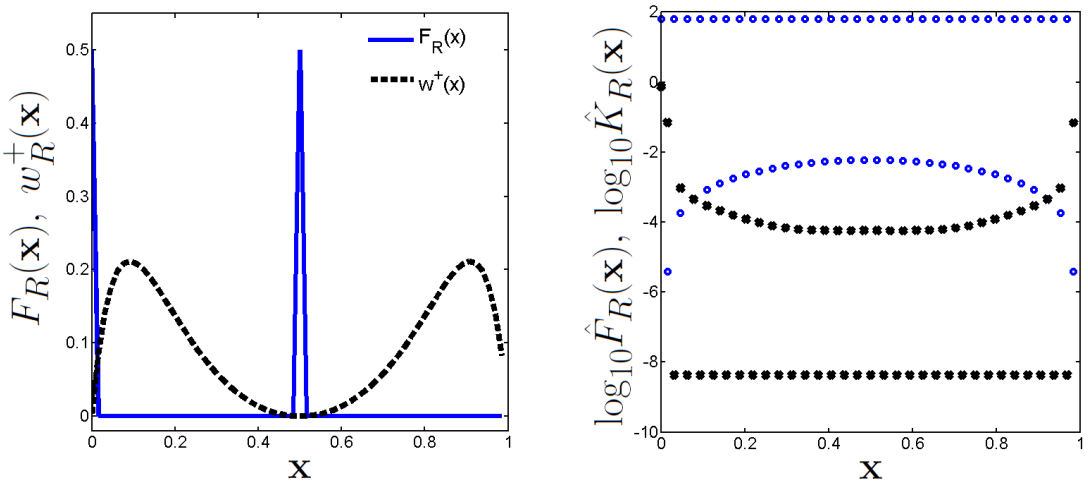

Figure 3.3 The figure is for parameters $(G, L) = (3, 0.2)$ which result in a non-classical solution F_R (which is two Dirac masses). Computations were done with $n = 64$ grid points. The figure highlights the complementary support of $\hat{F}_R(x)$ and $w_R^+(x)$ in real space (left), and complementary support of $\hat{F}_R(\mathbf{k})$ (blue circles) and $\hat{K}_R^+(\mathbf{k})$ (black crosses) in $\mathbf{k}$ space (right).

Remark 6. *Comparison of sufficient conditions and particle model) The gradient flow on equation (2.1) for the interaction potential (3.16) is defined by the ODE*

$$\dot{x}_j = -\nabla_{x_j} E_N, \quad 1 \le j \le N. \tag{3.17}$$

The long-time solution of the system (3.17) and the histogram of particle positions as $t \to \infty$ for parameters $(G, L) = (0.9, 1.5)$ for $N = 500$ particles with slightly perturbed uniform random initial data is shown in Figure 3.4. Comparison of the recovered approximate global minimizer, $\rho^*(x)$, that satisfies the sufficient conditions and the histogram from gradient flow shows that the support of the density is close to the width of the region which particles coalesce in particle model.

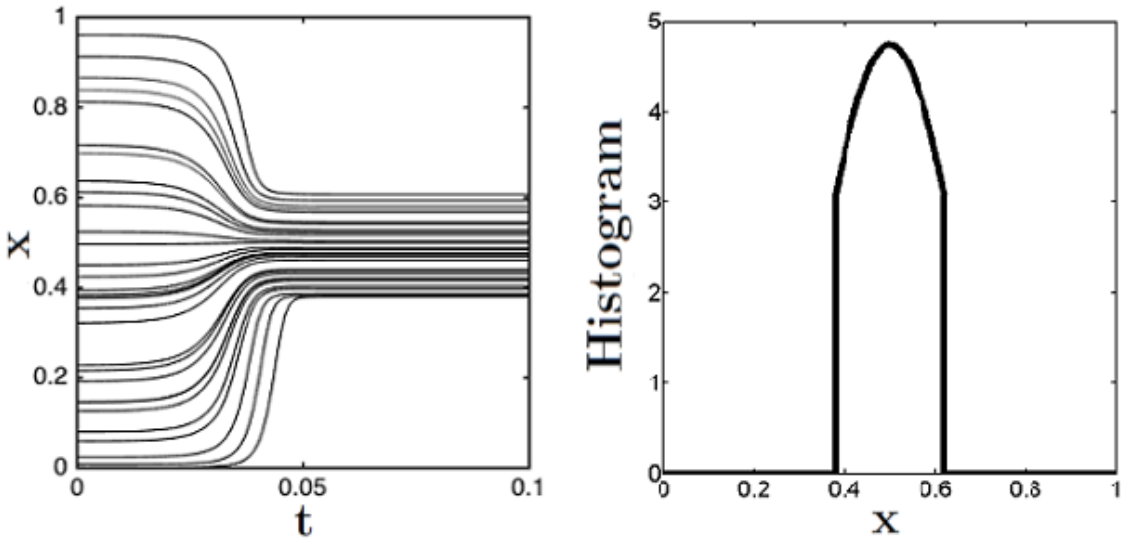

Figure 3.4 Gradient flow (i.e., time evolution of equation (3.17)) for $N = 500$ particles to a steady state (i.e., critical point) of (2.1) for a periodic Morse-type potential (3.16) (left), here only 30 particles shown; Particle density at the steady state obtained by differentiating the cumulative density function as was done in [6] (right). The parameters are $(G, L) = (0.9, 1.5)$.

CHAPTER 4

BASIC PROPERTIES OF LINEAR PROGRAMMING PROBLEMS

In this section we collect and review well-known properties of the solutions to linear programming (LP) problems in standard form. These properties will provide some insight into the nature of the numerical solutions to (R) and will be used in the subsequent chapters. The usual notation for a problem in standard form is:

$$\begin{aligned} \text{minimize} \quad & \mathbf{c}^{\mathrm{T}}\mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}, \\ & \mathbf{x} \geq \mathbf{0}, \end{aligned} \tag{4.1}$$

where $\mathbf{x}, \mathbf{c} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m, \mathbf{A} \in \mathbb{R}^{m \times n}$. Without loss of generality, the rows of $\mathbf{A}$ are assumed to be linearly independent so that they characterize independent equality constraints.

We denote the solution to the problem (4.1) as $\mathbf{x}^*$. The feasible set to the standard problem (4.1) is:

$$\mathcal{C} := \left\{ \mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \right\}. \tag{4.2}$$

The following subsections outline properties regarding solutions to LP problems of the form (4.1) (see, for instance, Chapter 3, [55]).

4.1 Constraint Qualifications

In this section we write the KKT conditions for (4.1). These conditions are important as they characterize solutions to (4.1).

Let the Lagrangian $\mathcal{L}$ associated with the standard problem (4.1) be:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) = \mathbf{c}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{A}\mathbf{x} - \mathbf{b}) - \mathbf{s}^{\mathrm{T}}\mathbf{x},$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\mathbf{s} \in \mathbb{R}^n$ are Lagrange multiplier vectors. A point $\mathbf{x}^*$ is a solution to (4.1) if and only if there exist points $(\mathbf{s}^*, \lambda^*)$ such that $(\mathbf{x}^*, \mathbf{s}^*, \lambda^*)$ satisfy the following KKT conditions [11].

$$\begin{aligned} \mathbf{c} + \mathbf{A}^T\boldsymbol{\lambda}^* - \mathbf{s}^* &= \mathbf{0}, \\ \mathbf{A}\mathbf{x}^* - \mathbf{b} &= \mathbf{0}, \\ \mathbf{s}^{*\mathrm{T}}\mathbf{x}^* &= 0, \\ \mathbf{x}^* &\geq \mathbf{0}, \\ \mathbf{s}^* &\geq \mathbf{0}. \end{aligned} \tag{4.3}$$

We use capital letters to denote the matrices of corresponding vectors, i.e., $\mathbf{S} = \mathrm{diag}(\mathbf{s})$. The following are conditions that guarantee a solution to (4.3): there is one point $\mathbf{s} > \mathbf{0}$ (strictly positive) and $\boldsymbol{\lambda}$ that satisfy the first equation in (4.3) [8]

$$\mathbf{A}^T\boldsymbol{\lambda} + \mathbf{c} = \mathbf{s}. \tag{4.4}$$

4.2 Solutions are Extreme Points

Definition 4.2.1. *(Extreme Points (Chapter 2, [7])) Let $\mathcal{S}$ be a non-empty convex set defined in $\mathbb{R}^n$. A vector $\mathbf{x} \in \mathcal{S}$ is called an extreme point if there are no two vectors $\mathbf{y}, \mathbf{z} \in \mathcal{S}$ and $\mathbf{y} \neq \mathbf{z}$ such that $\mathbf{x} = \alpha\mathbf{y} + (1-\alpha)\mathbf{z}$ for a scalar $\alpha \in (0, 1)$.*

For instance, when $\mathcal{S}$ is a polygon the extreme points are corners.

Theorem 4.2.2. *(LP Solutions are Extreme Points (Chapter 2, [7])) Assume that (4.1) has a solution. There is at least one extreme point of the set $\mathcal{C}$ that minimizes the linear objective function $\mathbf{c}^{\mathrm{T}}\mathbf{x}$.*

Proof. (a) Assume that the solution to the LP problem is unique. If the unique solution, $\mathbf{x}^*$, is not an extreme point then there are vectors $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ such that $\mathbf{x}^* = \alpha\mathbf{x} + (1-\alpha)\mathbf{y}$ for a scalar $\alpha \in (0, 1)$. The objective in (4.1) can be written as

$$\mathbf{c}^{\mathrm{T}}\mathbf{x}^* = \mathbf{c}^{\mathrm{T}}\alpha\mathbf{x} + (1-\alpha)\mathbf{c}^{\mathrm{T}}\mathbf{y} > \mathbf{c}^{\mathrm{T}}\alpha\mathbf{x}^* + (1-\alpha)\mathbf{c}^{\mathrm{T}}\mathbf{x}^*,$$

which implies that $\mathbf{c}^{\mathrm{T}}\mathbf{x}^* > \mathbf{c}^{\mathrm{T}}\mathbf{x}^*$, therefore, $\mathbf{x}^*$, is an extreme point by contradiction.

(b) If the solution to the LP problem is not unique, then the optimal set $\mathcal{C}_0 := \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{c}^{\mathrm{T}}\mathbf{x} \leq \mathbf{c}^{\mathrm{T}}\mathbf{y}, \text{ for all } \mathbf{x}, \mathbf{y} \in \mathcal{C}\right\}$ forms a closed convex set. Every closed convex set $\mathcal{C}_0$ has at least one extreme point (Chapter 2, [7]). □

Definition 4.2.3. *(Active and Inactive Sets) For any point $\mathbf{z} \in \mathcal{C}$, the active set is defined as $\mathcal{A} = \{j : z_j = 0\}$, the inactive set is $\mathcal{I} = \{j : z_j > 0\}$.*

Denote by $|\mathcal{A}|$ the number of elements in $\mathcal{A}$ (and the same for $|\mathcal{I}|$). Hence $|\mathcal{A}|$ denotes the number of zero entries of $\mathbf{z}$ and $|\mathcal{A}| + |\mathcal{I}| = n$. When $\mathbf{z}$ is in the interior of $\mathcal{C}$, $|\mathcal{A}| = 0$.

4.3 Upper Bounds on the Support of Solutions

Proposition 4.3.1. *(Upper bound on $|\mathcal{A}|$ (*Proposition 2.1.4 (b)*, [7])) A vector $\mathbf{v} \in \mathcal{C}$ defined in (4.2) is an extreme point of $\mathcal{C}$ if and only if the columns of $\mathbf{A}$ corresponding to the non-zero coordinates of $\mathbf{v}$ are linearly independent.*

Proof: ($\Longrightarrow$) Let $\mathbf{v} \in \mathcal{C}$ be a vector with k zero elements, (i.e., $|\mathcal{A}| = k$). If we write the constraint $\mathbf{A}\mathbf{v} = \mathbf{b}$ in block form with active and inactive sets as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{\mathcal{I}} & \mathbf{A}_{\mathcal{A}} \end{pmatrix}, \qquad \mathbf{v} = \begin{pmatrix} \mathbf{v}_{\mathcal{I}} \\ \mathbf{v}_{\mathcal{A}} \end{pmatrix} \quad \text{where } \mathbf{v}_{\mathcal{A}} = 0, \text{ and } \mathbf{v}_{\mathcal{I}} > 0,$$

then we obtain the (equivalent) equality constraint

$$\begin{pmatrix} \mathbf{A}_{\mathcal{I}} & \mathbf{A}_{\mathcal{A}} \end{pmatrix} \begin{pmatrix} \mathbf{v}_{\mathcal{I}} \\ \mathbf{0} \end{pmatrix} = \mathbf{b}.$$

If the columns of $\mathbf{A}_{\mathcal{I}}$ are linearly dependent, then $\mathbf{A}_{\mathcal{I}}\,\mathbf{w} = 0$, has a non-zero solution, $\mathbf{w} \neq 0$. For arbitrary $\beta \in \mathbb{R}$, we then have

$$\mathbf{A}_{\mathcal{I}}(\mathbf{v}_{\mathcal{I}} + \beta\mathbf{w}) = \mathbf{b}.$$

By taking $\beta > 0$ small enough, the vectors $\mathbf{v}_{\mathcal{I}} + \beta\mathbf{w} \geq 0$ and $\mathbf{v}_{\mathcal{I}} - \beta\mathbf{w} \geq 0$ are both non-negative. This implies that there exists a vector $\mathbf{w}' = \begin{pmatrix}\mathbf{w} & \mathbf{0}\end{pmatrix}^{\mathrm{T}}$ so that $\mathbf{v} + \beta\mathbf{w}' \in \mathcal{C}$ and $\mathbf{v} - \beta\mathbf{w}' \in \mathcal{C}$. We can write

$$\mathbf{v} = \frac{1}{2}\left(\mathbf{v} + \beta\mathbf{w}'\right) + \frac{1}{2}\left(\mathbf{v} - \beta\mathbf{w}'\right),$$

therefore, $\mathbf{v}$ is not an extreme point. Hence, the columns of $\mathbf{A}_{\mathcal{I}}$ must be linearly independent.

($\Longleftarrow$) Conversely, assume that $\mathbf{A}_{\mathcal{I}}$ has linearly independent columns. Suppose that $\boldsymbol{v}$ is not extreme. We can then write $\mathbf{v}$ as a convex combination of two other vectors $\mathbf{y}, \mathbf{z} \in \mathcal{C}$

$$\mathbf{v} = \alpha\mathbf{y} + (1-\alpha)\mathbf{z}, \quad \textit{for a scalar } \alpha \in (0,1).$$

Since $\mathbf{y} \geq 0$ and $\mathbf{z} \geq 0$, the only way for $v_j = 0$ is to also have $y_j = z_j = 0$. Since $\mathbf{v}$, $\mathbf{y}$, and $\mathbf{z}$ can be decomposed into blocks corresponding to the active and inactive sets of $\mathbf{v}$ (i.e., $\mathbf{v} = \begin{pmatrix}\mathbf{v}_{\mathcal{I}}^T & \mathbf{0}\end{pmatrix}^{\mathrm{T}}$), we have

$$\mathbf{A}\mathbf{v} = \mathbf{A}_{\mathcal{I}}\mathbf{v}_{\mathcal{I}} = \mathbf{b},$$

$$\mathbf{A}\mathbf{y} = \mathbf{A}_{\mathcal{I}}\mathbf{y}' = \mathbf{b},$$

$$\mathbf{A}\mathbf{z} = \mathbf{A}_{\mathcal{I}}\mathbf{z}' = \mathbf{b}.$$

It follows that $\mathbf{v}$, $\mathbf{y}$, and $\mathbf{z}$ are all solutions of a system with linearly independent columns, and by uniqueness, we have $\mathbf{v} = \mathbf{y} = \mathbf{z}$, implying that $\mathbf{v}$ is an extreme point of $\mathbf{C}$.

Remark 7. *(Unique solutions to (5.8) are determined by the set $\mathcal{A}$ or $\mathcal{I}$) Assume that $\mathbf{z}^*$ is the unique solution to (5.8) with active and inactive set $\mathcal{A}$ and $\mathcal{I}$. Then $\mathbf{z}^*$ is an extreme point. Without loss of generality, we can reorder the matrix $\mathbf{A}$ into blocks corresponding to the active and inactive set in $\mathbf{z} = \begin{pmatrix}\mathbf{0} & \mathbf{z}_{\mathcal{I}}\end{pmatrix}^{\mathrm{T}}$; $\mathbf{A} = \begin{pmatrix}\mathbf{A}_{\mathcal{A}} & \mathbf{A}_{\mathcal{I}}\end{pmatrix}$.*

Using the block structure, the equality constraint, i.e., $\mathbf{Az} = \mathbf{b}$*, can be written as*

$$\begin{pmatrix} \mathbf{A}_{\mathcal{A}} & \mathbf{A}_{\mathcal{I}} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{z}_{\mathcal{I}} \end{pmatrix} = \mathbf{A}_{\mathcal{I}}\mathbf{z}_{\mathcal{I}} = \mathbf{b}. \tag{4.5}$$

Since $\mathbf{z}^*$ *is an extreme point, by Proposition 4.3.1, the columns of* $\mathbf{A}$ *corresponding to non-zero coordinates of* $\mathbf{z}$ *are linearly independent, hence, (4.5) uniquely determines* $\mathbf{z}_{\mathcal{I}}$*.*

Remark 8. *(Geometry of the set* $\mathcal{C}$*, (4.2), and conditioning of the system (4.5)) Linear system (4.5) can be solved by knowing the active set corresponding to* $\mathbf{z}^*$*, i.e., solution to (5.8). The conditioning of the system (4.5) is determined by the matrix* $\mathbf{A}_{\mathcal{I}}$*, and is related to the geometry, i.e., the angles, at the vertex of the solution in the feasible set* $\mathcal{C}$*, defined at (4.2). Note that methods like the interior-point method hone in on the active set during the solution process. Therefore, the conditioning of* $\mathbf{A}_{\mathcal{I}}$ *has the potential to impact the use of iterative Krylov solvers in an interior-point method. This is because Krylov solvers may perform poorly by requiring many iterations on ill-conditioned problems.*

CHAPTER 5

CONIC OPTIMIZATION WITH FOURIER CONSTRAINTS

In this section we devise numerical methods for solving the sufficient conditions given by the relaxed problem (R). The specific form of the problem (R) we derived in the previous chapter involves minimizing a linear (variational) objective function, with a collection of linear constraints. Numerical discretizations of (R) will then take the form of a *linear programming* (LP) problem. This section will focus on devising numerical discretizations and corresponding *matrix-free* interior-point methods (IPMs) to solve the resulting LP problem for (R). Matrix-free methods that avoid having to build and directly solve large linear systems (such as those that occurring during the IPM solution process) provide an attractive avenue for improving the IPM computational solution time. For problems such as (R), reduction in computational time is crucial to increase the dimension (to $d = 2$ and 3) and resolution of the problems we may address.

5.1 Numerical Discretizations of the Convex Relaxation

In this section we present numerical details regarding discretization for the problem (R).

5.1.1 Periodic Domain: $\Omega = [0, 1]$

For the problem (R), we use an equispaced discretization using $n > 0$ grid points. In dimension $d = 1$, we have: $x_j = j/n, \quad \text{for } 0 \leq j \leq n-1$. Using the equispaced discretization, the functions $w(x)$ and $F(x)$ can be represented as vectors $\mathbf{w}, \mathbf{f} \in \mathbb{R}^n$ where

$$\mathbf{w}_j := w(x_j) \text{ and } \mathbf{f}_j \approx F(x_j). \tag{5.1}$$

Using the vectors (5.1) we discretize problem (R). Since we adopt an equispaced grid, the integrals in (R) can be discretized using simple quadrature.

- Mass constraint

$$\int_\Omega F(x)\mathrm{d}x = \langle F(x), 1\rangle \approx \frac{1}{n}\sum_{j=0}^{n-1} \mathbf{f}_j = 1. \tag{5.2}$$

- Non-negativity constraints, for all smooth $\phi(x) \geq 0$

$$\int F(x)\phi(x)\mathrm{d}x \geq 0 \qquad \Longrightarrow \qquad \mathrm{f}_j \geq 0. \tag{5.3}$$

- Fourier constraints

$$\int_\Omega F(x)e^{-i2\pi kx}\mathrm{d}x \geq 0, \text{ and real } (k \in \mathbb{Z}^d \setminus 0). \tag{5.4}$$

  Condition (5.4) is equivalent to the following cosine and sine expressions

$$\langle F(x), \cos(2\pi kx)\rangle \approx \sum_{j=0}^{n-1} \cos\left(2\pi kj\frac{1}{n}\right)\mathrm{f}_j \geq 0, \tag{5.5}$$

$$\langle F(x), \sin(2\pi kx)\rangle \approx \sum_{j=0}^{n-1} \sin\left(2\pi kj\frac{1}{n}\right)\mathrm{f}_j = 0. \tag{5.6}$$

The discrete (in)equalities (5.5), and (5.6) can be formulated using the *discrete Fourier transform*, which is defined as follows.

Definition 5.1.1. *(Discrete Fourier transform) The discrete Fourier transform (*DFT*) of the vector* $\mathbf{f}$*, with length* n*, and its inverse are defined as*

$$\hat{\mathrm{f}}_k = \sum_{j=0}^{n-1} \mathrm{f}_j\, \omega_n^{j\,k}, \qquad\qquad \mathrm{f}_j = \frac{1}{n}\sum_{k=0}^{n-1} \hat{\mathrm{f}}_k\, \omega_n^{-j\,k},$$

where $\omega_n = \mathrm{e}^{-\frac{2\pi i}{n}}$ *is the* n^{th} *root of unity.*

We use the notation $\hat{\mathbf{f}} = \text{fft}(\mathbf{f})$ and $\mathbf{f} = \text{ifft}(\hat{\mathbf{f}})$ to denote the fast Fourier transform algorithm, which computes the discrete Fourier transform using $\mathcal{O}(n \log n)$ flops. The DFT can be represented through matrix multiplication $\hat{\mathbf{f}} = \mathbf{F}\ \mathbf{f}$, where $\mathbf{F} \in \mathbb{C}^{n \times n}$ is given by ($\bar{\omega}$ is the complex conjugate of ω)

$$\mathbf{F}_{kj} = \omega_n^{jk}, \qquad \left(\mathbf{F}^{-1}\right)_{kj} = \frac{1}{n}\,\bar{\omega}_n^{jk},$$

or:

$$\mathbf{F} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega_n & \omega_n^2 & \dots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \dots & \omega_n^{2(n-1)} \\ \vdots & & \ddots & & \vdots \\ 1 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \dots & \omega_n^{(n-1)(n-1)} \end{pmatrix}.$$

Remark 9. *(Discrete Fourier Transform Norm) The matrices* $\mathbf{F}$ *and* $\mathbf{F}^{-1}$ *have orthogonal columns but are not unitary, and have the following matrix 2-norms*

$$\|\mathbf{F}\|_2 = \sqrt{n}, \qquad \|\mathbf{F}^{-1}\|_2 = \frac{1}{\sqrt{n}}.$$

Using the discretizations from (5.2–5.4), the discrete version of problem (R) is:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n}\mathbf{w}^{\mathrm{T}}\mathbf{f} \\ \text{subject to} \quad & \mathbf{f} \geq \mathbf{0}, \\ & \hat{\mathbf{f}} \geq \mathbf{0}, \\ & \mathbf{1}^{\mathrm{T}}\mathbf{f} = n, \\ & \frac{1}{\sqrt{n}}\,\mathbf{I}_-^{\mathrm{T}}\,\mathbf{F}\,\mathbf{f} - \hat{\mathbf{f}} = \mathbf{0}, \end{aligned} \tag{5.7}$$

where

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n, \ \mathbf{f} = \begin{pmatrix} \mathrm{f}_0 \\ \mathrm{f}_1 \\ \vdots \\ \mathrm{f}_{n-1} \end{pmatrix} \in \mathbb{R}^n, \ \hat{\mathbf{f}} = \begin{pmatrix} \hat{\mathrm{f}}_1 \\ \hat{\mathrm{f}}_2 \\ \vdots \\ \hat{\mathrm{f}}_{n-1} \end{pmatrix} \in \mathbb{C}^{n-1}, \ \mathbf{I}_- = \begin{pmatrix} \mathbf{0}^{\mathrm{T}} \\ \mathbf{I}_{n-1} \end{pmatrix} \in \mathbb{R}^{n\times(n-1)}.$$

Here, $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is the identity matrix, and $\mathbf{0}$ is a vector of zeros.

Note the slight abuse of notation that $\hat{\mathbf{f}}$ does not contain the zero mode $\hat{\mathrm{f}}_0$, and that $\mathbf{F}$ is also scaled by $\frac{1}{\sqrt{n}}$. This is to ensure that $\frac{1}{\sqrt{n}}\mathbf{F}$ has norm one, and also that in subsequent computations the application of $\mathbf{F}$ on a vector $\mathbf{v}$ is always just an FFT. We assume the problem (5.7) has a unique solution and denote it as $\mathbf{f}_n^*$ (i.e., $\mathbf{f}_n^* = \operatorname{argmin} \frac{1}{n}\mathbf{w}^{\mathrm{T}}\mathbf{f}$) and the optimal value $p_n^* = \frac{1}{n}\mathbf{w}^{\mathrm{T}}\mathbf{f}_n^*$.

The formulation of (5.7) warrants the following observations. First, we multiply the objective function in (5.7) by $\frac{1}{n}$, so that in the limit as $n \to \infty$ the optimal value $\frac{1}{n}\mathbf{w}^{\mathrm{T}}\mathbf{f}$ is a Riemann sum for the integral $\int w(x)F(x)\mathrm{d}x$. However, for any fixed value of n, replacing $\frac{1}{n}\mathbf{w}^{\mathrm{T}}\mathbf{f}$ with $\mathbf{w}^{\mathrm{T}}\mathbf{f}$ in (5.7) does not change the solution vector $\mathbf{f}_n^*$ (the minimum p_n^* would no longer converge as $n \to \infty$). Second, in (5.7) we have written $\hat{\mathbf{f}} \geq \mathbf{0}$ to mean that (i) $\hat{\mathbf{f}}$ is real and $\mathrm{Re}(\hat{\mathbf{f}}) \geq \mathbf{0}$. Note that $\mathrm{Im}(\hat{\mathbf{f}}) = 0$ is equivalent to $f_{n-j} = f_{j+1}$ for $j = 1, \cdots, \frac{n}{2} - 1$ (i.e., the vector $\mathbf{f}$ is symmetric).

The problem (5.7) involves constraints on $\mathbf{f}$ and its Fourier transform $\hat{\mathbf{f}}$. We now put the LP problem (5.7) into *standard form*[1]. This will allow us to write the IPM in a framework that can be generalized later to other discretizations and versions of our relaxed problem (R). It will also allow us to use standard LP theorems to provide some characterization on the support of the solutions to (5.7). To put the problem

[1]One standard form of an LP problem is: (i) the objective variables are non-negative; and (ii) the constraints are equalities.

into standard form, let:

$$\mathbf{u} = \begin{pmatrix} \mathbf{w} \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{z} = \begin{pmatrix} \mathbf{f} \\ \hat{\mathbf{f}} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \sqrt{n} \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} \frac{1}{\sqrt{n}}\mathbf{F} & -\mathbf{I}_{-} \end{pmatrix},$$

where $\mathbf{u} \in \mathbb{R}^{2n-1}$, $\mathbf{z} \in \mathbb{C}^{2n-1}$, $\mathbf{b} \in \mathbb{R}^{n}$, $\mathbf{A} \in \mathbb{C}^{n\times(2n-1)}$.

Problem (5.7) in standard form[2] is:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n}\mathbf{u}^{\mathrm{T}}\mathbf{z} \\ \text{subject to} \quad & \mathbf{A}\mathbf{z} = \mathbf{b}, \\ & \mathbf{z} \geq \mathbf{0}. \end{aligned} \tag{5.8}$$

Note that the constraint $\mathbf{z} \geq \mathbf{0}$ is subtle since it means that the values z_j are both real and $z_j \geq 0$ for all $0 \leq j \leq 2n-1$ (despite the fact that $\mathbf{A}$ is complex). In addition, the factor $\frac{1}{\sqrt{n}}$ in the definition of $\mathbf{A}$ implies that $\frac{1}{\sqrt{n}}\mathbf{F}$ has norm 1 (Remark 9), which ensures that the norm of $\mathbf{A}$ is bounded by 2 independent of n. Note that the rows of $\mathbf{A}$ are linearly independent since the invertible matrix $\mathbf{F}$ appears as a block in $\mathbf{A}$. Since $\mathbf{f}$ is symmetric, the values of $\hat{\mathbf{f}}$ are purely real. Therefore, the constraints on $\mathbf{f}$ ensure that $\mathbf{z} \in \mathbb{R}^{2n-1}$ is real. As we can see, the problem (5.8) has n equality constraints and $2n-1$ inequality constraints. We denote the solution to (5.8) as $\mathbf{z}^*$ which is unique due to the unique assumption on (5.7).

5.2 Interior-Point Methods

There are many methods for solving LP problems, with two of the most popular being the *simplex method* [20] and IPMs [11]. The difference between the simplex and IPM lies in how they traverse the feasible set to arrive at a solution. Specifically, for LP problems, the feasible set is a polyhedral set (i.e., consisting of linear inequalities and equalities) and the optimal solution (when unique) only occurs at the vertices, or

[2]With the exception that we include a $\frac{1}{n}$ prefactor in the objective function.

extreme points of the set (Theorem 4.2.2, [9]). In the worst case, the classical simplex method algorithm checks all vertices, leading to an exponential number of iterations (since polyhedral sets with m constraints can have 2^m vertices.) [40]. In practice, the simplex method performs much better than the worst-case scenario [58].

In 1979 L.G. Khachian proposed an ellipsoid algorithm with polynomial complexity, which in the worst case, has significantly fewer iterations than the simplex method [38]. Later, in 1984 N. Karmarkar discovered that the interior-point algorithm would outperform[3] the *ellipsoid method* [37]. See the review by M. Wright [58] for a detailed history of interior point methods. Unlike the simplex method, which improves by stepping around the vertices on the boundary of the feasible set, the interior-point method starts from a feasible interior point and improves along an interior path (known as the central path) toward the optimum point on the boundary. Generally, IPMs are a better approach to solve LP problems with many vertices, i.e., the feasible set is defined using a large number of bounding hyperplanes close to the optimum vertex [20].

A primary reason that we pursue IPMs over the simplex method is that IPMs generalize to *semidefinite programming* (SDP) problems, while the simplex method does not. Having methods that generalize to SDPs is an important future consideration: the sufficient conditions for the Helmholtz free energy involving multiple species of particles modifies the LPs in (R) to SDPs. Hence, advances in our understanding of IPMs for the LP problem (R) will extend to the more general SDP setting in the future.

Interior-point methods take (modified) Newton steps to optimize a penalized objective function and arrive at a solution. For large-scale problems (with a large number of variables and/or inequalities) the solution of a Newton step involves solving

[3]The complexity for Khachian's algorithm is $\mathcal{O}(n^6L^2)$, and the algorithm proposed by Karmarkar is $\mathcal{O}(n^{3.5}L^2)$, where n is the dimension of the problem and L is the problem data size (length of the input data stream) [21, 37].

a linear system with a Jacobian matrix and will often be the bottleneck for solving the optimization problem. If the Jacobian matrix is dense (as in our case), direct methods, such as Gaussian elimination, are very costly. For example, Gaussian elimination on Jacobian matrices that arise from numerical discretizations of (R) with n grid points may require $\mathcal{O}(n^3)$ floating point operations (flops) to solve. As a result, numerical discretizations of (R) using direct methods such as MATLAB's (version R2018a) built in optimization routines can handle problems with very limited resolution in two dimensions.

In our case, however, matrix-vector products involving the Jacobian can be computed quickly via the fast-Fourier transform due to the nature of the Fourier constraints. Hence, matrix-free Krylov subspace methods, that only require matrix vector products, offer an attractive route towards solving (R) in higher spatial dimensions. As with any Krylov method, a key challenge is to ensure that the linear system being solved is well-conditioned.

5.2.1 The Logarithmic Barrier Function

Interior-point methods (IPM) are a class of algorithms for solving optimization problems (with non-empty interiors or relative interiors) such as (4.1). Interior-point methods work by solving modified versions of the KKT conditions (4.3) using a sequence of Newton steps [11]. The purpose of adopting an interior-point method is to avoid imposing the inequality constraints in the KKT conditions, and instead to solve a system of equality constraints. This is because the solution to a system of equality constraints is often easier to implement numerically; for instance, using Newton or gradient descent methods.

One approach to solve constrained optimization problems with inequality constraints like the problem (5.8) is the *barrier method*, which is a particular interior-point method. The idea in the barrier method is to include the inequality

constraints directly into the objective function by adding a barrier function. The barrier function penalizes the objective function variables as they approach the boundary of the feasible set [11].

In order to solve (5.8) using the barrier method, the inequality constraints in (5.8) are penalized by a logarithmic barrier that is discussed in Definition 5.2.1.

Definition 5.2.1. *(Logarithmic barrier function) The logarithmic barrier function for the Problem (5.8) is:*

$$\phi(\mathbf{z}) := -\sum_{j=1}^{2n-1} \log(\mathrm{z}_j),$$

with $\mathbf{dom}\ \phi = \Big\{\ \mathbf{z} \in \mathbb{R}^n\ :\ \mathrm{z}_j > 0\ \ for\ \ j = 1, \cdots, 2n-1 \Big\}$.

We replace the inequality constraints in (5.8) with the logarithmic barrier to arrive at the following modified optimization problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n}\mathbf{u}^{\mathrm{T}}\mathbf{z} + t\phi(\mathbf{z}) \\ \text{subject to} \quad & \mathbf{A}\mathbf{z} = \mathbf{b}. \end{aligned} \tag{5.9}$$

The solution to problem (5.9) depends on the parameter $t > 0$ and is referred to as the *central path.*

Definition 5.2.2. *(Central Path) For each value of $t > 0$, denote the solution to (5.9) as $\mathbf{z}_t^*$, the locus of points $\mathbf{z}_t^*$ is referred to as the central path.*

The central path has many interesting properties [11]. If (5.8) has a solution, then the central path is well defined (since $\phi(\mathbf{z})$ is strictly convex on $\mathbf{z} > \mathbf{0}$, (5.9) has a unique minimizer).

The gradient of (5.9) is:

$$\nabla\Big(\frac{1}{n}\mathbf{u}^{\mathrm{T}}\mathbf{z} + t\phi(\mathbf{z})\Big) := \frac{1}{n}\mathbf{u} - t\mathbf{Z}^{-1}\mathbf{1}, \tag{5.10}$$

where $\mathbf{1} \in \mathbb{R}^{2n-1}$, and $\mathbf{Z}$ denotes diagonal matrix of the vector $\mathbf{z}$ (i.e., $\mathbf{Z} = \text{diag}(z_1, z_2, \ldots, z_{2n-1})$).

Remark 10. *(Properties of the barrier problem (5.9))*

- *Using the barrier method, one has the advantage of solving for an optimization problem without any inequality constraints with a desired method such as the Newton's method.*
- *The problem (5.9) is actually a family of nonlinear problems indexed by the parameter t.*
- *As the parameter t goes to zero, the solution to the barrier problem (5.9) becomes a better approximation to the solution of problem (5.8). Proposition 5.2.3 shows that as $t \to 0$, $\frac{1}{n}\mathbf{u}^{\mathrm{T}}\mathbf{z}_t^* \to \frac{1}{n}\mathbf{u}^{\mathrm{T}}\mathbf{z}^*$. Under a few additional assumptions on the problem (5.8), i.e., that $\mathbf{z}^*$ is unique, one can further show that $\mathbf{z}_t^* \to \mathbf{z}^*$ as $t \to 0$.*

Proposition 5.2.3. *(Convergence of the Barrier Problem (5.9), Chapter 11, [11]) In the solution to (5.9), $\mathbf{z}_t^*$ satisfies the following inequality*

$$\frac{1}{n}\mathbf{u}^{\mathrm{T}}(\mathbf{z}_t^* - \mathbf{z}^*) \leq t(2n-1). \tag{5.11}$$

Proof. The proof uses weak duality on the original problem (5.8). The Lagrangian corresponding to problem (5.8) has the form

$$L(\mathbf{z}, \mathbf{s}, \boldsymbol{\lambda}) = \frac{1}{n}\mathbf{u}^{\mathrm{T}}\,\mathbf{z} - \mathbf{s}^{\mathrm{T}}\,\mathbf{z} + \boldsymbol{\lambda}^{\mathrm{T}}\,(\mathbf{A}\mathbf{z} - \mathbf{b}). \tag{5.12}$$

To invoke weak duality and obtain an inequality, we introduce

$$\begin{aligned} g(\mathbf{s}, \boldsymbol{\lambda}) = \text{minimize} \quad & L(\mathbf{z}, \mathbf{s}, \boldsymbol{\lambda}) \\ \text{subject to} \quad & \mathbf{z} \geq \mathbf{0}. \end{aligned}$$

Weak duality says that for any $\mathbf{s} \geq \mathbf{0},\ \boldsymbol{\lambda} \in \mathbb{R}^n$ that is dual feasible, one has the lower bound on the objective function

$$g(\mathbf{s}, \boldsymbol{\lambda}) \leq \frac{1}{n}\mathbf{u}^{\mathrm{T}}\mathbf{z}^*. \tag{5.13}$$

Hence any choice of $\mathbf{s}, \boldsymbol{\lambda}$ that is dual feasible provides a lower bound. We now generate such an $\mathbf{s}$ and $\boldsymbol{\lambda}$ using the solution to the problem (5.9) to obtain the required inequality. Since $\mathbf{z}_t^*$ minimizes a convex objective function with only a linear constraint (no inequality constraint), the KKT equations for $\mathbf{z}_t^*$ involve only an equality constraint Lagrange multiplier which we call $\boldsymbol{\lambda}^*(t)$. Together $\mathbf{z}_t^*$ and $\boldsymbol{\lambda}^*(t)$ solve:

$$\mathbf{A}\mathbf{z}_t^* - \mathbf{b} = \mathbf{0}, \tag{5.14}$$

and

$$\frac{1}{n}\mathbf{u} - t\mathbf{Z}_t^{*^{-1}}\mathbf{1} + \mathbf{A}^{\dagger}\boldsymbol{\lambda}^*(t) = \mathbf{0}. \tag{5.15}$$

In equation (5.15) $\mathbf{Z}_t^* = \mathrm{diag}(\mathbf{z}_t^*)$, and $\mathbf{A}^{\dagger}$ is the conjugate transpose of a matrix $\mathbf{A}$. Now denote $\mathbf{s}^*(t) = t\mathbf{Z}_t^{*^{-1}}$. Here, $\mathbf{s}^*(t)$ and $\boldsymbol{\lambda}^*(t)$ are dual feasible for the original problem (5.8) because $\mathbf{s}^*(t) > 0$ (since $\mathbf{z}_t^* > 0$ is always in the interior), and $\boldsymbol{\lambda}^*(t) \in \mathbb{R}^n$ [11].

We now substitute the variables $\mathbf{s}^*(t)$ and $\boldsymbol{\lambda}^*(t)$ into $g(\mathbf{s}, \boldsymbol{\lambda})$ to obtain:

$$\begin{aligned} g(\mathbf{s}^*(t), \boldsymbol{\lambda}^*(t)) = \text{minimize} \quad & L(\mathbf{z}, \mathbf{s}^*(t), \boldsymbol{\lambda}^*(t)) \\ \text{subject to} \quad & \mathbf{z} \geq \mathbf{0}. \end{aligned}$$

Since $\mathbf{z}_t^*$ solves the equations (5.14–5.15), it is exactly the value needed to minimize the Lagrangian to obtain

$$\begin{aligned} g(\mathbf{s}^*(t), \boldsymbol{\lambda}^*(t)) &= L(\mathbf{z}_t^*, \mathbf{s}^*(t), \boldsymbol{\lambda}^*(t)) \\ &= \frac{1}{n}\mathbf{u}^{\mathrm{T}}\ \mathbf{z}_t^* - \mathbf{s}^*(t)\ \mathbf{z}_t^* + \boldsymbol{\lambda}^*(t)^{\mathrm{T}}\ (\mathbf{A}\mathbf{z}_t^* - \mathbf{b}) \\ &= \frac{1}{n}\mathbf{u}^{\mathrm{T}}\ \mathbf{z}_t^* - t(2n-1). \end{aligned} \tag{5.16}$$

Note that $2n-1$ is the dimension of $\mathbf{z}$ in (5.9). Finally, substituting (5.16) in (5.13) yields:

$$\begin{aligned} \frac{1}{n}\mathbf{u}^{\mathrm{T}}\ \mathbf{z}^* &\geq g(\mathbf{s}^*(t), \boldsymbol{\lambda}^*(t)) \\ &\geq \frac{1}{n}\mathbf{u}^{\mathrm{T}}\ \mathbf{z}_t^* - t(2n-1), \end{aligned}$$

which shows (5.11), and completes the proof. □

5.2.2 Primal-Dual Interior-Point Method

In this section we introduce the *primal-dual* interior-point method[4]. Since interior-point methods traverse the interior of the feasible set, we require that the interior of $\mathcal{C}$, i.e., $\mathrm{int}(\mathcal{C})$ (or more precisely, the relative interior of $\mathcal{C}$), to be non-empty.

The Lagrangian for the optimization problem (5.8) is

$$\mathcal{L} = \frac{1}{n}\mathbf{u}^{\mathrm{T}}\mathbf{z}\ + \boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{A}\mathbf{z} - \mathbf{b}) - \mathbf{s}^{\mathrm{T}}\mathbf{z},$$

where $\boldsymbol{\lambda} \in \mathbb{R}^n$ and $\mathbf{s} \in \mathbb{R}^{2n-1}$ are dual parameters.

In this primal-dual interior-point method $\mathbf{z}$, $\boldsymbol{\lambda}$, and $\mathbf{s}$ are strictly feasible in the sense that $\mathbf{z} > 0$, $\mathbf{s} > 0$, and $\boldsymbol{\lambda} \in \mathbb{R}^n$. The idea behind the primal-dual method is to

[4]The method is called primal-dual since both primal and dual variables are updated at each iteration [11]. One can also write primal-dual interior-point method using barrier parameter which is mentioned in Appendix B.

minimize (5.8) but instead relax the complementarity condition $\mathbf{z}^{\mathrm{T}}\mathbf{s} = \mathbf{0}$ and impose $\mathbf{ZS} = t\mathbf{I}$. With this modification, the KKT equations are:

$$\begin{aligned} \mathbf{g}_1 &:= \frac{1}{n}\mathbf{u} + \mathbf{A}^{\dagger}\boldsymbol{\lambda} - \mathbf{s}, \\ \mathbf{g}_2 &:= \mathbf{A}\mathbf{z} - \mathbf{b}, \\ \mathbf{g}_3 &:= \mathbf{ZS} - t\,\mathbf{I}. \end{aligned} \tag{5.17}$$

The equations that need to be solved are:

$$\mathbf{g}_1 = \mathbf{g}_2 = \mathbf{g}_3 = \mathbf{0}, \tag{5.18}$$

where we seek the solutions that satisfy $\mathbf{z} > \mathbf{0}$, and $\mathbf{s} > \mathbf{0}$. Note that the last equation in (5.17) is the relaxed form of the complimentary condition $\mathrm{z}_i\mathrm{s}_i = 0$ and allows for both $\mathrm{z}_i > 0$, and $\mathrm{s}_i > 0$ to be positive. As $t \to 0$, equation $\mathbf{g}_3$ converges to $\mathbf{g}_3 = \mathbf{0}$, so that the solution at $t = 0$ satisfies $\mathbf{ZS} = \mathbf{0}$.

Equations (5.17–5.18) are exactly the equations of the logarithmic barrier (5.14–5.15) obtained by introducing an extra variable $\mathbf{S} = t\mathbf{Z}^{-1}$. Although primal-dual methods and the logarithmic barrier method solve the same system of equations (i.e., the value of t has the same meaning in both systems), they differ slightly in the way that they generate sequences of values $\mathbf{z}, \boldsymbol{\lambda}, \mathbf{s}$ that solve (5.18). In the logarithmic barrier method, the values of $\mathbf{z}_t^*$ follow the central path and generate a corresponding set of $\mathbf{s}^*(t)$ that always exactly satisfies $\mathbf{g}_3 = 0$. Primal-dual methods allow for both $\mathbf{z}$ and $\mathbf{s}$ to vary independently so that when using an iterative method (i.e., Newton's method) to solve (5.17–5.18) the sequence of points $\mathbf{z}$ and $\mathbf{s}$ may not exactly satisfy $\mathbf{g}_3 = 0$ (and may not follow the central path). We continue to use the notation $\mathbf{z}_t$ to denote the sequence of points generated by the primal-dual method.

Primal-dual methods solve equations (5.18) simultaneously as $t \to 0$. The idea is to use a sequence of Newton steps to generate increments $\Delta\mathbf{z}$, $\Delta\boldsymbol{\lambda}$, and $\Delta\mathbf{s}$, and then to use these increments to move $\mathbf{z}$, $\mathbf{s}$, $\boldsymbol{\lambda}$ towards values that solve (5.18). A

typical Newton solver would fix a value of t and then solve (5.18) until the variables converge. However, it is more efficient to decrease the value of t in each Newton iteration. As a result, primal-dual IPM are not strictly Newton methods, since the equation $\mathbf{g}_3$ changes in each iteration of the Newton loop. The fact that the primal dual dynamics converge is an interesting problem in its own right.

One Newton step, starting at location $\mathbf{z}$, $\mathbf{s}$, $\boldsymbol{\lambda}$ applied to the equations (5.18) yields the following equation for the increment:

$$\begin{pmatrix} \mathbf{0} & \mathbf{A}^\dagger & -\mathbf{I} \\ \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{S} & \mathbf{0} & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Delta z} \\ \boldsymbol{\Delta \lambda} \\ \boldsymbol{\Delta s} \end{pmatrix} = \begin{pmatrix} -\mathbf{g}_1 \\ -\mathbf{g}_2 \\ -\mathbf{g}_3 \end{pmatrix}. \tag{5.19}$$

Note that the third equation in (5.19) yields:

$$\mathbf{S}\boldsymbol{\Delta z} + \mathbf{Z}\boldsymbol{\Delta s} = -\mathbf{g}_3. \tag{5.20}$$

Solving (5.20) for $\Delta\mathbf{s}$ in terms of $\Delta\mathbf{z}$, and substituting back into (5.19), yields the equivalent system

$$\overbrace{\begin{pmatrix} \mathbf{Z}^{-1}\mathbf{S} & \mathbf{A}^\dagger \\ \mathbf{A} & \mathbf{0} \end{pmatrix}}^{\mathbf{B}} \overbrace{\begin{pmatrix} \boldsymbol{\Delta z} \\ \boldsymbol{\Delta \lambda} \end{pmatrix}}^{\mathbf{d}} = \overbrace{\begin{pmatrix} -\mathbf{g}_1 - \mathbf{Z}^{-1}\mathbf{g}_3 \\ -\mathbf{g}_2 \end{pmatrix}}^{\mathbf{r}}. \tag{5.21}$$

For any vector $\tilde{\mathbf{d}}$ (that may or may not solve (5.21), we denote the residual as $\tilde{\mathbf{e}}$, i.e., $\tilde{\mathbf{e}} = \mathbf{B}\tilde{d} - \mathbf{r}$.

One potential issue with interior-point methods [30,41,58–60] is that the matrix $\mathbf{B}$ (usually) becomes ill-conditioned as the values of $\mathbf{z}$ and $\mathbf{s}$ approach their optimal points. As the values of $\mathbf{Z}$, $\mathbf{S}$ approach optimality ($t \to 0$), the product $z_j s_j \to 0$. As a result, the diagonal matrix $\boldsymbol{\Theta} := (\mathbf{Z})^{-1}\mathbf{S}$ (in the upper left block of $\mathbf{B}$) has values

that approach either $\Theta_{jj} \to 0$, if $s_j \to 0$, or $\Theta_{jj} \to \infty$, if $z_j \to 0$.

$$\lim_{t \to 0} \Theta_{jj} = z_j^{-1} s_j = \begin{cases} 0 & j \in \mathcal{I} \\ \\ \infty & j \in \mathcal{A} \end{cases}. \tag{5.22}$$

Hence, the norm $\|\boldsymbol{\Theta}\|_2 \to \infty$, which also causes the norm $\|\mathbf{B}\|_2 \to \infty$. Generally speaking, this will cause the condition (5.21) $\kappa(\mathbf{B}) \to \infty$ as $(\mathbf{z}, \boldsymbol{\lambda}, \mathbf{s})$ approach their optimal values. The ill-conditioning in $\mathbf{B}$, due to $\boldsymbol{\Theta}$, is generic and occurs in any IPM that has inequality constraints. Although IPM have an inherent ill-conditiong, the poor conditioning is not detrimental when direct linear solvers (such as Guassian elimination) are used on (5.21) [30, 41, 58–60]. In other words, Guassian elimination can be used to solve (5.21). If solving (5.21) is too computationally complex for Guassian elimination, and iterative Krylov solvers are used on (5.21) then the ill-conditioning due to the IPM does present a serious problem, and must be addressed.

5.3 Matrix-Free Methods for the Primal-Dual Algorithm

There has been a lot of interest in using matrix-free methods [4, 24–26, 32, 42]. For example, a fast, matrix-free implicit method has been developed to solve unsteady flow problems involving moving boundaries [46]. Matrix-free optimization methods have also been recently developed for compressed sensing problems [22, 31], as well for applications in *deep-learning* [61]. The inherent structure of the problem makes it necessary to investigate and propose an appropriate preconditioner in order to solve the Newton's step in the interior-point algorithm.

One way to remove the ill-condition in $\mathbf{B}$ due to the values of $\boldsymbol{\Theta}$ is through using a preconditioner (see §5.3.1, and §5.3.2). Using a matrix $\mathbf{P}$, in a symmetric fashion, we obtain the preconditioned system:

$$\begin{cases} \mathbf{P}^{-\frac{1}{2}} \mathbf{B} \mathbf{P}^{-\frac{1}{2}} \mathbf{y} = \mathbf{P}^{-\frac{1}{2}} \mathbf{r} \\ \\ \mathbf{P}^{-\frac{1}{2}} \mathbf{y} = \mathbf{d} \end{cases}. \tag{5.23}$$

When $\mathbf{P}$ is diagonal, with non-negative entries along the diagonal, then both $\mathbf{P}^{\frac{1}{2}}$ and $\mathbf{P}^{-\frac{1}{2}}$ are easily evaluated and well-defined.

Remark 11. *(Invertibility of the matrix* $\mathbf{B}$*) Matrix* $\mathbf{B}$ *defined in* (5.21) *is always invertible. This can be shown by finding* $\boldsymbol{\Delta z}$*, and* $\boldsymbol{\Delta\lambda}$ *in* (5.21) *directly. Note that by formulation, rows of the matrix* $\mathbf{A}$ *are linearly independent. In components the two equations in the linear system are:*

$$\boldsymbol{\Theta\Delta z} + \mathbf{A}^{\dagger}\boldsymbol{\Delta\lambda} = -\mathbf{g}_1 - \mathbf{Z}^{-1}\mathbf{g}_3, \tag{5.24}$$

and

$$\mathbf{A\Delta z} = -\mathbf{g}_2. \tag{5.25}$$

Note that $\boldsymbol{\Theta}$ *is diagonal and all entries are* > 0*, therefore, both sides of the equation* (5.24) *can be divided by* $\boldsymbol{\Theta}$*. Then multiply this equation through by* $\mathbf{A}$ *and using equation* (5.25) *gives*

$$\mathbf{A\Theta}^{-1}\mathbf{A}^{\dagger}\boldsymbol{\Delta\lambda} = -\mathbf{A\Theta}^{-1}(\mathbf{g}_1 + \mathbf{Z}^{-1}\mathbf{g}_3) + \mathbf{g}_2. \tag{5.26}$$

Let's denote $\mathbf{W} := \mathbf{A\Theta}^{-1}\mathbf{A}^{\dagger}$*. Provided that the matrix* $\mathbf{W}$ *is invertible (see Proposition 5.3.1), implies that equation* (5.26) *can be solved for* $\boldsymbol{\Delta\lambda}$*. This shows that the matrix* $\mathbf{B}$ *is invertible.*

Proposition 5.3.1. *(Invertibility of the matrix* $\mathbf{W}$*) The matrix* $\mathbf{W} := \mathbf{A\Theta}^{-1}\mathbf{A}^{\dagger}$ *is invertible.*

Proof. From definition the matrix $\mathbf{W}$ is symmetric. For $\mathbf{y} \in \mathbb{C}^n$, and $\mathbf{q} = \mathbf{A}^{\dagger}\mathbf{y}$, then we can write

$$\begin{aligned}\mathbf{y}^{\dagger}\mathbf{A}\mathbf{y} &= \mathbf{q}^{\dagger}\boldsymbol{\Theta}^{-1}\mathbf{q} \\ &= \sum_{j=1}^{n} |q_j|^2\theta_{jj}^{-1} \geq 0,\end{aligned}$$

since each term in the sum is positive. Hence, $\mathbf{W}$ is positive semi-definite. On the other hand $\mathbf{W}$ cannot have a zero eigenvalue, since $\mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{y} = \mathbf{0}$ implies that $\mathbf{A}^{\dagger}\mathbf{y} = \mathbf{0}$. But this cannot be true as rows of $\mathbf{A}$, i.e., columns of $\mathbf{A}^{\dagger}$, are linearly independent. Therefore, $\mathbf{W}$ is positive definite and so invertible. □

Remark 12. *(Solving the equation (5.21) using* MINRES*)*[5] *The left-hand side matrix in the equation (5.21) (i.e., $\mathbf{B}$) is symmetric, but not necessarily positive definite. Of the possible matrix-free standard Krylov subspace methods (*CG[6]*,* MINRES[7]*,* GMRES*), we focus on using* MINRES *since it is well-suited for symmetric matrices $\mathbf{B}$ that are not positive definite [48].*

Remark 13. *(Matrix-free approach) Matrix vector products involving $\mathbf{B}$ can be done using $\mathcal{O}(n\ \log n)$ flops. This is because $\mathbf{B}$ consists of: $\mathbf{\Theta}$, which is diagonal, and $\mathbf{A}$, which has a diagonal block and a* DFT *block. The* DFT *block in $\mathbf{A}$ can be computed using the* FFT *in $\mathcal{O}(n\ \log n)$.*

Note that in the primal-dual IPM algorithm explained in Algorithm 1, μ is the *centering parameter* to control the decrease of t. Also, $\epsilon_{\hat{\eta}}$ is the tolerance to restrict the surrogate duality gap, and ϵ_{NW} is the tolerance for the Newton method. At each Newton iteration we solve (5.21) using MINRES, where the stopping criteria is taken as $\|\tilde{\mathbf{e}}\|_{\infty} < \epsilon_{\mathrm{MR}}$.

Algorithm 1. *(Primal-dual interior-point method [11])*

given $\mathbf{z} > 0,\ t > 0,\ \mu > 1,\ \mathbf{s} > 0,\ \epsilon_{\hat{\eta}} > 0,\ \epsilon_{\mathrm{NW}} > 0,\ \epsilon_{\mathrm{MR}} > 0.$

repeat

1. *Determine t. Set $t := \frac{\hat{\eta}}{\mu}$, where $\hat{\eta}$ is the surrogate duality gap and computed as $\hat{\eta} = \frac{1}{2n-1}\ \sum_{j=1}^{2n-1} z_j\ s_j$.*

[5] *Minimal Residual Method, see the Appendix A for more information about Krylov subspace methods such as* MINRES.
[6] *Conjugate gradient*
[7] *Generalized minimal residual*

2. *Use preconditioned* MINRES *to solve* (5.21) *and obtain the Newton search direction* $\mathbf{d}$.

3. *Perform line search in direction* $\mathbf{d}$[8]*, to obtain the Newton increment and update* ($\mathbf{z}$, $\mathbf{s}$, $\boldsymbol{\lambda}$).

until $\|\mathbf{z}_i - \mathbf{z}_{i-1}\|_\infty < \epsilon_{\text{NW}}$ *and* $\hat{\eta} < \epsilon_{\hat{\eta}}$.

The primal-dual Algorithm 1 is a standard method [11]. However, MINRES is used in step 2 to solve Equation (5.21). The value $\hat{\eta}$ is the surrogate duality gap, which provides an estimate on how close the objective value at $\mathbf{z}_t$ is to the true objective function. The value of μ tries to iteratively drive the value of $\hat{\eta}$ to zero.

5.3.1 A Common Preconditioner

One common preconditioner that can be used to solve (5.21) is the diagonal (with positive entries) [49]

$$\mathbf{P_2} = \begin{pmatrix} \mathbf{Z}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{5.27}$$

An approach equivalent to using the preconditioner (5.27) is implemented in the software [50]. Specifically, this preconditioner can be implemented by substituting $\Delta\mathbf{z} = \mathbf{Z}^{\frac{1}{2}}\ \Delta\bar{\mathbf{z}}$ into the system (5.21), and then multiplying the first row by $\mathbf{Z}^{\frac{1}{2}}$. Therefore, in the variables $\Delta\bar{\mathbf{z}}, \boldsymbol{\Delta\lambda}$ the system is:

$$\begin{pmatrix} \mathbf{S} & \mathbf{Z}^{\frac{1}{2}}\mathbf{A}^\dagger \\ \mathbf{A}\mathbf{Z}^{\frac{1}{2}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Delta}\bar{\mathbf{z}} \\ \boldsymbol{\Delta\lambda} \end{pmatrix} = \begin{pmatrix} -\mathbf{Z}^{\frac{1}{2}}\mathbf{g}_1 - \mathbf{Z}^{-\frac{1}{2}}\mathbf{g}_3 \\ -\mathbf{g}_2 \end{pmatrix}. \tag{5.28}$$

Remark 14. *(Symmetric properties) The vector* $\mathbf{f}$ *is symmetric (and* $\hat{\mathbf{f}}$ *is real). Hence, the vector* $\mathbf{z}$ *inherits the same symmetries as well. It is straightforward to show*

[8] *The step length* α *along the search direction* $\mathbf{d}$ *is the maximum* α *so that* $\mathbf{z}, \mathbf{s}, \boldsymbol{\lambda}$ *stay strictly positive:* $(\mathbf{z}, \mathbf{s}, \boldsymbol{\lambda}) + \alpha(\Delta\mathbf{z}, \Delta\mathbf{s}, \Delta\boldsymbol{\lambda}) > 0$.

that the Krylov subspace $\mathbf{d}$, $\mathbf{Bd}$, $\mathbf{B}^2\mathbf{d}$ *preserves the symmetries in* $\mathbf{\Delta z}$. *Any choice of preconditioner should ensure that the (preconditioned) resulting Krylov subspace also preserves the symmetries in* $\mathbf{z}$ *and* $\boldsymbol{\lambda}$.

5.3.2 A New Preconditioner

In this section we introduce a new preconditioner for the system (5.23). We implement the preconditioner and solve the relaxed problem (R). We then study and compare the numerical performance and conditioning of the new preconditioner against other standard preconditioning approaches for several instances of problem (R).

We are particularly interested in understanding the performance scaling as the mesh n (problem size) grows, while ensuring convergence of the underlying solution $\mathbf{f}_n^* \to F_R(\mathbf{x})$.

We introduce and examine the following diagonal (with positive entries) preconditioner[9]

$$\mathbf{P_1} = \begin{pmatrix} \mathbf{I} + \mathbf{\Theta} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{5.29}$$

Applying (5.29) to (5.23), we have

$$\mathbf{P_1}^{-\frac{1}{2}}\mathbf{B}\mathbf{P_1}^{-\frac{1}{2}} = \begin{pmatrix} (\mathbf{I} + \mathbf{\Theta})^{-1}\mathbf{\Theta} & (\mathbf{I} + \mathbf{\Theta})^{-\frac{1}{2}}\mathbf{A}^\dagger \\ \mathbf{A}(\mathbf{I} + \mathbf{\Theta})^{-\frac{1}{2}} & \mathbf{0} \end{pmatrix}. \tag{5.30}$$

5.4 Performance of the Preconditioners: Asymptotic Study

In this subsection, we use asymptotics to understand how the different preconditioners $(\mathbf{P_1}, \mathbf{P_2})$ modify the equation (5.21) in the vicinity of an optimal solution (i.e., the extreme points of $\mathcal{C}$). In the vicinity of extreme points, the matrix $\mathbf{B}$ becomes poorly ill-conditioned, which is alleviated by the preconditioners. Krylov solvers,

[9] A non-diagonal preconditioner, $\mathbf{P_3}$, is also tested, but the results imply preference of using preconditioner $\mathbf{P_1}$ (see C for more details about implementation and comparison of preconditioner $\mathbf{P_3}$ to other preconditioners).

such as MINRES perform better on well-conditioned problems (see Appendix A), and motivates the current study to understand how the preconditioners improve conditioning.

In the limit as $(\mathbf{z}_t, \mathbf{s}_t)$ approach a solution $(\mathbf{z}, \mathbf{s})$, we may extract out different block terms in the matrix (5.30) using property (5.22)

$$\lim_{t\to 0} \frac{\Theta_{jj}}{1+\Theta_{jj}} = \frac{\mathrm{z}_j^{-1}\,\mathrm{s}_j}{1+\mathrm{z}_j^{-1}\,\mathrm{s}_j} = \frac{\mathrm{s}_j}{\mathrm{z}_j+\mathrm{s}_j} = \begin{cases} 0 & j\in\mathcal{I} \\ 1 & j\in\mathcal{A} \end{cases}. \tag{5.31}$$

$$\lim_{t\to 0} \frac{1}{\sqrt{1+\Theta_{jj}}} = \begin{cases} 1 & j\in\mathcal{I} \\ 0 & j\in\mathcal{A} \end{cases}. \tag{5.32}$$

In the MINRES algorithm, the matrix (5.30) will be applied as it appears. For the purpose of the asymptotic study, without loss of generality, we permute the rows and columns of $\mathbf{P_1}^{-\frac{1}{2}}(\mathbf{B})\mathbf{P_1}^{-\frac{1}{2}}$ so that they are in matrix blocks corresponding to the active and inactive sets. Specifically, $\mathbf{\Theta} = \mathrm{diag}(\mathbf{\Theta}_\mathcal{A}, \mathbf{\Theta}_\mathcal{I})$, $\mathbf{A} = (\mathbf{A}_\mathcal{A}, \mathbf{A}_\mathcal{I})$. Using (5.31), and (5.32), the matrix blocks in (5.30) simplify to

$$\mathbf{P_1}^{-\frac{1}{2}}(\mathbf{B})\mathbf{P_1}^{-\frac{1}{2}} = \begin{pmatrix} (1+\mathbf{\Theta})^{-1}\mathbf{\Theta} & (1+\mathbf{\Theta})^{\frac{-1}{2}}\mathbf{A}^\dagger \\ \mathbf{A}(1+\mathbf{\Theta})^{\frac{-1}{2}} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_\mathcal{I}^\dagger \\ \mathbf{0} & \mathbf{A}_\mathcal{I} & \mathbf{0} \end{pmatrix}, \tag{5.33}$$

we denote the submatrix $\mathbf{M}$ as

$$\mathbf{M} = \begin{pmatrix} \mathbf{0} & \mathbf{A}_\mathcal{I}^\dagger \\ \mathbf{A}_\mathcal{I} & \mathbf{0} \end{pmatrix}.$$

For every finite value of $t > 0$, the matrix $\mathbf{P_1}^{-\frac{1}{2}}\mathbf{B}\mathbf{P_1}^{-\frac{1}{2}}$ is invertible. However, in the limit $t \to 0$, the matrix $\mathbf{P_1}^{-\frac{1}{2}}$ becomes a projection, and the matrix $\mathbf{M}$ may not be invertible. The fact that $\mathbf{M}$ is not invertible is not a fundamental problem

because the preconditioned linear system is still solvable (since $\mathbf{P_1}^{-\frac{1}{2}}$ also multiplies the right-hand side vector as well).

To gain some insight into performance of the preconditioned Krylov methods, we examine the (effective) conditioning of the matrix $\mathbf{M}$. If the matrix $\mathbf{A}_{\mathcal{I}}$ is invertible, then $\mathbf{M}$ is also invertible and $\kappa(\mathbf{M})$ is well-defined. If $\mathbf{A}_{\mathcal{I}}$ is not invertible (for instance if $\mathbf{A}_{\mathcal{I}}$ is not square then $\mathbf{M}$ is not invertible), then we will examine the effective conditioning number:

$$\text{(Effective conditioning)} \qquad \kappa(\mathbf{M}) := \frac{\sigma_{max}(\mathbf{M})}{\sigma_{min}(\mathbf{M})}$$

where $\sigma_{min}(\mathbf{M})$ is the smallest non-zero singular value of $\mathbf{M}$.

Although $\mathbf{M}$ may not be invertible, we expect that the linear system $\mathbf{P_1}^{-\frac{1}{2}}(\mathbf{B})\mathbf{P_1}^{-\frac{1}{2}}\mathbf{d} = \mathbf{P_1}^{-\frac{1}{2}}\mathbf{b}$ will always be solvable which motivates the study of the effective conditioning.

Remark 15. *(Some properties of the matrix $\mathbf{M}$)*

- *The matrix $\mathbf{M}$ is symmetric but not positive definite. The (effective) condition number of $\mathbf{M}$ is related to $\mathbf{A}_{\mathcal{I}}$ as:*

$$\kappa(\mathbf{M}) = \frac{\sigma_{max}(\mathbf{M})}{\sigma_{min}(\mathbf{M})} = \frac{\sigma_{max}(\mathbf{A}_{\mathcal{I}})}{\sigma_{min}(\mathbf{A}_{\mathcal{I}})},$$

  where σ's are singular values and $\sigma_{min}(\mathbf{A}_{\mathcal{I}})$ is the smallest non-zero singular value of $\mathbf{A}_{\mathcal{I}}$.

- *One can derive a symmetric matrix which is positive semi-definite*[10]

$$\mathbf{M}^2 = \mathbf{M}^\dagger\mathbf{M} = \begin{pmatrix} \mathbf{A}_{\mathcal{I}}^\dagger\mathbf{A}_{\mathcal{I}} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{\mathcal{I}}\mathbf{A}_{\mathcal{I}}^\dagger \end{pmatrix},$$

  with the condition number

$$\kappa(\mathbf{M}^2) = \frac{\sigma_{max}(\mathbf{M}^2)}{\sigma_{min}(\mathbf{M}^2)} = \frac{|\lambda_{max}(\mathbf{M}^2)|}{|\lambda_{min}(\mathbf{M}^2)|} = \frac{\sigma_{max}^2(\mathbf{M})}{\sigma_{min}^2(\mathbf{M})},$$

[10] *For any vector $\mathbf{x} \neq \mathbf{0}$ and invertible matrix $\mathbf{M}$ (i.e., $\mathbf{Mx} \neq \mathbf{0}$), the product $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ is positive semi-definite since $\mathbf{x}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{Mx} = \|\mathbf{Mx}\|^2 \geq 0$ [54].*

where λ's are eigenvalues. Note also that $\mathbf{A}_{\mathcal{I}}^{\dagger}\mathbf{A}_{\mathcal{I}}$ is invertible because $\mathbf{A}_{\mathcal{I}}$ has linearly independent columns and that $\mathbf{A}_{\mathcal{I}}^{\dagger}\mathbf{A}_{\mathcal{I}}$ and $\mathbf{A}_{\mathcal{I}}\mathbf{A}_{\mathcal{I}}^{\dagger}$ have the same non-zero eigenvalues (which are positive).

Remark 16. *(Eigenvalues of $\mathbf{A}_{\mathcal{I}}^{\dagger}\mathbf{A}_{\mathcal{I}}$) The origin of the effective conditioning of $\mathbf{M}$ is due to the (square-root) of the eigenvalues of $\mathbf{A}_{\mathcal{I}}^{\dagger}\mathbf{A}_{\mathcal{I}}$. We can further identify the eigenvalues of $\mathbf{A}_{\mathcal{I}}^{\dagger}\mathbf{A}_{\mathcal{I}}$ in terms of partial Fourier matrices.*

Specifically, without loss of generality, the matrix $\mathbf{A}_{\mathcal{I}}$ has block form $\mathbf{A}_{\mathcal{I}} = (\hat{\mathbf{F}}_p \ \hat{\mathbf{I}}_p)$ where $\hat{\mathbf{F}}$ is a subset of the columns of the DFT matrix $\frac{1}{\sqrt{n}}\mathbf{F} \in \mathbb{C}^{n\times\alpha}$, and $\hat{\mathbf{I}}_p \in \mathbb{R}^{n\times\beta}$ is a subset of the columns of the matrix $\mathbf{I}_-$. The number of columns $\alpha+\beta = |\mathcal{I}| \leq n$ is the size of the support of the inactive set (which is bounded by n since the columns of $\mathbf{A}_{\mathcal{I}}$ are linearly independent). Hence $\hat{\mathbf{F}}_p^{\dagger}\hat{\mathbf{F}}_p = \mathbf{I}_\alpha$ and $\hat{\mathbf{I}}_p^T\hat{\mathbf{I}}_p = \mathbf{I}_\beta$ have orthonormal columns. Introduce $\mathbf{G} := \mathbf{I}_p^T\hat{\mathbf{F}}_p \in \mathbb{C}^{\alpha\times\beta}$. A simple calculation shows that (dropping subscribes α and β on the identity matrices):

$$\mathbf{A}_{\mathcal{I}}^{\dagger}\mathbf{A}_{\mathcal{I}} = \begin{pmatrix} \mathbf{I} & \mathbf{G}^{\dagger} \\ \mathbf{G} & \mathbf{I} \end{pmatrix}, \tag{5.34}$$

The maximum and minimum eigenvalues of (5.34) can be worked out directly in terms of the norm $\|\mathbf{G}\|$. Specifically, since the eigenvalues of

$$\begin{pmatrix} \mathbf{0} & \mathbf{G}^{\dagger} \\ \mathbf{G} & \mathbf{0} \end{pmatrix}$$

are $\pm\sigma(\mathbf{G})$ where σ are the singular values of $\mathbf{G}$, we have

$$\lambda_{\max}(\mathbf{A}_{\mathcal{I}}^{\dagger}\mathbf{A}_{\mathcal{I}}) = 1 + \|\mathbf{G}\|, \qquad \lambda_{\min}(\mathbf{A}_{\mathcal{I}}^{\dagger}\mathbf{A}_{\mathcal{I}}) = 1 - \|\mathbf{G}\|$$

which yields

$$\kappa(\mathbf{A}_{\mathcal{I}}^{\dagger}\mathbf{A}_{\mathcal{I}}) = \frac{1+\|\mathbf{G}\|}{1-\|\mathbf{G}\|}, \qquad \kappa(\mathbf{M}) = \sqrt{\frac{1+\|\mathbf{G}\|}{1-\|\mathbf{G}\|}}. \tag{5.35}$$

The matrix $\mathbf{G}$ *is a partial Fourier matrix — it is comprised of sampling the orthonormal matrix* $\frac{1}{\sqrt{n}}\mathbf{F}$ *at columns and rows related to the support of the extreme point solution* $\boldsymbol{z}^*$*. In other words, it is a submatrix of* $\frac{1}{\sqrt{n}}\mathbf{F}$*. Hence* $\|\mathbf{G}\| \leq 1$ *and the numerator in* (5.35) *is bounded by* 2*. Therefore, any potential ill-conditioning of* $\kappa(\mathbf{M})$ *depends on whether* $\|\mathbf{G}\|$ *stays bounded away from* 1 *as* $n \to \infty$*.*

5.4.1 Test Case when $F_R(x)$ is One Dirac Mass

We consider the potential $w_{PM}(x)$ introduced in (3.16) with values $(G, L) = (2, 1.5)$ which yields a solution $F_R(x)$ that is one Dirac mass. For example, when $n = 2^2$, the IPM converges in 16 Newton iterations (with parameters $\epsilon_{\text{NW}} = \epsilon_{\hat{\eta}} = 10^{-4}$, and $\epsilon_{\text{MR}} = 10^{-6}$). Below is the numerical solution $\mathbf{z}_t^*$ at step 17 substituted into the equation $\mathbf{A}\mathbf{z}_t^* = \mathbf{b}$:

$$\underbrace{\begin{pmatrix} 0.5 & 0.5 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0.5 & -0.5i & -0.5 & 0.5i & -1 & 0 & 0 \\ 0.5 & -0.5 & 0.5 & -0.5 & 0 & -1 & 0 \\ 0.5 & 0.5i & -0.5 & -0.5i & 0 & 0 & -1 \end{pmatrix}}_{\mathbf{A}} \times \underbrace{\begin{pmatrix} 3.99981 \\ 0.00007 \\ 0.00005 \\ 0.00007 \\ 1.99988 \\ 1.99986 \\ 1.99988 \end{pmatrix}}_{\mathbf{z}_t^*} = \underbrace{\begin{pmatrix} 2 \\ 0 \\ 0 \\ 0 \end{pmatrix}}_{\mathbf{b}} \tag{5.36}$$

Equation (5.36) shows that the inactive set is converging to $\mathcal{I} = \{1, 5, 6, 7\}$. The submatrix $\mathbf{A}_{\mathcal{I}}$ of $\mathbf{A}$ derives from choosing columns of $\mathbf{A}$ related to the inactive set:

$$\mathbf{A}_{\mathcal{I}} = \begin{pmatrix} 0.5 & 0 & 0 & 0 \\ 0.5 & -1 & 0 & 0 \\ 0.5 & 0 & -1 & 0 \\ 0.5 & 0 & 0 & -1 \end{pmatrix}, \quad \text{also} \quad \mathbf{A}_{\mathcal{I}}^{\dagger}\mathbf{A}_{\mathcal{I}} = \begin{pmatrix} 1 & -0.5 & -0.5 & -0.5 \\ -0.5 & 1 & 0 & 0 \\ -0.5 & 0 & 1 & 0 \\ -0.5 & 0 & 0 & 1 \end{pmatrix}$$

Table 5.1 Conditioning and Singular Values versus Problem Size when $F_R(\mathbf{x})$ is one Dirac Mass

n	$\kappa(\mathbf{M})$	$\sigma_{max}(\mathbf{A}_\mathcal{I})$	$\sigma_{min}(\mathbf{A}_\mathcal{I})$
2^2	3.73	1.366	0.366
2^3	5.47	1.391	0.254
2^4	7.87	1.403	0.178
2^5	11.22	1.408	0.125
2^6	15.94	1.411	0.088
2^7	22.58	1.413	0.062
2^8	31.97	1.413	0.044
2^9	45.23	1.414	0.031
2^{10}	63.98	1.414	0.022
2^{11}	90.445	1.414	0.016
2^{12}	127.99	1.414	0.011
2^{13}	181.03	1.414	0.0078

For arbitrary values of n, the inactive set $\mathcal{I}$ of discrete vectors that converge to $F_R(x) = \delta(x)$ is:

$$\mathcal{I} = \{1, n+1, n+2, \cdots, 2n-1\}.$$

Using the inactive set $\mathcal{I}$, we numerically compute $\mathbf{A}_\mathcal{I}$, the singular values, and corresponding corresponding conditioning number $\kappa(\mathbf{M})$. Table 5.1 shows the scaling of the conditioning number of $\mathbf{A}_\mathcal{I}$ for different values of n. Figure 5.1 plots $\kappa(M) \sim 2\sqrt{n}$ with the asymptotic behavior.

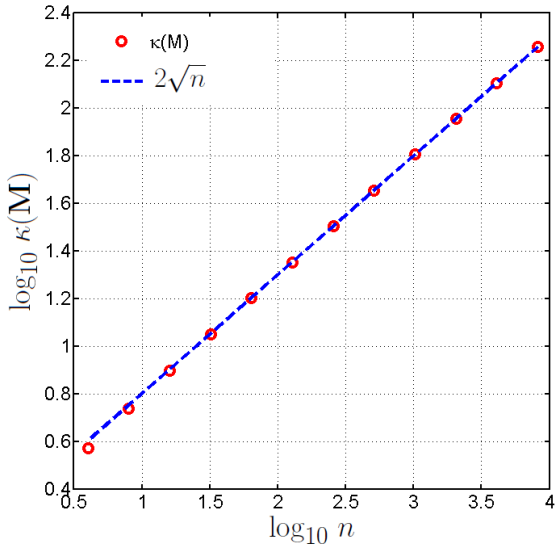

Figure 5.1 Condition number $\kappa(\mathbf{M}) \sim 2\sqrt{n}$ versus problem size n when the solution $F_R(\mathbf{x})$ is one Dirac mass.

5.4.2 Test Case when $F_R(x)$ is Two Dirac Masses

We consider the potential $w_{PM}(x)$ introduced in (3.16) with values $(G, L) = (3, 0.2)$ which yields a solution $F_R(x)$ that is two Dirac masses. For example, when $n = 2^2$, the IPM converges in 16 Newton iterations (with parameters $\epsilon_{\text{NW}} = \epsilon_{\hat{\eta}} = 10^{-4}$, and $\epsilon_{\text{MR}} = 10^{-6}$). Below is the numerical solution $\mathbf{z}_t^*$ at step 20 substituted into the equation $\mathbf{A}\mathbf{z}_t^* = \mathbf{b}$:

$$\underbrace{\begin{pmatrix} 0.5 & 0.5 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0.5 & -0.5i & -0.5 & 0.5i & -1 & 0 & 0 \\ 0.5 & -0.5 & 0.5 & -0.5 & 0 & -1 & 0 \\ 0.5 & 0.5i & -0.5 & -0.5i & 0 & 0 & -1 \end{pmatrix}}_{\mathbf{A}} \times \underbrace{\begin{pmatrix} 2.00005 \\ 0.00001 \\ 1.99993 \\ 0.00001 \\ 0.00006 \\ 1.99998 \\ 0.00006 \end{pmatrix}}_{\mathbf{z}_t^*} = \underbrace{\begin{pmatrix} 2 \\ 0 \\ 0 \\ 0 \end{pmatrix}}_{\mathbf{b}} \tag{5.37}$$

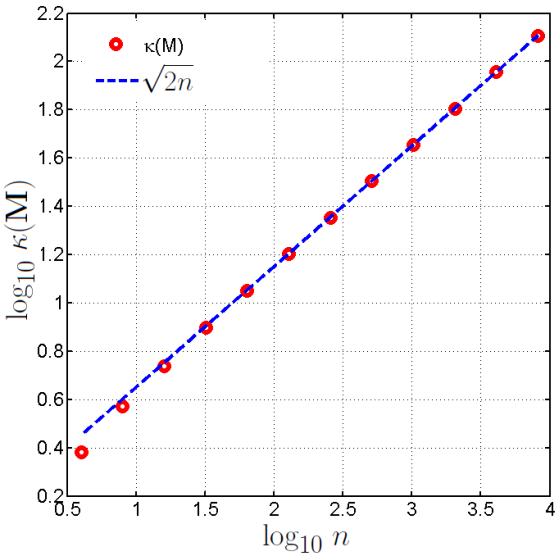

Figure 5.2 Condition number $\kappa(\mathbf{M}) \sim \sqrt{2n}$ versus problem size n when the solution $F_R(\mathbf{x})$ is two Dirac masses.

Equation (5.37) shows that the inactive set is converging to $\mathcal{I} = \{1, 3, 6\}$. The submatrix $\mathbf{A}_{\mathcal{I}}$ of $\mathbf{A}$ derives from choosing columns of $\mathbf{A}$ related to the inactive set:

$$\mathbf{A}_{\mathcal{I}} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & -0.5 & 0 \\ 0.5 & 0.5 & -1 \\ 0.5 & -0.5 & 0 \end{pmatrix}, \quad \text{also,} \quad \mathbf{A}_{\mathcal{I}}^{\dagger}\mathbf{A}_{\mathcal{I}} = \begin{pmatrix} 1 & 0 & -0.5 \\ 0 & 1 & -0.5 \\ -0.5 & -0.5 & 1 \end{pmatrix}. \tag{5.38}$$

The inactive set in the case when $F_R(\mathbf{x})$ is two Dirac masses for general n has the form:

$$\mathcal{I} = \left\{1, \frac{n}{2} + 1, n + 2, n + 4, \cdots, 2n - 1\right\}.$$

Table 5.2 and Figure 5.2 show $\kappa(\mathbf{M})$ which is computed using singular values corresponding to the inactive sets of $\mathbf{A}$ for the case where $F_R(\mathbf{x})$ is two Dirac masses. Note there is a minor change in the condition number of the submatrix $\mathbf{M}$ changes from $\kappa(\mathbf{M}) \sim 2\sqrt{n}$ for the case when $F_R(\mathbf{x})$ is one Dirac mass to $\kappa(\mathbf{M}) \sim \sqrt{2n}$ for case when $F_R(\mathbf{x})$ is two Dirac masses.

Table 5.2 Conditioning and Singular Values versus Problem Size when $F_R(\mathbf{x})$ is Two Dirac Masses

n	$\kappa(\mathbf{M})$	$\sigma_{max}(\mathbf{A}_{\mathcal{I}})$	$\sigma_{min}(\mathbf{A}_{\mathcal{I}})$
2^2	2.41	1.306	0.541
2^3	3.73	1.366	0.366
2^4	5.47	1.391	0.254
2^5	7.87	1.402	0.178
2^6	11.22	1.409	0.125
2^7	15.94	1.411	0.088
2^8	22.58	1.413	0.062
2^9	31.97	1.414	0.044
2^{10}	45.23	1.414	0.031
2^{11}	63.98	1.414	0.022
2^{12}	90.45	1.414	0.016
2^{13}	127.99	1.414	0.011

5.4.3 Test Case when $F_R(x)$ is Four Dirac Masses

We also consider a test case when the numerics converge to a solution $F_R(x)$ that is four Dirac masses. In order to check the singular value of $\mathbf{A}_{\mathcal{I}}$, as before, we identify the inactive set $\mathcal{I}$ when $n = 2^3$:

$$\mathcal{I} = \{1, 3, 5, 7, 12\}\,.$$

From $\mathcal{I}$ we have

$$\mathbf{A}_{\mathcal{I}} = \begin{pmatrix} \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & 0 \\ \frac{1}{\sqrt{8}} & \frac{-i}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & \frac{i}{\sqrt{8}} & 0 \\ \frac{1}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & 0 \\ \frac{1}{\sqrt{8}} & \frac{i}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & \frac{-i}{\sqrt{8}} & 0 \\ \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & -1 \\ \frac{1}{\sqrt{8}} & \frac{-i}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & \frac{i}{\sqrt{8}} & 0 \\ \frac{1}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & 0 \\ \frac{1}{\sqrt{8}} & \frac{i}{\sqrt{8}} & \frac{-1}{\sqrt{8}} & \frac{-i}{\sqrt{8}} & 0 \end{pmatrix}. \tag{5.39}$$

We can see that columns 2 and 4 are complex conjugates. Table 5.3 and Figure 5.3 show the conditioning numbers, and indicates that condition number in this case scales like $\mathcal{O}(\sqrt{n})$.

The fact that $\kappa(\mathbf{M}) \sim \sqrt{n}$ is not unreasonable and gives an expected bound on the MINRES algorithm to converge in $\mathcal{O}(\sqrt{n})$ iterations. There may in fact be a tighter bound.

The singular values in Figure 5.3 are obtained by considering $\mathbf{A}_{\mathcal{I}}$ as a matrix over complex vectors, i.e., $z_j \in \mathbb{C}$. Incorporating the additional linear constraints that z_j is real, induces a restriction that $\mathbf{A}_{\mathcal{I}}$ act on symmetric vectors $\mathbf{z}$. In the case of $n = 8$, we have $z_3 = z_7$ in solving the problem $\mathbf{Az} = \mathbf{b}$. Therefore the modified

Table 5.3 Conditioning and Singular Values versus Problem Size when $F_R(\mathbf{x})$ is Four Dirac Masses

n	$\kappa(\mathbf{M})$	$\sigma_{max}(\mathbf{A}_\mathcal{I})$	$\sigma_{min}(\mathbf{A}_\mathcal{I}) \neq 0$
2^3	2.41	1.31	0.54
2^4	3.73	1.37	0.37
2^5	5.47	1.39	0.25
2^6	7.87	1.4	0.18
2^7	11.22	1.41	0.12
2^8	15.94	1.41	0.09
2^9	22.58	1.41	0.06
2^{10}	31.97	1.41	0.44
2^{11}	45.23	1.41	0.31
2^{12}	63.98	1.41	0.02
2^{13}	90.50	1.41	0.01

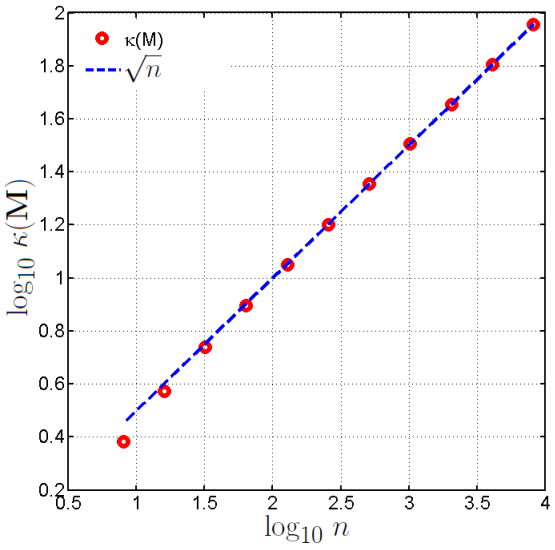

Figure 5.3 Condition number $\kappa(\mathbf{M}) \sim \sqrt{n}$ versus problem size n when the solution $F_R(\mathbf{x})$ is four Dirac masses.

form of $\mathbf{A}_{\mathcal{I}}$ over this (symmetric) vector space is:

$$\mathbf{A}_{\mathcal{I}}^{*} = \begin{pmatrix} \frac{1}{\sqrt{8}} & 2 & \frac{1}{\sqrt{8}} & 0 \\ \frac{1}{\sqrt{8}} & 0 & \frac{-1}{\sqrt{8}} & 0 \\ \frac{1}{\sqrt{8}} & -2 & \frac{1}{\sqrt{8}} & 0 \\ \frac{1}{\sqrt{8}} & 0 & \frac{-1}{\sqrt{8}} & 0 \\ \frac{1}{\sqrt{8}} & 2 & \frac{1}{\sqrt{8}} & -1 \\ \frac{1}{\sqrt{8}} & 0 & \frac{-1}{\sqrt{8}} & 0 \\ \frac{1}{\sqrt{8}} & -2 & \frac{1}{\sqrt{8}} & 0 \\ \frac{1}{\sqrt{8}} & 0 & \frac{-1}{\sqrt{8}} & 0 \end{pmatrix} \tag{5.40}$$

The matrix $\mathbf{A}_{\mathcal{I}}^{*}$ has (non-zero) singular values $\sigma(\mathbf{A}_{\mathcal{I}}^{*}) = \{4.39, 3.18, 2.83, 1.62\}$. Table 5.4 and Figure 5.4 show that in the case of using the $\mathbf{A}_{\mathcal{I}}^{*}$ the largest and smallest (non-zero) singular values are bounded independent of n, so that the effective condition number is bounded as $n \to \infty$. This suggests that the convergence of MINRES may be independent of the problem size n (which would be good). In the following sections, we perform a numerical study to determine the practical performance of the preconditioners.

5.5 Performance of the Preconditioners: Numerical Study

This section presents a numerical investigation for different preconditioners used to solve Equation (5.8) with the primal-dual interior-point method. In each Newton iteration, the MINRES algorithm with different choices of preconditioners is used to solve Equation (5.21). Of particular interest is the total number of matrix-vector products (MATVECs), added up over all the Newton iterations of the primal-dual algorithm, required by the different preconditioners to compute a solution to a given accuracy. This is because the total computational complexity scales with the number of MATVECs (each MATVEC costing $\mathcal{O}(n \log n)$ flops).

Table 5.4 Restricted Conditioning and Singular Values versus Problem Size when $F_R(\mathbf{x})$ is Four Dirac Masses

n	$\kappa(\mathbf{M}^*)$	$\sigma_{max}(\mathbf{A}_{\mathcal{I}}^*)$	$\sigma_{min}(\mathbf{A}_{\mathcal{I}}^*)$
2^3	3.13	2.04	0.65
2^4	3.90	2.06	0.53
2^5	4.48	2.07	0.46
2^6	4.86	2.07	0.43
2^7	5.09	2.08	0.41
2^8	5.21	2.08	0.40
2^9	5.28	2.08	0.39
2^{10}	5.31	2.08	0.39
2^{11}	5.33	2.08	0.39
2^{12}	5.34	2.08	0.39
2^{13}	5.34	2.08	0.39

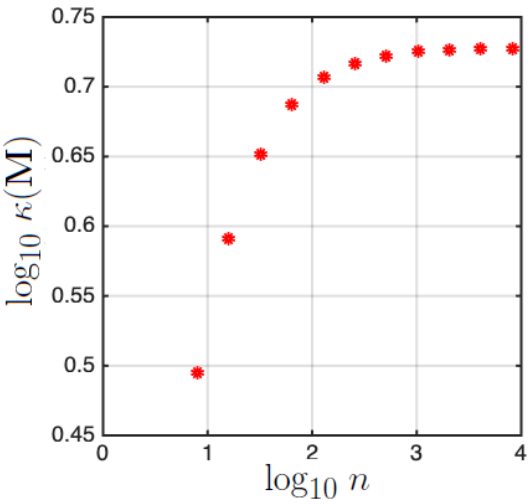

Figure 5.4 Condition number for the matrix $\mathbf{A}_{\mathcal{I}}$ restricted to symmetric vectors $\mathbf{z}$, i.e., $\kappa(\mathbf{A}^*)$ versus problem size n. The case if for a $F_R(\mathbf{x})$ that is four Dirac masses. The restricted condition number is bounded an converges as $n \to \infty$.

Note that all the results in this section are for a periodic Morse type potential as defined in (3.16), i.e.,

$$w(\mathbf{x}) = -GLe^{-\frac{1}{L}\sin(\pi|\mathbf{x}|)} + e^{-\sin(\pi|\mathbf{x}|)}, \qquad G, L > 0.$$

To compare the performance of $\mathbf{P_1}$ and $\mathbf{P_2}$ in the matrix-free primal-dual algorithm we use the following criteria.

- Stopping criteria for the primal-dual Newton iteration loop: $\|\mathbf{f}_t - \mathbf{f}_{ref}\|_\infty < \epsilon_{\text{NW}}$ and $\hat{\eta} < \epsilon_{\hat{\eta}}$ where $\mathbf{f}_{ref}$ (computed via MATLAB) is a precomputed reference solution. Note that this is different than the criteria stated in Algorithm 1 (which does not require knowledge of the solution) that $\|\mathbf{z}_i - \mathbf{z}_{i-1}\|_\infty < \epsilon_{\text{NW}}$.

- Stopping criteria for the MINRES linear solver: we use $\|\tilde{\mathbf{f}}\| < \epsilon_{\text{MR}}$ where $\tilde{\mathbf{e}} = \mathbf{B}\tilde{d} - \mathbf{b}$ is the residual. The MINRES routine appears inside the primal-dual Newton loop. To ensure that the MINRES error does not impact the Newton iterations we take $\epsilon_{\text{MR}} \ll \epsilon_{\text{NW}}$.

- The parameters in the interior-point algorithm throughout this section are taking as follows:

  - Tolerance for MINRES algorithm: $\epsilon_{\text{MR}} = 10^{-8}$;
  - Tolerance for the duality gap: $\epsilon_{\hat{\eta}} = 10^{-2}$;
  - Tolerance for Newton's method: $\epsilon_{\text{NW}} = 10^{-2}$.

5.5.1 Choice of the Centering Parameter μ

The parameter t in the primal-dual interior-point method depends on a factor $\mu > 1$ (centering parameter). In this section two empirical studies are performed to quantify the effect of μ on the convergence and performance of the matrix-free primal-dual interior-point method.

Figure 5.5 compares how the number of MATVECs and number of Newton iterations in the matrix-free primal-dual algorithm scale with different values of μ. Since μ roughly controls the geometric rate that $t \to 0$, as expected, smaller values of μ decrease t slowly and require more Newton iterations. Large values of μ create large changes in the effective KKT equations in each iteration and also require more Newton steps.

As we can see from the Figure 5.5, choosing a parameter $\mu = 80$, the total number of MATVECs is (roughly) less than other values. Therefore, for the rest of the computations in this section we take $\mu = 80$.

Figure 5.6 shows the convergence of the primal-dual interior-point method variables versus the number of Newton iterations for different values of μ — and explains our preference for choosing a parameter $\mu \geq 10$. We can see from the Figure 5.6 that the primal-dual algorithm converges faster using $\mu = 10$ (left plot), and, the surrogate duality gap decreases faster for $\mu = 10$ (middle plot). Also, choosing a very small parameter like $\mu = 1.5$, makes the algorithm converge in more Newton steps, and causes the surrogate duality gap to decrease slower.

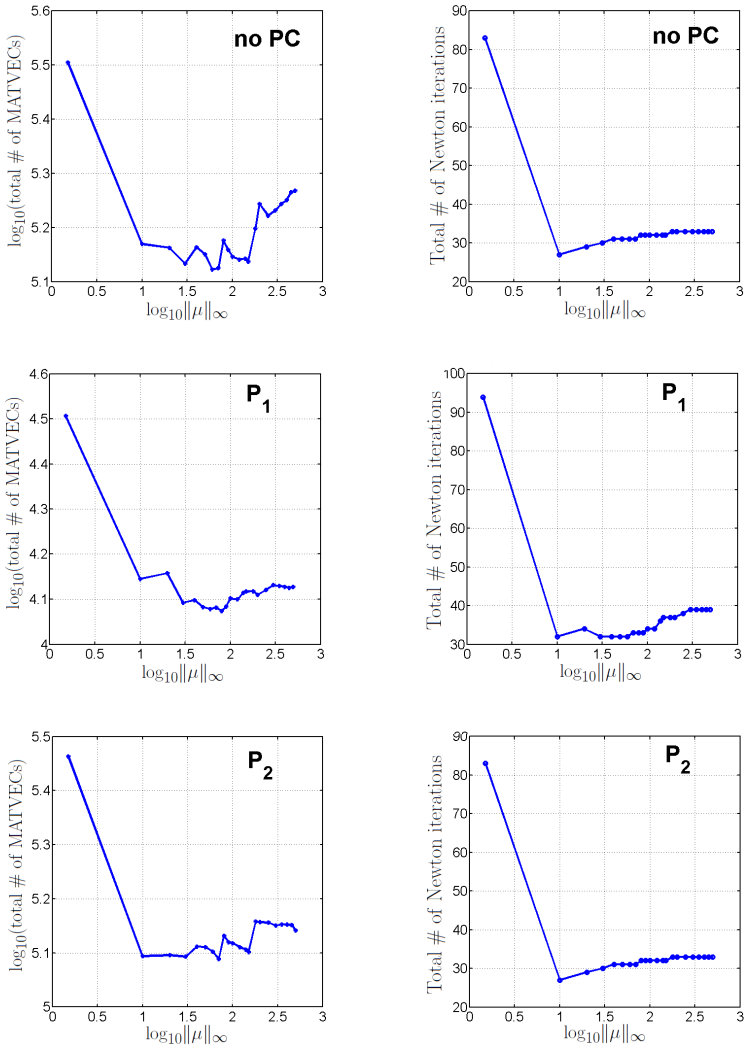

Figure 5.5 Trade-off in choice of parameter μ in the primal-dual interior-point algorithm in solving the problem (5.21) using no preconditioner (top row), preconditioner $\mathbf{P_1}$ (middle) and preconditioner $\mathbf{P_2}$ (bottom). The test case uses (3.16) with $(G, L) = (0.9, 1.5)$ yielding an $F_R(x)$ that is a continuous function. Test parameters are: $n = 2^8$, $\epsilon_{\text{NW}} = \epsilon_{\hat{\eta}} = 10^{-2}$, $\epsilon_{\text{MR}} = 10^{-8}$.

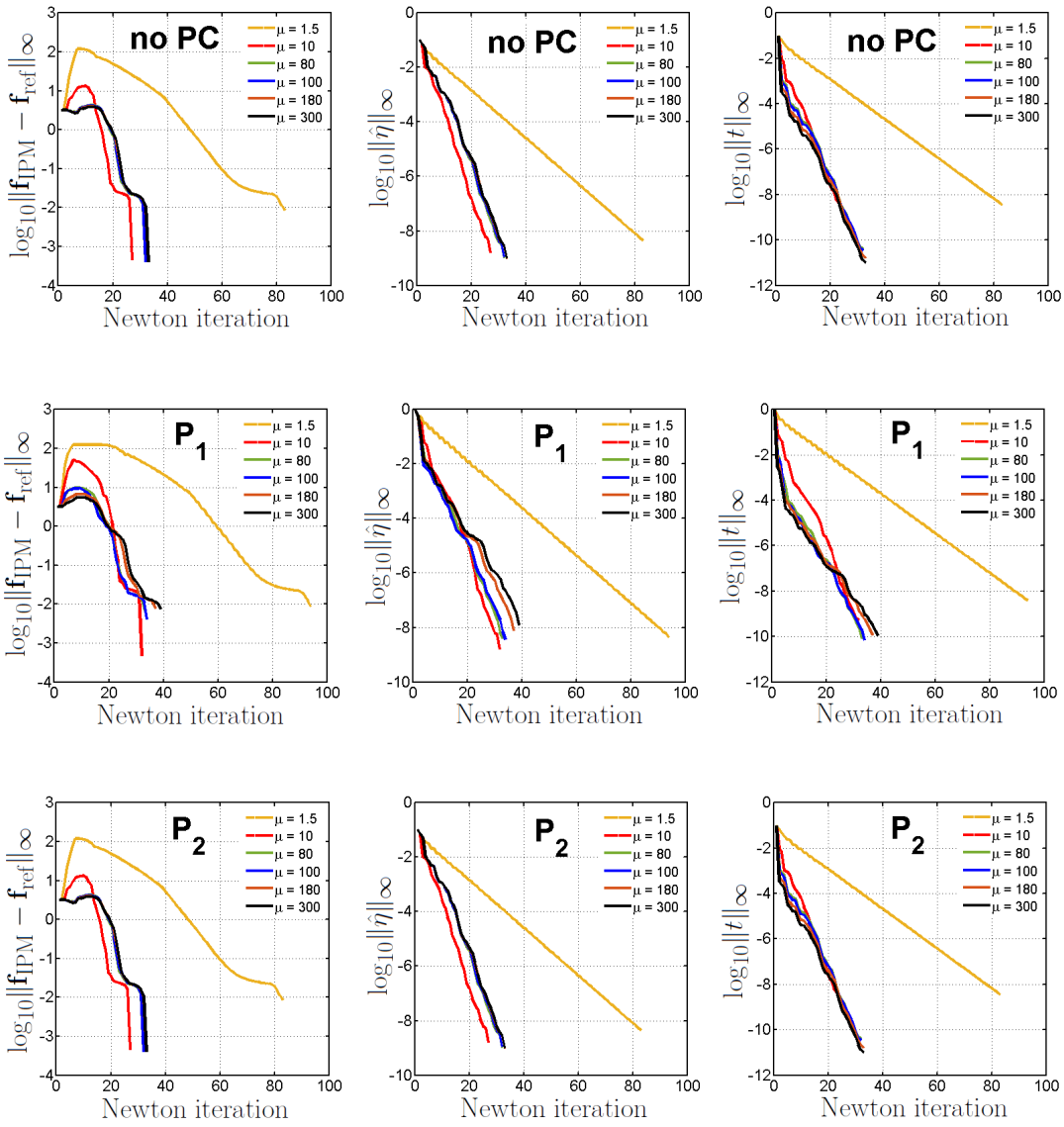

Figure 5.6 Convergence of the primal-dual interior-point algorithm for different parameters μ: Convergence of the solution v.s. Newton steps (left); Surrogate duality gap ($\hat{\eta}$) v.s. Newton steps (middle); Interior parameter (t) v.s. Newton steps (right). The results shown are using no preconditioners (top row), and preconditioners $\mathbf{P_1}$, $\mathbf{P_2}$ (middle and bottom rows respectively). The test case uses (3.16) with $(G, L) = (0.9, 1.5)$ yielding an $F_R(x)$ that is a continuous function. Test parameters are: $n = 2^8$, $\epsilon_{\text{NW}} = \epsilon_{\hat{\eta}} = 10^{-2}$, $\epsilon_{\text{MR}} = 10^{-8}$. The reference solution $\mathbf{f}_{ref}$ is computed (to high accuracy) using MATLABs optimization routine.

5.5.2 Test Case when $F_R(\mathbf{x})$ is a Continuous Function

This section tests the performance of the different preconditioners when the solution $F_R(x)$ is a continuous function (obtained when the parameters in $w_{PM}(x)$ are $(G, L) = (0.9, 1.5)$).

Figure 5.7 presents the numerical solution for $F_R(x)$ (right) for different preconditioners, and demonstrates that all three preconditioners are able to obtain the solution at the present accuracy (and appear indistinguishable under visual inspection). Figure 5.7 (left), provides a histogram plotting the number of Newton iterations required against the number of MATVECs. The histogram hints at the computational advantages of $\mathbf{P_1}$ compared to $\mathbf{P_2}$. Specifically, $\mathbf{P_1}$ has more Newton iterations that require fewer MATVECs.

Figure 5.8 compares the convergence (in sup norm of the solution error, and primal-dual parameter t) in the solution of (5.21). The preconditioners arise in the MINRES computation, and only mildly impact (through the choice of MINRES tolerance) the resulting increments $\mathbf{\Delta z}, \mathbf{\Delta s}, \mathbf{\Delta}\lambda$. This is why there is little difference between the panels in Figure 5.8 — the three preconditioners travel (roughly) along the same central path and converge at almost the same rates. Three points along the central path at Newton iterations 1, 16 and 32, shown by red dots in Figure 5.8, and selected to investigate the performance of the MINRES solver. The companion Figure 5.9 plots the MINRES residual, i.e. $\tilde{\mathbf{e}} = \mathbf{Bd} - \mathbf{b}$ from (5.21), versus the number of MINRES iterations. Figure 5.9 clearly shows that at every Newton step, the MINRES algorithm using preconditioner $\mathbf{P_1}$ converges at the fastest rate, thereby requiring fewer MATVECs than either $\mathbf{P_2}$ or no preconditioner.

Finally, Figure 5.10 compares the total number of Newton iterations and MATVECSs of the three preconditioners needed to compute the solution for different problem sizes n. The figure shows that, in practice, the preconditioner $\mathbf{P_1}$ provides a significant improvement in the slope of the number of MATVECs versus n, thereby

improving computational cost in large problems. Note that for the purposes of the test in Figure 5.10, we consider each value of n as an independent optimization problem. In reality, the problems (5.8) are different discretizations of the same underlying continuum problem. Hence, one could try to exploit this fact to improve the overall computational complexity.

This subsection demonstrated that the preconditioner $\mathbf{P_1}$ outperformed other preconditioners when $F_R(x)$ is a continuous function. In the following subsections we examine the performance when $F_R(x)$ is a different critical point (which may change the conditioning of the matrix $\mathbf{A}_\mathcal{I}$ and impact the performance of the matrix-free methods).

5.5.3 Test Case when $F_R(\mathbf{x})$ is One Dirac mass

In this subsection we perform a numerical test when $F_R(x)$ is one Dirac mass, specifically with $w_{PM}(x)$ and $(G, L) = (2, 1.5)$. Figure 5.11 compares the number of Newton iterations and MATVECs for the different preconditioners. The figure shows that, in practice, the preconditioner $\mathbf{P_1}$ provides a significant improvement in the slope of the number of MATVECs versus n, thereby improving computational cost in large problems.

5.5.4 Test Case when $F_R(\mathbf{x})$ is Two Dirac Masses

In this subsection we perform a numerical test when $F_R(x)$ is two Dirac masses, specifically with $w_{PM}(x)$ and $(G, L) = (3, 0.2)$. Figure 5.12 compares the number of Newton iterations and MATVECs for the different preconditioners. The figure shows that, in practice, the preconditioner $\mathbf{P_1}$ provides a significant improvement in the number of MATVECs versus n, thereby improving computational cost in large problems.

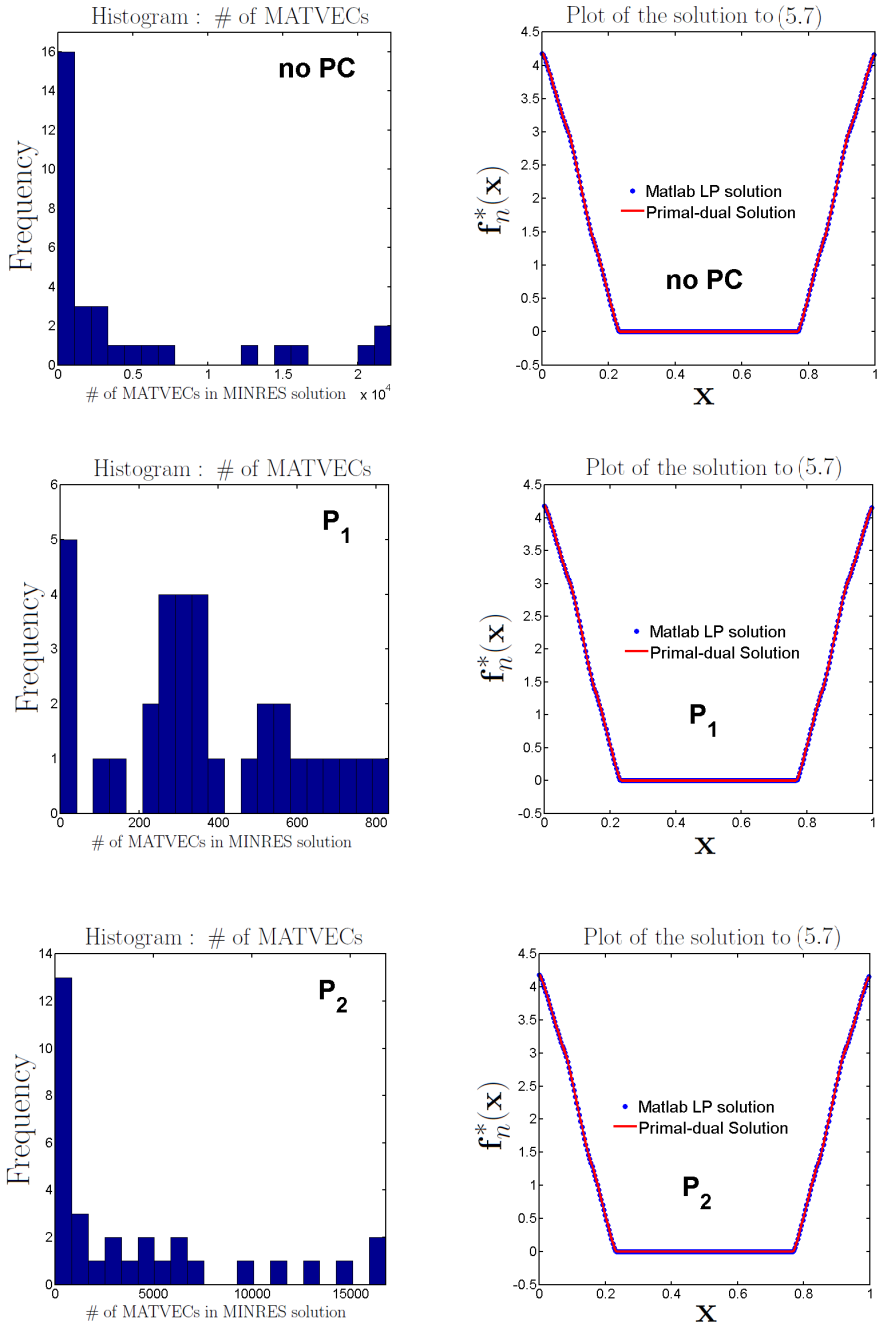

Figure 5.7 Comparison of the total number of MATVECs required to solve the primal-dual algorithm for the problem (5.7) using no preconditioner (top row), $\mathbf{P_1}$ (middle) and $\mathbf{P_2}$ (bottom). The histograms on the left shows $\mathbf{P_1}$ requires a much less number of MATVECs. The right plots visually show that solutions $\mathbf{f}(\mathbf{x})$ fully converge to a continuous function. The test parameters are: $(G, L) = (0.9, 1.5)$, $n = 2^8$, $\mu = 80$, $\epsilon_{\text{NW}} = \epsilon_{\hat{\eta}} = 10^{-2}$, $\epsilon_{\text{MR}} = 10^{-8}$.

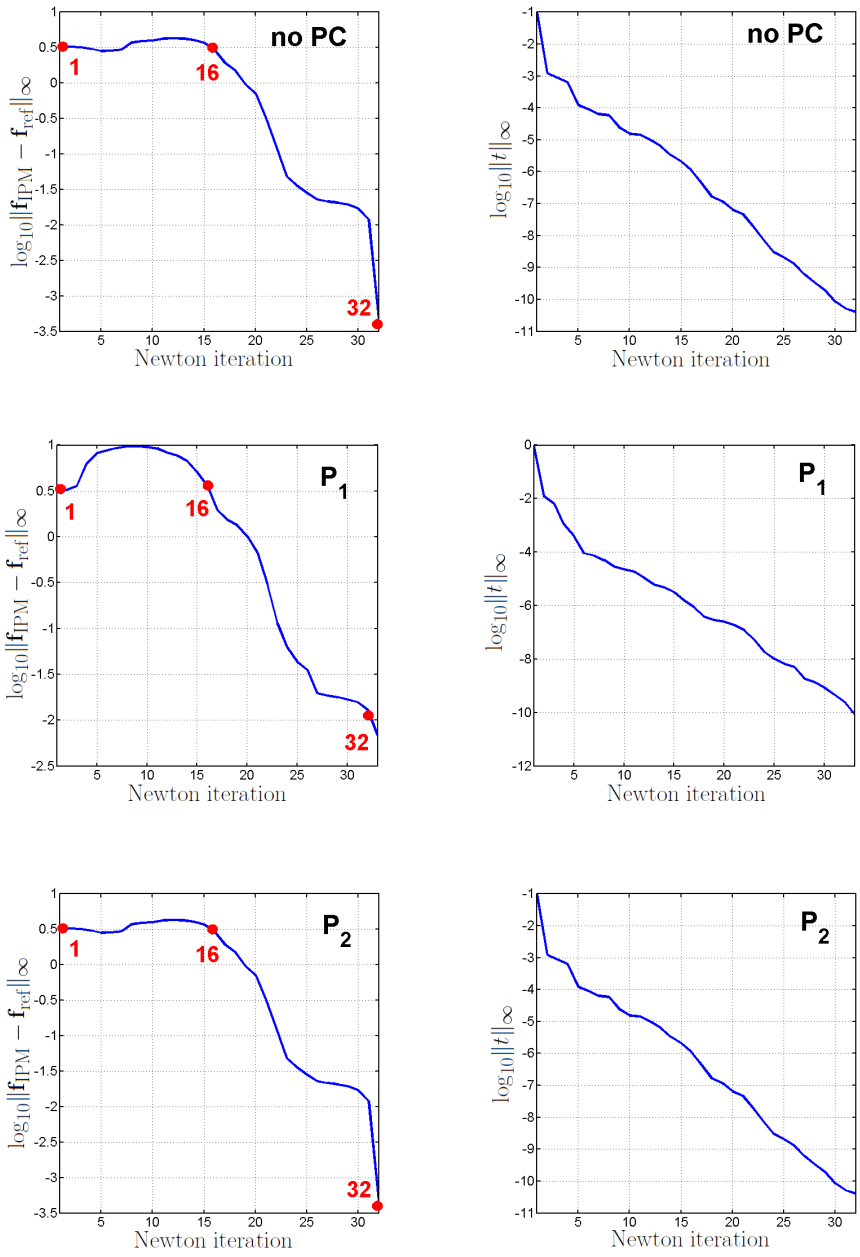

Figure 5.8 Interior-point method convergence versus Newton iteration when $F_R(x)$ is a continuous function, for no preconditioner, and preconditioners $\mathbf{P_1}$ and $\mathbf{P_2}$. The left figures plot the sup norm of the solution error with respect to a reference solution (computed via MATLAB). The right figures plot the convergence of the parameter t. Test parameters are: $(G, L) = (0.9, 1.5)$, $\epsilon_{\mathrm{NW}} = \epsilon_{\hat{\eta}} = 10^{-2}$, $\epsilon_{\mathrm{MR}} = 10^{-8}$.

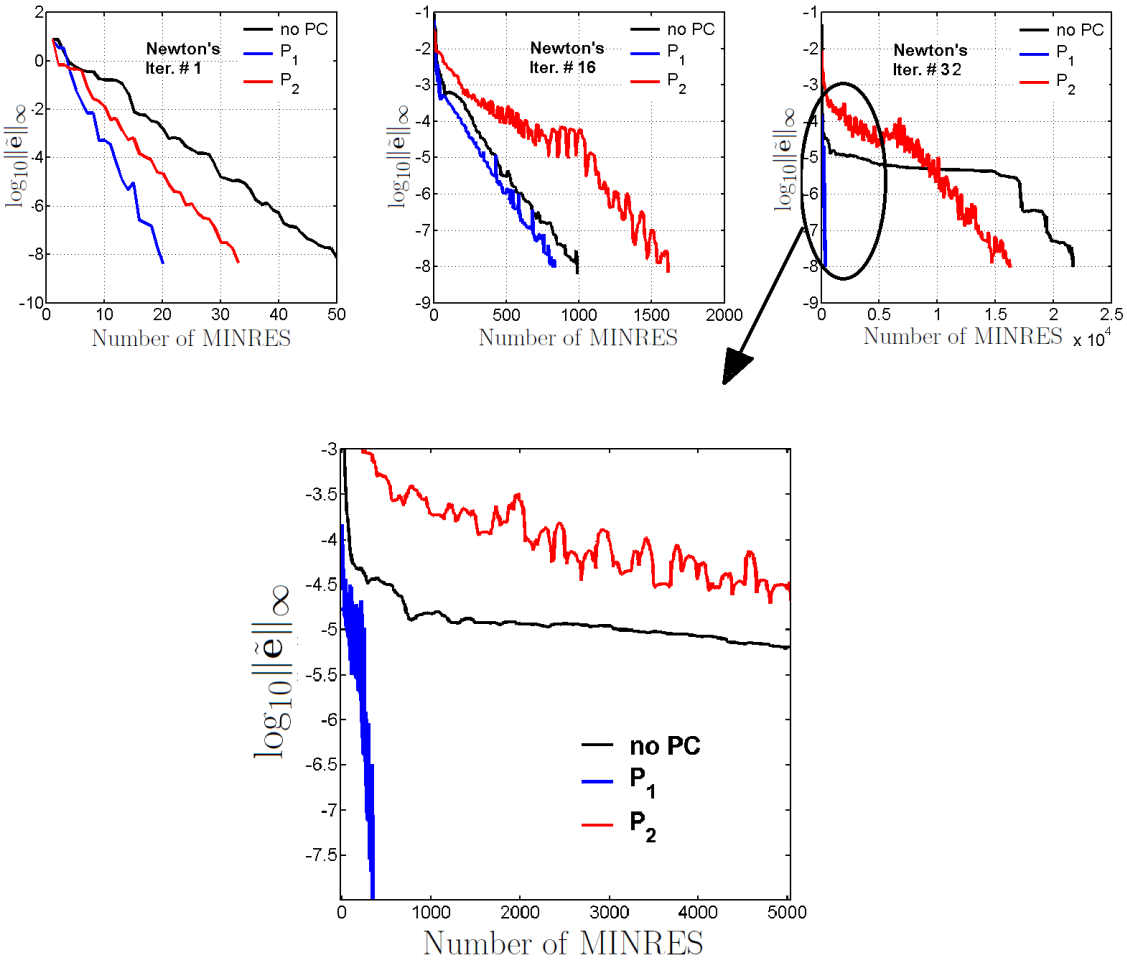

Figure 5.9 Comparison of the MINRES residual versus number of MINRES iterations to equation (5.21). The plots are for three points along the central path in the interior-point method (the points are the red circles in Figure 5.8). The three curves compare convergence without using a preconditioner and with using preconditioners $\mathbf{P_1}$ and $\mathbf{P_2}$. Test parameters are: $(G, L) = (0.9, 1.5)$ (which yields a continuous solution $F_R(x)$), $\epsilon_{\text{NW}} = \epsilon_{\hat{\eta}} = 10^{-2}$, $\epsilon_{\text{MR}} = 10^{-8}$. The preconditioner $\mathbf{P_1}$ outperforms the other preconditioners.

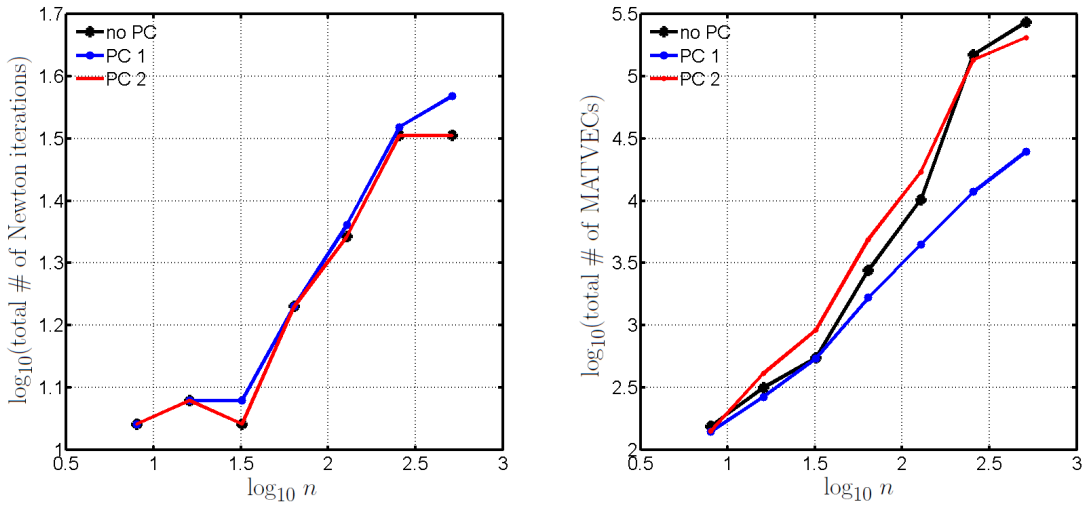

Figure 5.10 Performance comparison of different preconditioners when $F_R(x)$ is a continuous function. The number of Newton iterations (left) and MATVECs (right) required by matrix-free interior-point methods are plotted versus problem size n. The preconditioner $\mathbf{P_1}$ outperforms the other preconditioners by requiring fewer MATVECs. Test parameters are: $(G, L) = (0.9, 1.5)$, $\epsilon_{\text{NW}} = \epsilon_{\hat{\eta}} = 10^{-2}$, $\epsilon_{\text{MR}} = 10^{-8}$.

5.5.5 Test Case when $F_R(\mathbf{x})$ is Four Dirac Masses

In this subsection we perform a numerical test when $F_R(x)$ is four Dirac masses, specifically with $w_{PM}(x)$ and $(G, L) = (2, 0.15)$. Figure 5.13 compares the number of Newton iterations and MATVECs for the different preconditioners. The figure shows that, although the number of MATVECs versus n have similar slopes for $\mathbf{P_1}$ and $\mathbf{P_2}$, in practice, the preconditioner $\mathbf{P_1}$ provides a significant improvement in the total number of MATVECs versus n, thereby improving computational cost in large problems.

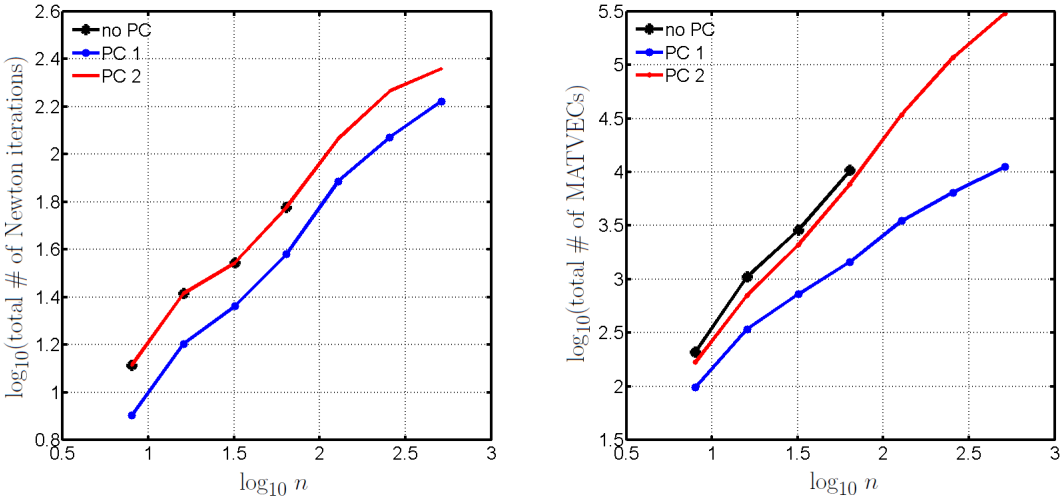

Figure 5.11 Performance comparison of different preconditioners when $F_R(x)$ is one Dirac mass. The number of Newton iterations (left) and MATVECs (right) required by matrix-free interior-point methods are plotted versus problem size n. The preconditioner $\mathbf{P_1}$ outperforms the other preconditioners by requiring fewer MATVECs. Test parameters are: $(G, L) = (2, 1.5)$, $\epsilon_{\text{NW}} = \epsilon_{\hat{\eta}} = 10^{-2}$, $\epsilon_{\text{MR}} = 10^{-8}$.

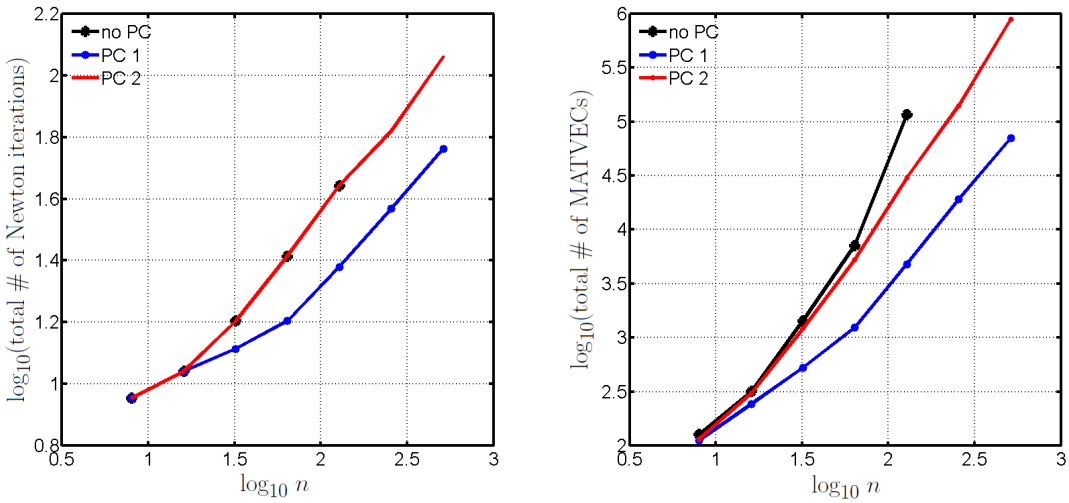

Figure 5.12 Performance comparison of different preconditioners when $F_R(x)$ is two Dirac masses. The number of Newton iterations (left) and MATVECs (right) required by matrix-free interior-point methods are plotted versus problem size n. The preconditioner $\mathbf{P_1}$ outperforms the other preconditioners by requiring fewer MATVECs. Test parameters are: $(G, L) = (3, 0.2)$, $\epsilon_{\text{NW}} = \epsilon_{\hat{\eta}} = 10^{-2}$, $\epsilon_{\text{MR}} = 10^{-8}$.

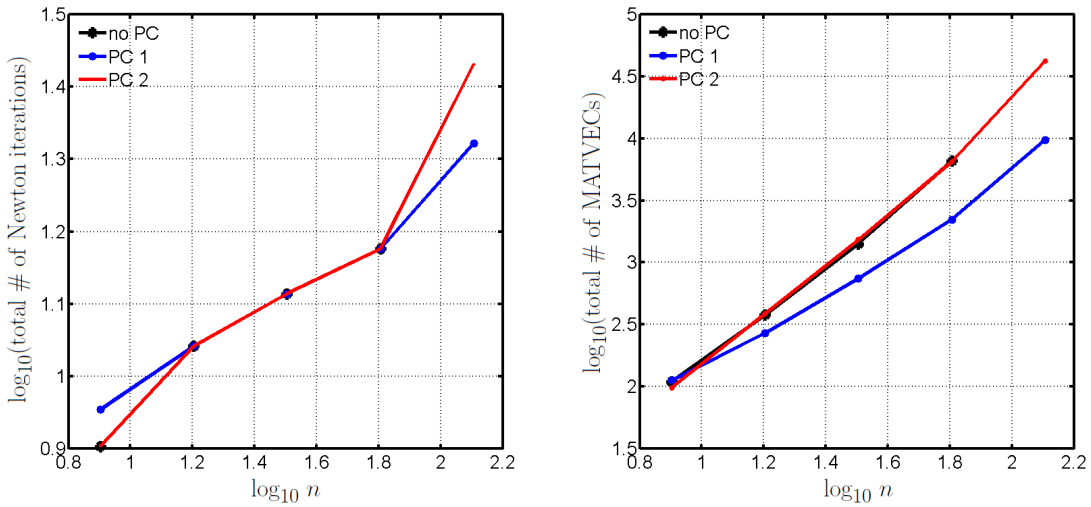

Figure 5.13 Performance comparison of different preconditioners when $F_R(x)$ is four Dirac masses. The number of Newton iterations (left) and MATVECs (right) required by matrix-free interior-point methods are plotted versus problem size n. The preconditioner $\mathbf{P_1}$ outperforms the other preconditioners by requiring fewer MATVECs. Test parameters are: $(G, L) = (2, 0.15)$, $\epsilon_{\text{NW}} = \epsilon_{\hat{\eta}} = 10^{-2}$, $\epsilon_{\text{MR}} = 10^{-8}$.

5.6 Convergence of Discrete Solution Under Mesh Refinement

In this section we investigate the convergence of the discrete solution $\mathbf{f}_n^*$ to the continuum problem $F_R(\mathbf{x})$, under the refinement of the grid, i.e., $n \to \infty$. Continuum variational problems give rise to a sequence of discrete optimization problems parameterized by the number of grid points n. The goal is to understand the limit as $n \to \infty$.

As mentioned in §5.2.2, the problem (5.7) can be found using the primal-dual interior-point method, and letting the parameter $t \to 0$. We already studied the convergence of the interior-point method as $t \to 0$ in §5.3.2.

In the continuum problem (R), $F_R(\mathbf{x})$ admits two types of solutions that have fundamentally different characteristics. In one case, we observe that $F_R(\mathbf{x})$ is a continuous (but nonsmooth) function; while in other cases we observe that $F_R(\mathbf{x})$ may contain Dirac masses and hence is not a classical function. When $F_R(\mathbf{x})$ is

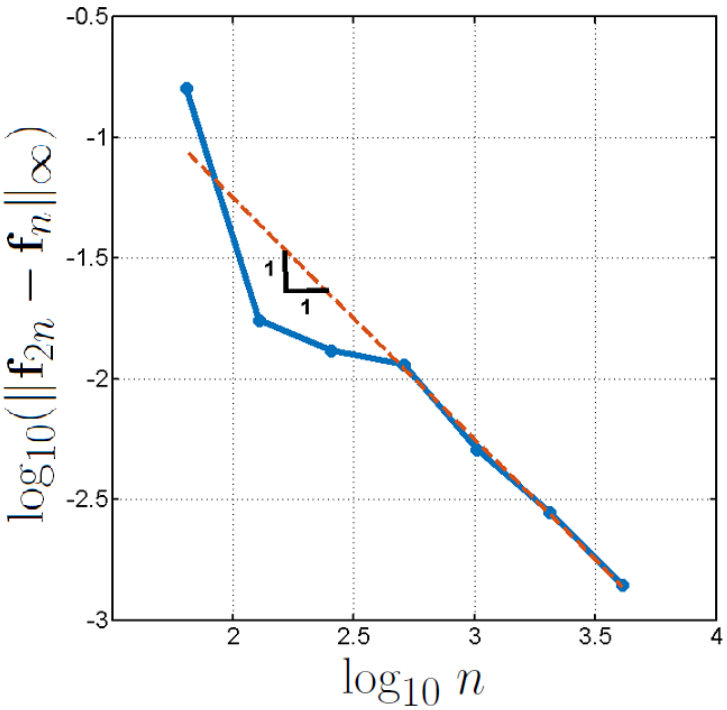

Figure 5.14 Linear convergence of solutions $\mathbf{f}_n^*$ to problem (5.7). As n gets large G_n converges to zero. The test case is for an interaction potential (3.16) with parameters $(G, L) = (0.9, 1.5)$ which yields a continuous minimizer $F_R(\mathbf{x})$ Tolerances for the interior-point algorithm are: $\epsilon_{\hat{\eta}} = 10^{-10}, \epsilon_{\text{MR}} = 10^{-12}$, which are small enough to remove any errors introduced by the interior-point algorithm.

continuous we will investigate a notion of strong convergence, i.e., does $\mathbf{f}_n^*$ converge uniformly to $F_R(\mathbf{x})$? In order to test the convergence of solutions to the problem (5.7), $\mathbf{f}_n^*$, we examine the sup norm of the difference between two solutions on consecutive (nested) grids as:

$$G_n := \left\| \mathbf{f}_{2n}^* - \mathbf{f}_n^* \right\|_{n,\infty} = \max_{1 \le j \le n} \left[(\mathbf{f}_{2n}^*)_{2j-1} - (\mathbf{f}_n^*)_j \right]. \tag{5.41}$$

If G_n converges to zero, then the sequence of $\mathbf{f}_n$ behaves somewhat like a Cauchy sequence. Figure 5.14 shows the convergence of G_n, in the case when $F_R(\mathbf{x})$ is a continuous function, for $n = 2^5 - 2^{12}$. The figure shows that as $n \to \infty$, the discrete solution $\mathbf{f}_n^*$ converges to $F_R(\mathbf{x})$ linearly in n.

CHAPTER 6

CONCLUSION AND OUTLOOK

This chapter summarizes the conclusions and results presented throughout the thesis. In addition, we also present some future works that generalize the results from this thesis to other problems of interest.

6.1 Conclusions and Results

Conclusion 1. *(Global minimizers of the Helmholtz free energy functional) In §2, we presented the Helmholtz free energy functional as a continuoum model and showed it arose from a large deviations principle to the Boltzmann distribution. Consequently, global minimizers to the Helmholtz free energy functional characterize the long-time behavior of systems with many particles at zero temperature.*

Conclusion 2. *(Sufficient condition for optimality) In §3, we used a convex relaxation to formulate sufficient conditions for global optimality for the nonconvex Helmholtz energy. The sufficient conditions take the form of an infinite dimensional linear variational problem.*

Conclusion 3. *(Efficient numerical solver) We developed a fast numerical method in §5 to solve the conic programming problem from §3 using a primal-dual interior-point algorithm. The proposed method uses a* MINRES *algorithm, and exploits the Fourier structure of the problem for an efficient matrix-free method.*

Conclusion 4. *(Computational cost of primal-dual interior-point method) We applied a non-common preconditioner in order to alleviate the ill-conditioning that arises in the matrix-free interior-point algorithm. A comparison of the results shows the effectiveness of the proposed preconditioner. On the test problems we examined, the*

total computational cost is estimated to be $\mathcal{O}(n^2 \log n)$*, which is faster than other approaches.*

6.2 Future Work

- More general cases of pair interaction problems contains an external potential that we did not include in our problem, and need to incorporate for more complex models;
- We only studied the Helmholtz energy for zero temperature, however, by adding an entropy term one can study the Helmholtz energy (and Boltzmann distribution) for finite temperatures;
- Extend the developed solver to problems with multiple species, and to higher dimensional geometries (such as two and three dimensions, or molecular configuration spaces).

APPENDIX A

ITERATIVE ALGORITHMS FOR LINEAR SYSTEMS

Consider the following system of linear equations

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \qquad \text{where} \quad \mathbf{A} \in \mathbb{C}^{n\times n}, \text{ and } \mathbf{x}, \mathbf{b} \in \mathbb{C}^{n}. \tag{A.1}$$

Solving the problem (A.1) using noniterative methods (i.e., Gaussian elimination) may require $\mathcal{O}(n^3)$ work, which is computationally expensive as n gets larger. If matrix vector products $\mathbf{A}\mathbf{v}$ can be computed quickly, an attractive alternative is to use an iterative Krylov method to solve (A.1). Table A.1 shows the most common iterative algorithms depending on the corresponding matrix structure. For symmetric positive definite matrices, the conjugate gradient method is usually preferred over MINRES, however, in some cases it may be better to use MINRES [29]. The system (5.21) that we are trying to solve in §5.2.2 is symmetric and not necessarily positive definite, therefore, we implement MINRES in the interior-point method solver.

Table A.1 Different Iterative Algorithms for Solving (A.1) based on Properties of the Matrix A

Iterative algorithm	Properties of $\mathbf{A}$
CG	Symmetric positive definite
MINRES	Symmetric
GMRES	Nonsymmetric

A.1 Convergence and Cost of Conjugate Gradient

Although we use MINRES in our algorithms, we provide here a few details on the rate of convergence (i.e., number of MATVECS) for the similar conjugate gradient (CG) algorithm.

For a symmetric positive definite matrix $\mathbf{A}$, let $\kappa(\mathbf{A}) = \dfrac{\lambda_{max}(\mathbf{A})}{\lambda_{min}(\mathbf{A})}$ be the condition number. Given κ, a well-known upper bound on the error from the conjugate gradient method is:

$$\frac{\|\mathbf{e}_n\|_{\mathbf{A}}}{\|\mathbf{e}_0\|_{\mathbf{A}}} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n \simeq 2\, e^{\dfrac{-2n}{\sqrt{\kappa}}}, \tag{A.2}$$

where $\mathbf{e}_k$ is the error defined as $\mathbf{e}_k := \mathbf{x}_k - \mathbf{x}_{k-1}$, and $\|\mathbf{x}\|_{\mathbf{A}}$ for a positive definite matrix $\mathbf{A}$ is defined as

$$\|\mathbf{x}\|_{\mathbf{A}} = (\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x})^{\frac{1}{2}}.$$

From (A.2), the error (in the weighted $\mathbf{A}$-norm) decays exponentially in the number of iterations. Since $e^{-2} \approx 0.14$, we expect that (roughly) after $\sqrt{\kappa}$ of steps the solution accuracy of $\mathbf{x}$ will improve by one digit. It is also worth noting that although conjugate gradient is used as an iterative method, it is actually an exact method — if $\mathbf{A}$ has n distinct eigenvalues then conjugate gradient and MINRES converge (using exact arithmetic) in at most n steps [56].

Note that at each step of CG, and MINRES we have a matrix vector product which (usually) dominates the computation cost of the algorithm. The time complexity is therefore $\mathcal{O}(m\sqrt{\kappa})$, where m is the cost (i.e., number of flops) of computing a matrix vector product [29, 34, 48, 56].

A.2 The Minimal Residual Algorithms

This section provides the MINRES and PMINRES (preconditioned MINRES) algorithms [2, 17, 18, 33], which are used in §5, and C of this thesis respectively.

In the algorithms below, $\mathbf{P}$ is the preconditioner of $\mathbf{A}$. Since MINRES requires that $\mathbf{A}$ is symmetric, one might expect that a preconditioned MINRES requires the computation of $\mathbf{P}^{\frac{1}{2}}$ to ensure that the precondition matrix $\mathbf{P}^{-\frac{1}{2}}\mathbf{A}\mathbf{P}^{-\frac{1}{2}}$ remains symmetric. The advantage of PMINRES is that it

(i) avoids having to compute $\mathbf{P}^{-\frac{1}{2}}$; and

(ii) only requires the computation of $\mathbf{P}^{-1}$ once every iteration.

Algorithm 2. *(The minimal residual* MINRES *[17])*

given $\mathbf{A}$, $\mathbf{b}$, $\epsilon_{\mathrm{MR}} > 0$

set

$\mathbf{x}_0 = \mathbf{v}_0 = \mathbf{d}_0 = \mathbf{d}_{-1} = \mathbf{0}$,

$\gamma_0 = \gamma_1 = 1, \quad \sigma_0 = \sigma_1 = 0$,

$\beta_1 = \|\mathbf{b}\|_2$.

repeat

$\mathbf{v}_i = (1/\beta_i)\mathbf{v}_i, \quad \alpha_i = \mathbf{v}_i^T \mathbf{A}\mathbf{v}_i, \quad \mathbf{v}_{i+1} = \mathbf{A}\mathbf{v}_i - \alpha_i \mathbf{v}_i - \beta_i \mathbf{v}_{i-1}, \quad \beta_{i+1} = \|\mathbf{v}_{i+1}\|_2,$

$\delta = \gamma_i \alpha_i - \gamma_{i-1}\sigma_i \beta_i, \quad \rho_1 = \sqrt{\delta^2 + \beta_{i+1}^2}, \quad \rho_2 = \sigma_i \alpha_i + \gamma_{i-1}\gamma_i \beta_i, \quad \rho_3 = \sigma_{i-1}\beta_i,$

$\gamma_{i+1} = \delta/\rho_1, \quad \sigma_{i+1} = \beta_{i+1}/\rho_1, \quad \mathbf{d}_i = (\mathbf{v}_i - \rho_3 \mathbf{d}_{i-2} - \rho_2 \mathbf{d}_{i-1})/\rho_1,$

$\mathbf{x}_i = \mathbf{x}_{i-1} + \gamma_{i+1}\, \eta\, \mathbf{d}_i, \quad \|r_i\|_2 = |\sigma_{i+1}| \;\; \|r_{i-1}\|_2, \quad \eta = -\sigma_{i+1}\eta,$

until $\|\mathbf{r}_i\|_\infty < \epsilon_{\mathrm{MR}}$.

Remark 17. *(Matrix-vector product computation,* $\mathbf{A}\mathbf{v}_i$*) Note that in applying Algorithm 2 to the problem in §5, we do not build and store the matrix* $\mathbf{A}$.

Algorithm 3. *(The preconditioned minimal residual* PMINRES *[2, 17, 18, 33])*

given $\mathbf{A}$, $\mathbf{b}$, $\mathbf{P}$, $\epsilon_{\text{MR}} > 0$

set

$\mathbf{z}_0 = \mathbf{0}$, $\mathbf{z}_1 = \mathbf{b}$, $\mathbf{q}_1 = \mathbf{P}^{-1}\mathbf{z}_1$, $\beta_1 = \sqrt{\mathbf{b}^T\mathbf{q}_1}$

$\delta_1^{(1)} = 0$, $\mathbf{x}_0 = \mathbf{d}_0 = \mathbf{d}_{-1} = \mathbf{0}$, $c_0 = -1$, $s_0 = 0$

repeat

$\mathbf{p}_k = \mathbf{A}\mathbf{q}_k$, $\alpha_k = (1/\beta_k^2)\mathbf{q}_k^T\mathbf{p}_k$, $\mathbf{z}_{k+1} = (1/\beta_k)\mathbf{p}_k - (\alpha_k/\beta_k)\mathbf{z}_k - (\beta_k/\beta_{k-1})\mathbf{z}_{k-1}$

$\mathbf{q}_{k+1} = \mathbf{P}^{-1}\mathbf{z}_{k+1}$, $\beta_{k+1} = \sqrt{\mathbf{q}_{k+1}^T\mathbf{z}_{k+1}}$, $\delta_k^{(2)} = c_{k-1}\delta_k^{(1)} + s_{k-1}\alpha_k$

$\gamma_k^{(1)} = s_{k-1}\delta_k^{(1)} - c_{k-1}\alpha_k$, $\epsilon_{k+1}^{(1)} = s_{k-1}\beta_{k+1}$, $\delta_{k+1}^{(1)} = -c_{k-1}\beta_{k+1}$

SymOrtho$(\gamma_k^{(1)}, \beta_{k+1}) \to c_k, s_k, \gamma_k^{(2)}$, $\tau_k = c_k\phi_{k-1}$, $\phi_k = s_k\phi_{k-1}$

if $\gamma_k^{(2)} \neq 0$

$\mathbf{d}_k = (1/\gamma_k^{(2)})\big((1/\beta_k)\mathbf{q}_k - \gamma_k^{(2)}\mathbf{d}_{k-1} - \epsilon_k^{(1)}\mathbf{d}_{k-2}\big)$, $\mathbf{x}_k = \mathbf{x}_{k-1} + \tau_k\mathbf{d}_k$

end

until $\|\mathbf{A}\mathbf{x}_i - \mathbf{b}\|_\infty < \epsilon_{\text{MR}}$.

Remark 18. *(Matrix-vector product computation,* $\mathbf{A}\mathbf{q}_k$ *and* $\mathbf{P}^{-1}\mathbf{z}_{k+1}$*) In applying Algorithm 3 to the problem in §5, we do not build and store the matrix* $\mathbf{A}$*, or the preconditioner* $\mathbf{P}$*.*

Algorithm 4. *(SymOrtho [17])*

given $a,\ b \in \mathbb{R}$

if $b = 0,$

$s = 0, \quad r = |a|, \quad$ ***if*** $\quad a = 0, \quad c = 1, \quad$ ***else*** $\quad c = \operatorname{sign}(a) \quad$ ***end***

elseif $a = 0,$

$c = 0, \quad s = \operatorname{sign}(b), \quad r = |b|$

elseif $|b| > |a|,$

$\tau = a/b, \quad s = \operatorname{sign}(b)/\sqrt{1+\tau^2}, \quad c = s\tau, \quad r = b/s$

elseif $|b| > |a|,$

$\tau = b/a, \quad s = \operatorname{sign}(a)/\sqrt{1+\tau^2}, \quad s = c\tau, \quad r = a/c$

end

APPENDIX B

QUADRATIC PENALTY FUNCTION METHODS

An alternative approach [31] to strongly enforcing the constraint $A\mathbf{z} = \mathbf{b}$, is to weakly enforce the constraint using a quadratic penalty function:

$$\begin{aligned} &\text{minimize} \quad \mu\, \mathbf{u}^{\mathrm{T}}\mathbf{z} + \frac{1}{2}||\mathbf{A}\mathbf{z} - \mathbf{b}||^2 \\ &\text{subject to} \quad \mathbf{z} \geq \mathbf{0}. \end{aligned} \tag{B.1}$$

In (B.1), a quadratic penalty function is introduced to enforce $\mathbf{Az} = \mathbf{b}$ (approximately). The variable $\mu > 0$ is an additional penalty parameter (not to be confused with the primal-dual centering parameter). Applying a primal-dual method to (B.1) requires solving a Newton step at each iteration to obtain the primal and dual increments

$$\begin{pmatrix} \mathbf{A}^{\dagger}\mathbf{A} & -\mathbf{I}_{2n-1} \\ \mathbf{S} & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \mathbf{\Delta z} \\ \mathbf{\Delta s} \end{pmatrix} = \begin{pmatrix} -\mathbf{g}_1 \\ -\mathbf{g}_2 \end{pmatrix}, \tag{B.2}$$

where

$$\begin{aligned} \mathbf{g}_1 &:= \mathbf{A}^{\dagger}\mathbf{A}\mathbf{z} + \mu\mathbf{u} - \mathbf{A}^{T}\mathbf{b} - \mathbf{s}, \\ \mathbf{g}_2 &:= \mathbf{ZS} - t\mathbf{1}. \end{aligned}$$

Here, $\mathbf{Z}$, and $\mathbf{S}$ are diagonal matrices of the vectors $\mathbf{z}$, and $\mathbf{s}$.

After eliminating $\mathbf{\Delta s}$ in (B.2), the linear system for $\mathbf{\Delta z}$ becomes symmetric positive definite, so that conjugate gradient may be use. This is in contrast to the approach in Chapter 5 which resulted in a symmetric but not positive definite matrix and required an alternative to conjugate gradient (i.e., MINRES). Note that the primal-dual algorithm for solving (B.1) has a loop over two parameters (i.e., μ and

t) as opposed to just one for the primal-dual method used in Chapter 5. To avoid the added complications of having two parameters, we prefer to use the primal-dual method in Chapter 5 and handle the equality constraints with Lagrange multipliers (at the expense of implementing MINRES).

APPENDIX C

STUDY OF A NON-DIAGONAL PRECONDITIONER

In this Appendix we introduce another preconditioner, $\mathbf{P_3}$, and compare it with the two other preconditioners mentioned in §5.2.2. Here we solve the equation (5.19) using a preconditioner that incorporates non-diagonal elements of matrix $\mathbf{B}$:

$$\mathbf{P_3} = \begin{pmatrix} \mathbf{I} + \mathbf{\Theta_1} & \mathbf{0} & c\,\mathbf{I} \\ \mathbf{0} & \mathbf{I} + \mathbf{\Theta_2} & \mathbf{0} \\ c\,\mathbf{I} & \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{C.1}$$

In (C.1), c is a positive constant, while $\mathbf{\Theta} \in \mathbb{R}^{(2n-1)\times(2n-1)}$ is defined as before, i.e., $\mathbf{\Theta} := (\mathbf{Z})^{-1}\mathbf{S}$, and contains two submatrices $\mathbf{\Theta}_1 \in \mathbb{R}^{n\times n}$, and $\mathbf{\Theta}_2 \in \mathbb{R}^{(n-1)\times(n-1)}$,

$$\mathbf{\Theta} = \begin{pmatrix} \mathbf{\Theta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Theta}_2 \end{pmatrix}.$$

The inverse of $\mathbf{P_3}$ is:

$$\mathbf{P_3}^{-1} = \begin{pmatrix} \mathbf{D} & \mathbf{0} & -c\,\mathbf{D} \\ \mathbf{0} & (\mathbf{I} + \mathbf{\Theta_2})^{-1} & \mathbf{0} \\ -c\,\mathbf{D} & \mathbf{0} & \mathbf{I} + c^2\,\mathbf{D} \end{pmatrix}, \tag{C.2}$$

where $\mathbf{D} = \big((1-c^2)\mathbf{I} + \mathbf{\Theta}_1\big)^{-1}$.

Note that the preconditioner is not diagonal. To avoid having to compute $\mathbf{P_3}^{-\frac{1}{2}}$ we do not use the method (5.23) for implementation. Instead, we use the PMINRES (preconditioned MINRES) algorithm [2, 17, 18, 33] which only requires application of $\mathbf{P_3}^{-1}$.

Figures C.1 and C.2 show the comparison of using $\mathbf{P_3}$ with different values of c against two other preconditioners for test problems when $F_R(x)$ is a continuous

function, and $F_R(x) = \frac{1}{2}\delta(x) + \frac{1}{2}\delta(x - \frac{1}{2})$ respectively. Based on Figures C.1 and C.2, we can see that the number of MATVECs required by using $\mathbf{P_3}$ is more than $\mathbf{P_1}$ in both cases; and as $c \to 0$ the difference between the number of MATVECs gets smaller, i.e., (# of MATVECs with $\mathbf{P_3}$) $\to$ (# of MATVECs with $\mathbf{P_1}$).

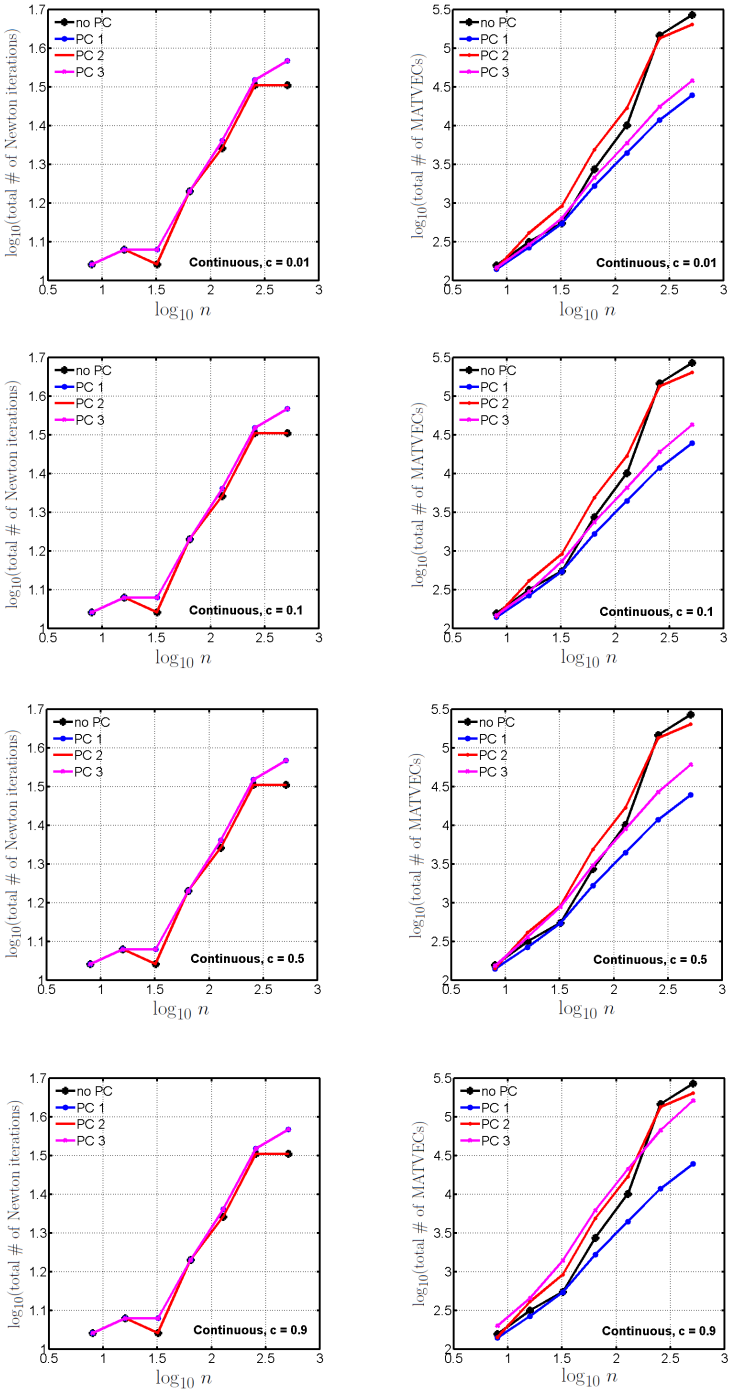

Figure C.1 Number of Newton iterations (left) and MATVECs (right) for different values of problem size n using no preconditioner, $\mathbf{P_1}$, $\mathbf{P_2}$, and $\mathbf{P_3}$. The test is for the interaction potential (3.16) with $(G, L) = (0.9, 1.5)$ and results in a continuous minimizer $F_R(x)$. Tolerance parameters are $\epsilon_{\text{NW}} = \epsilon_{\hat{\eta}} = 10^{-2}$, $\epsilon_{\text{MR}} = 10^{-8}$.

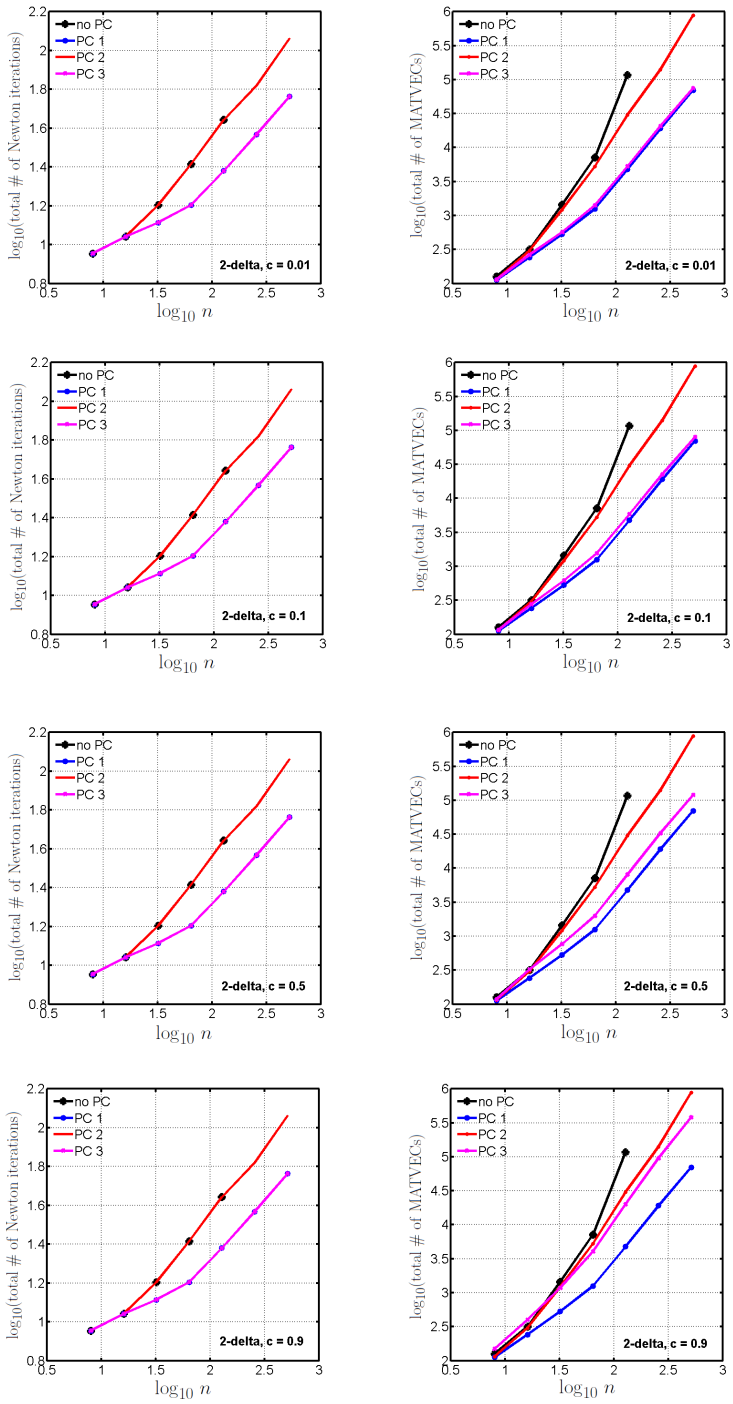

Figure C.2 Number of Newton iterations (left) and MATVECs (right) for different values of problem size n using no preconditioner, $\mathbf{P_1}$, $\mathbf{P_2}$, and $\mathbf{P_3}$. The test is for the interaction potential (3.16) with $(G, L) = (3, 0.2)$ and results in $F_R(x) = \frac{1}{2}\delta(x) + \frac{1}{2}\delta(x - \frac{1}{2})$. Tolerance parameters are $\epsilon_{\text{NW}} = \epsilon_{\hat{\eta}} = 10^{-2}$, $\epsilon_{\text{MR}} = 10^{-8}$.

BIBLIOGRAPHY

[1] P. N. Alcain and C. O. Dorso. The neutrino opacity of neutron rich matter. *Nuclear Physics A*, 961:183–199, 2017.

[2] H. Avron, A. Gupta, and S. Toledo. Solving hermitian positive definite systems using indefinite incomplete factorizations. *Journal of Computational and Applied Mathematics*, 243:126–138, 2013.

[3] M. Bandegi and D. Shirokoff. Approximate global minimizers to pairwise interaction problems via convex relaxation. *Society for Industrial and Applied Mathematics (SIAM) Journal on Applied Dynamical Systems*, 17(1):417–456, 2018.

[4] S. Becker, Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3:165–218, September 2011.

[5] A. J. Bernoff and C. M. Topaz. A primer of swarm equilibria. *SIAM Journal on Applied Dynamical Systems*, 10(1):212–250, 2011.

[6] A. J. Bernoff and C. M. Topaz. Nonlocal aggregation models: A primer of swarm equilibria. *SIAM Review*, 55(4):709–747, 2013.

[7] D. P. Bertsekas. *Convex optimization theory.* Athena Scientific, Belmont, MA, 2009.

[8] D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods.* Academic Press, Cambridge, MA, 2014.

[9] D. P. Bertsekas, A. Nedi, and A. E. Ozdaglar. *Convex analysis and optimization.* Athena Scientific, Belmont, MA, 2003.

[10] S. Bhattacharjee, M. Elimelech, and M. Borkovec. DLVO interaction between colloidal particles: beyond Derjaguins approximation. *Croatica Chemica Acta*, 71(4):883–903, 1998.

[11] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, New York, NY, 2004.

[12] J. A. Cañizo, J. A. Carrillo, and F. S. Patacchini. Existence of compactly supported global minimisers for the interaction energy. *Archive for Rational Mechanics and Analysis*, 217(3):1197–1217, 2015.

[13] M. E. Caplan, A. S. Schneider, C. J. Horowitz, and D. K. Berry. Pasta nucleosynthesis: Molecular dynamics simulations of nuclear statistical equilibrium. *Physical Review C*, 91(6):065802, 2015.

[14] J. A. Carrillo, M. Chipot, and Y. Huang. On global minimizers of repulsive–attractive power-law interaction energies. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2028):20130399, 2014.

[15] J. A. Carrillo, A. Figalli, and F. S. Patacchini. Geometry of minimizers for the interaction energy with mildly repulsive potentials. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 34(5):1299–1308, 2017.

[16] D. Chafaï, N. Gozlan, P. Zitt, et al. First-order global asymptotics for confined particles with singular pair repulsion. *The Annals of Applied Probability*, 24(6):2371–2413, 2014.

[17] S. C. Choi. *Iterative methods for singular linear equations and least-squares problems.* Doctoral Dissertation, Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, 2006.

[18] S. C. T. Choi, C. C. Paige, and M. A. Saunders. MINRES-QLP: A krylov subspace method for indefinite or singular symmetric systems. *SIAM Journal on Scientific Computing*, 33(4):1810–1836, 2011.

[19] R. Choksi, R. C. Fetecau, and I. Topaloglu. On minimizers of interaction functionals with competing attractive and repulsive potentials. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 32(6):1283–1305, 2015.

[20] G. B. Dantzig and M. N. Thapa. *Linear programming 1: introduction.* Springer Science and Business Media, New York, NY, 2006.

[21] G. B. Dantzig and M. N. Thapa. *Linear programming 2: theory and extensions.* Springer Science and Business Media, New York, NY, 2006.

[22] I. Dassios, K. Fountoulakis, and J. Gondzio. A preconditioner for a primal-dual newton conjugate gradient method for compressed sensing problems. *SIAM Journal on Scientific Computing*, 37(6):A2783–A2812, 2015.

[23] B. V. Deraguin and L. Landau. Theory of the stability of strongly charged lyophobic sols and of the adhesion of strongly charged particles in solution of electrolytes. *Acta Physicochim: USSR*, 14:633–662, 1941.

[24] S. Diamond and S. Boyd. Convex optimization with abstract linear operators. *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE) International Conference on Computer Vision*, 2015.

[25] S. Diamond and S. Boyd. Matrix-free convex optimization modeling. *Optimization and Its Applications in Control and Data Sciences: in Honor of Boris T. Polyaks 80th Birthday. Springer Optimization and Its Applications*, 115:221–264, 2016.

[26] S. Diamond and S. Boyd. Stochastic matrix-free equilibration. *Journal of Optimization Theory and Applications*, 172(2):436–454, 2016.

[27] A. Dumbo and O. Zeitouni. *Large deviation techniques and applications*. Springer, New York, NY, 1998.

[28] A. Einstein. Investigations on the theory of the Brownian movement. *Annalen der Physik*, 17:549, 1905.

[29] D. C. L. Fong and M. A. Saunders. CG versus MINRES: An empirical comparison. *Sultan Qaboos University Journal for Science*, 17(1):44–62, 2012.

[30] A. Forsgren, P. E. Gill, and J. R. Shinnerl. Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):187–211, 1996.

[31] K. Fountoulakis, J. Gondzio, and P. Zhlobich. Matrix-free interior point method for compressed sensing problems. *Mathematical Programming Computation*, 6(1):1–31, 2014.

[32] J. Gondzio. Matrix-free interior point method. *Computational Optimization and Applications*, 51(2):457–480, 2012.

[33] R. Herzog and K. M. Soodhalter. A modified implementation of MINRES to monitor residual subvector norms for block systems. *SIAM Journal on Scientific Computing*, 39(6):A2645–A2663, 2017.

[34] M. R. Hestenes and E. Stiefel. *Methods of conjugate gradients for solving linear systems*, volume 49. National Bureau of Standards, Washington, DC, 1952.

[35] C. J. Horowitz, M. A. Perez-Garcia, and J. Piekarewicz. Neutrino-pasta scattering: The opacity of nonuniform neutron-rich matter. *Physical Review C*, 69(4):045804, 2004.

[36] I. Karatzas and S. E. Shreve. Brownian motion. In *Brownian Motion and Stochastic Calculus*, pages 47–127. Springer, New York, NY, 1998.

[37] N. Karmarkar. A new polynomial-time algorithm for linear programming. *Proceedings of the sixteenth annual Association for Computing Machinery (ACM) symposium on theory of computing*, pages 302–311, 1984.

[38] L. G. Khachiyan. A polynomial algorithm in linear programming. *Doklady Academii Nauk SSSR*, 244:1093–1096, 1979.

[39] S. Klaus and K. Ioan. *Lecture Notes on Non-Equilibrium statistical mechanics*. https://www.ks.uiuc.edu/ kosztin/PHYCS498NSM. Cited 07/10/2019.

[40] V. Klee and G. I. Minty. How good is the simplex algorithm? *Inequalities*, III:159–175, 1979.

[41] M. Kocvara, D. Loghin, and J. Turner. Constraint interface preconditioning for topology optimization problems. *SIAM Journal on Scientific Computing*, 38(1):A128–A145, 2016.

[42] M. Kočvara and M. Stingl. On the solution of large-scale SDP problems by the modified barrier method using iterative solvers. *Mathematical Programming*, 120(1):285–287, 2009.

[43] R. Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255, 1966.

[44] A. J. Leverentz, C. M. Topaz, and A. J. Bernoff. Asymptotic dynamics of attractive-repulsive swarms. *SIAM Journal on Applied Dynamical Systems*, 8(3):880–908, 2009.

[45] C. Lin and L. A. Segel. *Mathematics applied to deterministic problems in the natural sciences*, volume 1. SIAM, Philadelphia, PA, 1988.

[46] H. Luo, J. D. Baum, and R. Löhner. An accurate, fast, matrix-free implicit method for computing unsteady flows on unstructured grids. *Computers and Fluids*, 30(2):137–159, 2001.

[47] A. Mogilner, L. Edelstein-Keshet, L. Bent, and A. Spiros. Mutual interactions, potentials, and individual distance in a social aggregation. *Journal of mathematical biology*, 47(4):353–389, 2003.

[48] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.

[49] M. A. Saunders. Management science and engineering 318 (computational and mathematical engineering 338) Large-scale numerical optimization notes 7: PDCO – Primal-dual interior methods, Spring 2013. Standford University, Management Science and Engineering (and Institute for Computational and Mathematical Engineering), web.stanford.edu/group/SOL/software/pdco/pdco.pdf. Cited 07/10/2019.

[50] M. A. Saunders, B. Kim, C. Maes, S. Akle, and M. Zahr. (PDCO: Primal-dual interior method for convex objectives, 2013. web.stanford.edu/group/SOL/software/pdco. Cited 07/10/2019.

[51] A. S. Schneider, C. J. Horowitz, J. Hughto, and D. K. Berry. Nuclear pasta formation. *Physical Review C*, 88(6):065807, 2013.

[52] S. Serfaty. Systems of points with Coulomb interactions. *Proceedings International Congress of Math, Rio de Janeiro*, pages 935–978, 2018.

[53] R. Simione, D. Slepčev, and I. Topaloglu. Existence of ground states of nonlocal-interaction energies. *Journal of Statistical Physics*, 159(4):972–986, 2015.

[54] G. Strang. *Linear Algebra and its Application*. Thomson, Brooks/Cole, Belmont, CA, 2006.

[55] J. K. Strayer. *Linear programming and its applications*. Springer Science and Business Media, New York, NY, 2012.

[56] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*, volume 50. SIAM, Philadelphia, PA, 1997.

[57] E. J. W. Verwey. Theory of the stability of lyophobic colloids. *The Journal of Physical Chemistry*, 51(3):631–636, 1947.

[58] M. H. Wright. The interior-point revolution in optimization: history, recent developments, and lasting consequences. *Bulletin of the American mathematical society*, 42(1):39–56, 2005.

[59] S. Wright. Stability of augmented system factorizations in interior-point methods. *SIAM Journal on Matrix Analysis and Applications*, 18(1):191–222, 1997.

[60] S. J. Wright. Modified cholesky factorizations in interior-point algorithms for linear programming. *SIAM Journal on Optimization*, 9(4):1159–1191, 1999.

[61] W. Zhou, I. G. Akrotirianakis, S. Yektamaram, and J. D. Griffin. A matrix-free line-search algorithm for nonconvex optimization. *Optimization Methods and Software*, 34(1):1–24, 2019.

[62] R. Zwanzig. *Nonequilibrium statistical mechanics*. Oxford University Press, Oxford, UK, 2001.