Title

Contributions to Scientific Computing and Mathematical Modelling: Stochastic Simulation, Constrained Optimization, and Infectious Disease

Author

Landeros, Alfonso

Publication Date

2021

UNIVERSITY OF CALIFORNIA
Los Angeles

Contributions to Scientific Computing and Mathematical Modelling:
Stochastic Simulation, Constrained Optimization, and Infectious Disease

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biomathematics

by

Alfonso Landeros

2021

ABSTRACT OF THE DISSERTATION

Contributions to Scientific Computing and Mathematical Modelling: Stochastic Simulation, Constrained Optimization, and Infectious Disease

by

Alfonso Landeros
Doctor of Philosophy in Biomathematics
University of California, Los Angeles, 2021
Professor Kenneth L. Lange, Chair

The advent of large scale data, particularly from the biological sciences, has accelerated interest in developing computational methods for analysis and prediction. Implementing such methods often requires software to either automate computational tasks or to carry out calculations that elude analytic techniques. This work focuses on the latter while paying respect to useful and elegant abstractions from mathematical theory.

The diverse set of topics in this dissertation span applied probability, mathematical optimization, and epidemiological modelling. First, we investigate simulation techniques for stochastic processes. Second, we elaborate on the proximal distance technique of constrained optimization as a computational framework with examples in orthogonal projection, clustering, regression, and imaging. Third, we use deterministic equation modelling to evaluate school reopening strategies under pandemic conditions. Finally, we conclude with preliminary work on an application of the proximal distance method to hierarchical linear models.

The dissertation of Alfonso Landeros is approved.

Gottfried E. Konecny

Hua Zhou

Janet S. Sinsheimer

Mary S. Sehl

Kenneth L. Lange, Committee Chair

University of California, Los Angeles

2021

A los que tienen que hacerlo con ánimo.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The research presented in this dissertation was funded by a Predoctoral Training Grant from the National Human Genome Research Institute (T32HG002536), financial support from UCLA Biomathematics, and generous funding from Kenneth Lange and Mary Sehl.

First, I would like to thank my primary mentors Kenneth Lange, Mary Sehl, and Janet Sinsheimer for taking a gamble on me. I have been painfully aware of my shortcomings ever since I began my journey in graduate school. Their concern for both my intellectual development and well-being helped me mature into a capable scholar. I could not have gotten to this point without their encouragement. The other two members of my committee were invaluable to my graduate school experience. Both Hua Zhou and Ken Lange are guilty of introducing me to the Julia programming language at a very early stage in my career. Gottfried Konecny has been a wonderful collaborator. Though our joint effort did not bear any fruit, I cherish his willingness to sit down and discuss topics in cancer.

The faculty, staff, and students of the Biomathematics program, past and present, are entitled to my thanks. Though they may not realize it, the camaraderie from my Biomath cohort — Benjamin Chu, Alexander Fisher, and Jason Lin — helped keep me sane through the low points in my career. I would also like to single out Bhaven Mistry, Timothy Stutz, Gabriel Hassler, and Mauricio Cruz Loya for being instigators of bawdy conversation and sharing their wisdom in thought-provoking discourse. I look forward to observing everyone's academic careers flourish for many years to come.

My life was charmed by a number of teachers and mentors well before my time at UCLA. I insist that this document should bear their names even if it is deemed embarrassing or unprofessional. In semi-chronological order, I wish to extend my gratitude:

To Sandra Orozco, who helped me transition into English-only classrooms at age 8.

To Mrs. Finch, who noticed that I could not read her blackboard.

To Ms. Tanaka, who insisted I learn touch typing and enroll in music.

To Eduardo Villanueva and Fantima Luu, for inspiring me to pursue mathematics.

To Andy Castillo, Clarissa Banda, and Lorraine Langevin, who have helped several generations in the Montebello community succeed in higher education.

To Antonio Castro, Michael Lopez, Leonard Narumi, Anne Rardin, and Kurt Richardson, for the lessons in music, volunteerism, and human relationships that have comforted me in adulthood.

Finally, I am forever indebted to my friends and family for their unwavering emotional support over the last 5 years. My sister Laura and brother-in-law Daniel were always kind enough to let me sneak away to their home, far away from Los Angeles, every once in a while. Natalia Pfaff was my fountain of hope through each precarious moment at which I contemplated giving up on graduate school altogether. Most of all I am thankful for my parents, Alfonso and Elsa. They have been extraordinarily patient and unconditional in their love towards me in spite of my inability to communicate what I do and why I enjoy it. I am glad that they lived to see the end of this chapter in my life, so that they may witness both their children thrive.

VITA

2013	B.S. (Mathematics/Applied Science), University of California, Los Angeles. Specialization in Computing.
2014-2016	Programmer/Analyst. UCLA Health.
2017–2019	GATP Trainee, University of California, Los Angeles.

PUBLICATIONS

Sehl ME, Shimada M, Landeros A, Lange K, Wicha MS. (2015) *Modeling of Cancer Stem Cell State Transitions Predicts Therapeutic Response.* PLOS ONE. 10(9):e0135797.

Landeros A, Stutz T, Keys KL, Alekseyenko A, Sinsheimer JS, Lange K, Sehl, ME. (2018) *BioSimulator.jl: Stochastic simulation in Julia.* Computer Methods and Programs in Biomedicine. 167:23–35.

Landeros A, Padilla OHM, Zhou H, Lange K. (2020) *Extensions to the Proximal Distance Method of Constrained Optimization*, Under review.

Landeros A, Ji X, Lange K, Stutz TC, Xu J, Sehl ME, Sinsheimer JS. (2020) *An Examination of School Reopening Strategies During the SARS-CoV-2 Pandemic*, Under review.

Stutz TC, Landeros A, Xu J, Sinsheimer JS, Sehl ME, Lange K. (2021) *Stochastic Simulation Algorithms for Interacting Particle Systems.* PLOS ONE. 16(3):e0247046.

INTRODUCTION

The growing size and complexity of biological data sets defy ordinary human understanding. Computational tools are increasingly employed in order to identify novel patterns in data and make salient predictions. The marriage of mathematical rigor with algorithmic thinking is necessary in implementing software to aid scientific inquiry. Guided by unbridled curiosity, this dissertation investigates a broad and diverse set of topics at the intersection of the life sciences, stochastic and deterministic modelling, and mathematical optimization. The subsequent paragraphs outline the journey undertaken.

In Chapter 1 we present an abbreviated overview of simulation techniques for Markov jump processes. The broad class of algorithms include the celebrated Gillespie algorithm, also known as the direct method, the family of first reaction methods, τ-leaping, and several variants. We also discuss extensions to spatial-stochastic models based on interacting particle systems. The chapter culminates in a brief introduction to a software package, BioSimulator.jl, that was developed as a suite of stochastic simulation algorithms in the Julia programming language. Software is available at `https://github.com/alanderos91/BioSimulator.jl`. The majority of this chapter is an abridgment of an article published in *Computer Methods and Programs in Biomedicine.* The extension to interacting particle systems is a collaborative project led by Timothy Stutz and is included for completeness.

Chapter 2 studies the problem of minimizing a loss model $f(\boldsymbol{x})$ subject to constraints of the form $\boldsymbol{D}\boldsymbol{x} \in S$, where S is a closed set, convex or not, and $\boldsymbol{D}$ is a fusion matrix. To tackle this generic class of problems, we combine the Beltrami-Courant penalty method of optimization with the proximal distance principle. These ingredients are used to derive iterative descent methods that illustrate the utility of the general framework. We adapt the alternating direction method of multipliers and compare performance metrics on extensive numerical tests including problems in metric projection, convex regression, convex clustering, total variation image denoising, and matrix condition numbers. We provide code in the

Julia language to replicate all of our experiments at `https://github.com/alanderos91/ProximalDistanceAlgorithms.jl`. This work is under review.

Chapter 3 is a study of an important public health issue that emerged as the result of the SARS-CoV-2 pandemic in the year 2020. Using simulation studies based on compartmental models of infectious disease, we examine reopening strategies for schools during a pandemic with goals of cataloging qualitative behavior under various scenarios, identifying important transmission modalities, and evaluating policies. This work is under review.

Finally, Chapter 4 presents preliminary results on an application that grew out of the material in Chapter 2. Here we motivate the use of sparsity sets as a data-driven approach to regularization in regression modelling with hierarchical data.

CHAPTER 1

Simulation of Markov jump processes

1.1 Background and Significance

Differential equations are fundamental modelling tools in applied mathematics and systems biology. For example, compartmental models describe the spread of a disease in a population, rate equations summarize reactions in biochemical networks, and master equations enumerate probabilistic flows that contribute to stochastic phenomena. The complex overlapping feedback and feed-forward loops of biological systems often subvert elegant theorems and so one must be content to make progress with approximations and numerical simulations. A class of probabilistic models based on Markov jump processes naturally give rise to systems of differential equations, the Kolmogorov forward and backward equations [Lan10a, Nor98]. The former is sometimes referred to as the *master equation* governing the stochastic behavior of a dynamical system.

A probabilistic description of a biological system is helpful in cases where (a) rare events, such as extinction or mutation, occur naturally and influence system dynamics, (b) population compartments, such as numbers of biochemical molecules, are present in small numbers, and (c) population cycles arise from demographic stochasticity [Lan10a, ELS02, Wil06]. Examples of stochastic models include gene expression networks [MW11], cancer dynamics [SZS11], and demographic and ecological systems [CS12]. However, exact closed-form solutions are available in all but the simplest of models.

This difficulty can be overcome by starting with simpler models. Yet many interesting systems produce a master equation involving either a large number or infinitely-many coupled differential equations. *Generating functions* allow one to make progress provided the state space of the Markovian system is not too complicated. Functional transformations, namely the *Laplace transform*, can sometimes reduce the problem to simpler recurrence relations. A third approach involves *truncating the state space* itself, which simplifies the problem to linear algebra and the matrix exponential [MK06, CSL16]. The error due to truncation must be carefully captured and the truncation scheme itself is not obvious. A fourth approach is simulation of statistically exact realizations of a stochastic process [Gil77]. Simulation algorithms leverage the theory of continuous-time Markov chains to generate accurate time-evolution trajectories, test the sensitivity of models to key parameters, and quantify frequencies of rare events. Computational power, time, and the availability of quality, reliable software packages are the only limiting factors. Indeed, these types of simulations are embarrassingly parallel. Yet stochastic simulation is not a panacea. Models with a high degree of variability require more realizations than can be realized on finite computing machines in a reasonable amount of time. Model inference questions can exceed the computational power of the entire observable universe [KM08, RGP10].

The chapter is organized as follows. Section 2 provides a brief overview of theory and culminates in the issue of intractable master equations. In Section 3 we review various simulation algorithms in the literature which are applied to well-mixed stochastic processes. We then extend the general simulation principle to a special class of spatial systems called *interacting particle systems* in Section 4. Next, Section 5 briefly discusses of BioSimulator.jl, a software implementation of various simulation algorithms. Section 6 illustrates our software with an example from prokaryotic gene regulation. Finally, Section 7 concludes with an overview of potential improvements to our software.

1.2 Markov Jump Processes

Before we discuss simulation specifics, we describe the time evolution of a Markov jump process from the perspective of Markov chains [Nor98, Lan10a]. The underlying Markov chain follows a column vector $\mathbf{X}_t$ whose i-th component X_{ti} is the number of particles of type i at time $t \geq 0$. The components of $\mathbf{X}_t$ track species counts and are necessarily nonnegative integers. The system starts at time $t = 0$ and evolves via a succession of random reactions. Let c denote the number of possible jumps, hereafter referred to as reaction channels, and d the number of particle types. Channel j is characterized by a propensity function $r_j(\mathbf{x})$ depending on the current vector of counts $\mathbf{x}$. In a small time interval of length s, we expect $r_j(\mathbf{x})s + o(s)$ reactions of type j to occur. Reaction j changes the count vector by a fixed integer vector $\mathbf{v}^j$. Some components v_k^j of $\mathbf{v}^j$ may be positive, some 0, and some negative.

The chain waits an exponential length of time in between jumps. If the chain is currently in state $\mathbf{x} \equiv \mathbf{X}_t$, then the intensity of the waiting time until the next reaction is given by $r_0(\mathbf{x}) = \sum_{j=1}^{c} r_j(\mathbf{x})$. Once the decision to jump is made, the chain jumps to the neighboring state $\mathbf{x} + \mathbf{v}^j$ with probability $r_j(\mathbf{x})/r_0(\mathbf{x})$. Table 1.1 lists typical reactions, their propensities $r(\mathbf{x})$, and increment vectors $\mathbf{v}$. In the table, S_i denotes a single particle of type i. Only the nonzero increments v_i are shown. The reaction propensities invoke the law of mass action and depend on rate constants a_i [Hig08]. The vocabulary used to describe particular types of reactions varies across disciplines.

Often one is interested in the finite-time transition probabilities of a Markov chain. Let $p_{\mathbf{x},\mathbf{y}}(t)$ denote the probability that a Markov chain starting at state $\mathbf{x}$ transitions to state $\mathbf{y}$ by time t. The *master equation* reads

$$p_{\mathbf{x},\mathbf{y}}(t + dt) = p_{\mathbf{x},\mathbf{y}}(t) \left[1 - \sum_{j=1}^{c} r_j(\mathbf{y})dt\right] + \sum_{j=1}^{c} p_{\mathbf{x},\mathbf{y}-\mathbf{v}^j}(t)\ r_j(\mathbf{y} - \mathbf{v}^j)dt + o(dt),$$

Name	Reaction	$r(\mathbf{x})$	$\mathbf{v}$
Immigration	$0 \to S_1$	a_1	$v_1 = 1$
Decay	$S_1 \to 0$	$a_2 x_1$	$v_1 = -1$
Dimerization	$S_1 + S_1 \to S_2$	$a_3 \binom{x_1}{2}$	$v_1 = -2, v_2 = 1$
Isomerization	$S_1 \to S_2$	$a_4 x_1$	$v_1 = -1, v_2 = 1$
Dissociation	$S_2 \to S_1 + S_1$	$a_5 x_2$	$v_1 = 2, v_2 = -1$
Budding	$S_1 \to S_1 + S_2$	$a_6 x_1$	$v_2 = 1$
Replacement	$S_1 + S_2 \to S_2 + S_2$	$a_7 x_1 x_2$	$v_1 = -1, v_2 = 1$
Complex Reaction	$S_1 + S_2 \to S_3 + S_4$	$a_8 x_1 x_2$	$v_1 = v_2 = -1, v_3 = v_4 = 1$

Table 1.1: **Propensities $r(\mathbf{x})$ and increment vectors $\mathbf{v}$ for some typical reactions.** Here, S_i denotes a single particle of type i, and a_i denotes the reaction rate constant.

or equivalently, in differential form

$$\frac{d}{dt} p_{\mathbf{x},\mathbf{y}}(t) = \sum_{j=1}^{c} \left[p_{\mathbf{x},\mathbf{y}-\mathbf{v}^j}(t)\ r_j(\mathbf{y}-\mathbf{v}^j) - p_{\mathbf{x},\mathbf{y}}(t)\ r_j(\mathbf{y}) \right].$$

Here positive terms correspond to probabilistic paths along the state space that contribute to the target state $\mathbf{y}$ whereas negative flux represents paths away from this destination. Solving the master equation is an enormous endeavor except in the simplest of models because it is often an infinite system of coupled differential equations. Alternatively, simulating multiple realizations of the process yields data that can then be used to estimate transition probabilities and summary statistics.

1.3 Kinetic Monte Carlo Methods

This section provides a review of various simulation techniques. The list is not exhaustive. We provide descriptions to develop an intuition for the general simulation approach, the mathematics involved, and interesting algorithmic considerations. A recent textbook provides a comprehensive review of different methods [MPT17].

1.3.1 The Family of Direct Methods

The original direct method is a statistically exact algorithm that operates on the wait and jump mechanism for simulating a continuous-time Markov chain [Gil76]. At each step, the algorithm evaluates the propensities $r_j(\mathbf{x})$ of each reaction channel and generates two random deviates. One of these is an exponential deviate indicating the time to the next reaction based on $r_0(\mathbf{x})$. The second is a uniform random number $U(0,1)$ determining which reaction fires next based on the ratios $r_j(\mathbf{x})/r_0(\mathbf{x})$. This general notion of simulating a Markov jump process is quite similar to solving a differential equation via Euler's method. In contrast to the deterministic setting, a Markov jump process has random jump times (step size) and a random, discrete jumps (instantaneous rate of change). Indeed, the two main computational steps in the direct method are:

1. Generate a *random deviate* W representing the time to the next event by sampling from an exponential distribution with rate $r_0(\mathbf{x})$. This provides the update $t \mapsto t + W$.
2. Generate a *random index* J denoting the reaction that occurred by sampling from a set of categories with probabilities $r_j(\mathbf{x})/r_0(\mathbf{x})$ for each reaction indexed by $j = 1, 2, \ldots, c$. This provides the update $\mathbf{x} \mapsto \mathbf{x} + \mathbf{v}^J$.

There are various techniques to simulate these random numbers [MT00]. However, some propensities $r_j(\mathbf{x})$ change after each event, so the distributions underlying each step will necessarily change. An interesting feature of the direct method and its variants is that the two steps are coupled only by the propensities r_j and the cumulative intensity r_0. Thus, the *family of direct methods* emphasize (i) minimizing the number of times propensities are evaluated and (ii) searching efficiently on the reaction rates to sample the index J. *Dependency graphs* are key data structures that address (i) by describing how the firing of reaction j affects the propensity of all other reactions k [MPT17, GB00, TZP15]. Search methods address

(ii) by bisection techniques and iteratively sorting reaction channels according to their firing frequency [MPC06].

1.3.2 The First Reaction Method and its Derivatives

Gillespie proposed the *first reaction method* as an alternative to the direct method [Gil77]. The main difference is the time to the next reaction

$$W = \min_{1 \leq j \leq c} \{W_1, \ldots, W_c\}$$

defined by independent exponentially distributed waiting times $W_1, \ldots, W_c$ each with rates $r_1, \ldots, r_c$, respectively. Here c again denotes the total number of reaction channels. The premise of the algorithm is to compute the minimum of c exponential random variables explicitly. This approach must generate c exponential deviates per iteration compared to the 2 random numbers required for the direct method. However, it provides an alternative way of thinking about simulation by closely coupling the random time to the next reaction and the reaction channel itself.

The *next reaction method*, also known as the *Gibson-Bruck method*, improves on the poor efficiency of the first reaction approach [GB00]. At time $t = 0$, the algorithm seeds each reaction channel j a firing time τ_j and stores them inside a *priority queue*. In this context, a priority queue is a data structure that sorts pairs (j, W_j) according to the value of W_j in increasing order. That is, if W_J is the minimum time, then the pair (J, W_J) appears at the top of the queue. Thus, the next reaction is J and its firing time is τ_J; all other reactions fire at some future time. After reaction J fires, the next reaction method updates the state vector $\mathbf{x} \to \mathbf{x} + \mathbf{v}^J$. The next firing time W_J is also updated by an appropriate exponential deviate:

$$W_{J,\text{new}} = W_{J,\text{old}} + \text{Exponential}(r_{J,\text{new}}(\mathbf{x})),$$

where $r_{J,\text{new}}(\mathbf{x})$ is the new propensity value. The remaining firing times change according to the recipe

$$W_{j,\text{new}} = t + \frac{r_{j,\text{old}}}{r_{j,\text{new}}}(W_{j,\text{old}} - t), \quad j \neq J,$$

based on the lack of memory property of the exponential distribution. The next reaction method relies on dependency graphs to evaluate reaction propensities that are then used to update firing times. Indeed, the technique is an example of an algorithm that is sensitive to data structures given how poorly its performance is characterized relative to theoretical guarantees [CLP04]. Recent work has brought this issue to light and shown that one can achieve a constant-time complexity by using appropriate implementations of data structures [SO15].

1.3.3 τ-leaping Methods

Algorithms that simulate every reaction ultimately succumb to the high computational expense of large models. The τ-leaping framework attempts to accelerate simulation by lumping reaction events together within a time leap τ, selected to be as large as possible [Gil01, GP03, CGP06]. The basic τ-leaping formula is

$$\mathbf{X}_{t+\tau} = \mathbf{X}_t + \sum_{j=1}^{c} \mathbf{v}^j Y_j.$$

where Y_j is a Poisson random variable with rate $r_j(\mathbf{X}_t)\tau$. The main challenge in τ-leaping is selecting the step size as large as possible while satisfying a *leap condition* that provides some guarantee that reaction propensity functions do not change significantly over the leap interval. A commonly used heuristic is

$$|r_j(\mathbf{X}_{t+\tau}) - r_j(\mathbf{X}_t)| \leq \epsilon, \quad j = 1, 2, \ldots, c,$$

which states that the propensity for each reaction j is approximately constant over a leap of size τ. Here, $\epsilon \in (0, 1]$ is a prescribed acceptable change in propensities that controls the accuracy of sample paths generated by a τ-leaping algorithm. A larger ϵ allows for larger leaps, while a smaller ϵ restricts leap size.

As with differential equations, a stochastic system is said to be *stiff* if the dynamics force a simulation procedure to take "small" steps. Stiffness arises for a variety of reasons such as the presence of separate time scales that split the system between "fast" and "slow" reactions, with the former occurring in a nearly deterministic fashion. In any case, this issue causes the number of simulated events to increase in exact methods. Stiffness poses a second threat to τ-leaping methods: in addition to decreasing the leap size, it causes the leaping procedure to generate an excess of events due to the unbounded nature of the Poisson distribution. There are precautionary measures to protect against aberrant behavior in selecting τ [CGP05, SAL09].

As with exact simulation methods, there are a variety of *tau*-leaping techniques. The Step Anticipation τ-Leaping (SAL) algorithm subsumes the original Poisson-based τ-leaping procedure [SAL09]. In the SAL algorithm, one approximates each propensity by a first-order Taylor polynomial around t with starting value $r_j(\mathbf{x})$. The number of reactions of type j is then sampled from a Poisson distribution with mean

$$\omega_j(t, t+\tau) = \int_0^t \left[r_j(\mathbf{x}) + \frac{d}{dt} r_j(\mathbf{x}) s \right] \mathrm{d}s = r_j(\mathbf{X})\tau + \frac{d}{dt} r_j(\mathbf{x}) \frac{1}{2} \tau^2.$$

One can use a higher-order approximation at the cost of having to contend with time derivatives of a stochastic process.

1.4 Interacting Particle Systems

The methods of the previous section were developed for spatially homogeneous stochastic processes and emphasized state spaces involving population counts. Many interesting biological phenomena involve interactions between subsystems separated by distance and barriers, such as pattern formation during embryonic development and island diversification. The theory of Markov chains can be applied to more abstract state spaces to model such phenomena. In particular, interacting particle systems (IPS) consider a stochastic process on a *configuration space* Ω determined by a *local state space* S and a *site space* typically represented by a countable graph G [Lig05]. The local state space describes attributes of interacting particles or agents that take up space in G. For example, in the *voter model* one considers the range of voter attitudes S on a topic, say liberal or conservative fiscal policy, which may be continuous or discrete [Lig99]. Each voter's home lies at some site of the site space, $G = \mathbb{Z}^2$, a two-dimensional discrete grid. Each voter randomly reconsiders his or her views due to the influence of lively debates with neighbors. In this model both the spatial distribution of voters communities and their local, internal dynamics play a role in emergent properties. The configuration space is therefore $\Omega = S^{\mathbb{Z}^2}$ as it describes all possible ways of arranging voters on grid G while accounting for each possible view enumerated in S. There is a sizeable literature describing the dynamics and mathematical properties of different IPSs, but it is clear that configuration spaces quickly become unwieldy if the local state space considers many attributes.

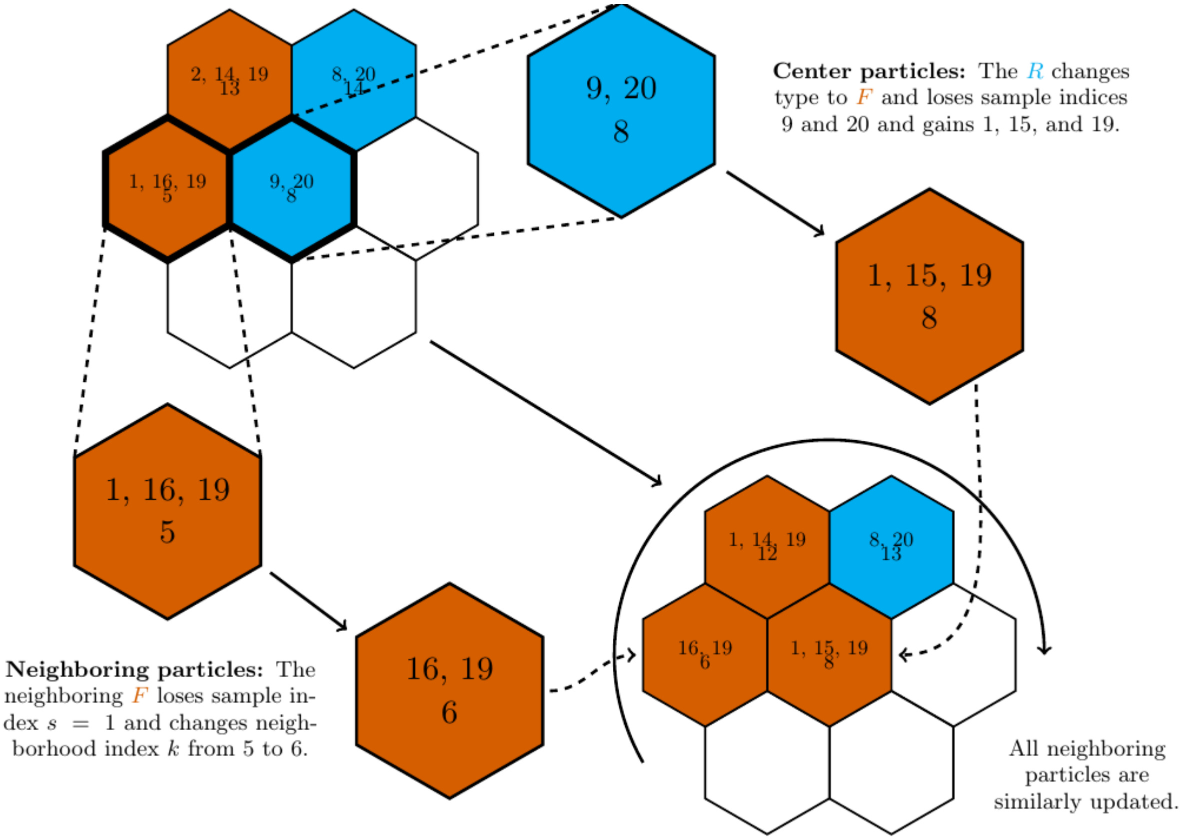
The most straightforward method of simulating IPSs assigns a reaction set to each site on the lattice. At each step of the algorithm, one evaluates reaction propensities at each site which are then used to determine the next event and its firing time in similar fashion as the Direct method. This approach is prohibitively expensive as system size increases because the sampling steps scales with the number of sites. If we limit our attention to pairwise interactions based on discrete nearest-neighbor interactions, then we can improve simulation by defining equivalence classes on configurations. For example, *Von Neumann neighborhoods* on a 2-D grid involve only interactions with neighbors east, west, north, or south of a site and have a *neighborhood capacity* $C = 4$. In general, the neighborhood capacity will depend on the dimension of the lattice d and its topology. If an IPS has L types of particles and a neighborhood capacity C, then one can simply count the number of particles of each type within a given neighborhood to form an equivalence class. Including open sites as phantom particles, the total space of configurations reduces to equivalence classes of the form $(n_1, n_2 \ldots, n_{L+1})$ with $\sum_{k=1}^{L+1} n_k = C$. The exponents of each term in the multinomial $(x_1 + \ldots x_{L+1})^C$ encode all possible neighborhood classes. The sum of their coefficients count the total number of neighborhood classes:

$$K = \sum_{n_1+\ldots+n_{L+1}=C} \binom{C}{n_1, \ldots, n_{L+1}}.$$

Evaluating the propensity function of each reaction channel requires us to track the number of particles of each type that fall within a specific neighborhood class. This makes possible defining reaction channels on the *neighborhood class* rather than specific sites, effectively reducing the number of reaction channels. This approach is sometimes referred to as the n-fold method for Kinetic Monte Carlo simulations [BKL75], and is analogous to network-free simulation techniques [SFE11, GM18, SML19]. The approach depicted in Figure 1.1 based on the model in Table 1.2 further reduces the number of virtual reaction channels organizing particles into *sample classes* determined by neighborhood configurations and interacting particle types.

Name	Diagram	Type
Predation	$F + R \to F + F$	Pairwise
Reproduction	$R + \emptyset \to R + R$	Pairwise
Migration	$F + \emptyset \to \emptyset + F$	Pairwise
	$R + \emptyset \to \emptyset + R$	Pairwise
Death	$F \to \emptyset$	On-site
	$R \to \emptyset$	On-site

Table 1.2: **A predator-prey model as an IPS**. Foxes (F) and rabbits (R) interact on a 2-D hexagonal lattice with open sites ($\emptyset$). Reactions are either on-site, if involving a single animal interacting only with itself, or pairwise, if involving an animal interacting with an adjacent site.



Center particles: The R changes type to F and loses sample indices 9 and 20 and gains 1, 15, and 19.

Neighboring particles: The neighboring F loses sample index $s = 1$ and changes neighborhood index k from 5 to 6.

All neighboring particles are similarly updated.

Figure 1.1: **Schematic of neighborhood updates.** Each animal is assigned a unique *neighborhood index* k (bottom number) based on neighborhoods of the form (O, F, R) counting the number of adjacent open sites, foxes, and rabbits, respectively. The set of possible interactions (Table 1.2) and the neighborhood index determine *sampling classes* s (top numbers). Once a simulation algorithm samples a particular reaction channel operating on particular sampling class s, it is easy to sample the appropriate "center" particle which is guaranteed to have the correct interacting neighboring particle. The number of sample classes reflects the number of possible interactions that an animal can participate in. In this example a fox eats a nearby rabbit. It can no longer hunt after this event and therefore loses a sample class.

1.5 Software Implementation

BioSimulator.jl is a user-friendly stochastic simulation tools that implements the methods described in Section 1.3 [LSK18]. It is written in Julia, a programming language targeted at scientific computing. The software package is open-source and available at `https://www.github.com/alanderos91/BioSimulator.jl.git`.

In the interest of space, we omit a detailed description of the software and present performance benchmarks against three software packages: StochPy, StochKit2, and Gillespie.jl (Table 1.3). Each software package is selected for its similarity to BioSimulator.jl. Specifically, these tools are domain-independent, general purpose simulation engines for Markov jump processes involving population counts. Descriptions of the benchmark methodology, the models used, and software packages is available in our published manuscript [LSK18].

Software	StochPy	StochKit2	Gillespie.jl	BioSimulator.jl	
Method	Direct	Direct (dep. graph)	Direct	Direct	
Mode	Serial	Parallel	Serial	Serial	Parallel
Kendall's process		174.31 (173.81 – 174.95)		1.04 (1.03 – 1.05)	0.56 (0.48 – 0.66)
	257 (223 – 296)		1285 (1280 – 1293)	1.38 (1.37 – 1.40)	0.79 (0.68 – 0.93)
Michaelis-Menten		210.42 (209.89 – 211.14)		0.783 (0.788 – 0.808)	0.45 (0.44 – 0.46)
	214 (213 – 216)		1.15 (1.12 – 1.21)	0.918 (0.909 – 0.936)	0.47 (0.46 – 0.48)
Auto-regulation		432 (425 – 445)		426 (424 – 432)	135 (124 – 149)
	8.23 (7.87 – 8.75); $\times 10^4$		824 (762 – 834)	1809 (1734 – 1860)	211 (193 – 240)
Dimer-decay		216.98 (216.52 – 217.68)		1.76 (1.75 – 1.79)	0.81 (0.80 – 0.82)
	483 (480 – 488)		2.68 (2.66 – 2.84)	2.39 (2.37 – 2.42)	1.03 (1.02 – 1.05)
Yeast		260 (254 – 267)		2.56 (2.53 – 2.63)	1.38 (1.10 – 1.42)
	469 (465 – 474)		2.25 (2.21 – 2.63)	2.92 (2.89 – 2.99)	1.24 (1.20 – 1.27)

Table 1.3: **Median runtimes and interquartile ranges for StochPy, StochKit2, Gillespie.jl, and BioSimulator.jl.** Here each model is simulated 1000 times and timings are reported in milliseconds (ms). Each sample measured the time to generate 10 realizations of a given stochastic process. Results using fixed-interval and fixed output options are recorded in the first and second rows of each cell, respectively. Each simulation tool is used with its default settings. For example, StochKit2 automatically parallelizes simulation tasks involving multiple realizations and uses a dependency graph by default. We note that both StochKit2 and BioSimulator.jl used 8 threads for the parallel simulation benchmarks. Direct comparisons based on these results are not possible, in the sense that slower performance does not necessarily indicate a particular tool is poorly implemented. Rather, our results reflect natural trade-offs in optimizing software for particular goals. **Note**: The StochPy benchmark on the *auto-regulation* model is based on only 100 samples.

1.6 Example: An Auto-Regulatory System

The influence of noise at the cellular level is difficult to capture in deterministic models. Stochastic simulation is appropriate for the study of regulatory mechanisms in genetics, where key species may be present in low numbers.

There are eight possible reactions: (1) gene transcription into RNA, (2) translation of the protein, (3) dimerization of the protein with itself, (4) dissociation of the protein dimer, (5) binding to the gene, (6) unbinding from the gene, (7) RNA degradation, and (8) protein degradation. There are five species to track — the free copies of the gene, transcribed RNA, protein molecules, dimer molecules, and blocked copies of the gene. The model is easily implemented in BioSimulator.jl:

```
using BioSimulator

# Initialization.
model <= Network("negative auto-regulation")

# Add species to model.
model <= Species("gene", 10) # assume 10 copies of the gene are present
model <= Species("RNA", 0)   # transcribed from the underlying gene
model <= Species("P", 0)     # protein
model <= Species("P2", 0)    # protein dimer
model <= Species("P2_gene")  # gene repression

# Add reactions to model.
model <= Reaction("transcription",              0.01, "gene --> gene + RNA")
model <= Reaction("translation",                10.0, "RNA --> RNA + P")
model <= Reaction("dimerization",               1.0,  "P + P --> P2")
model <= Reaction("dissociation",               1.0,  "P2 --> P + P")
model <= Reaction("repression binding",         1.0,  "gene + P2 --> P2_gene")
model <= Reaction("reverse repression binding", 10.0, "P2_gene --> gene + P2")
model <= Reaction("RNA degradation",            0.1,  "RNA --> 0")
model <= Reaction("protein degradation",        0.01, "P --> 0")

# Run the simulation.
result = simulate(model, StepAnticipation(), Val(:full), time = 500.0, trials = 100)
```

RNA typically has a limited lifetime. Thus, the per particle reaction rates governing protein production are balanced to favor translation events following transcription. Moreover, the reaction rates for dimerization and dissociation reflect an assumption that the protein favors neither the monomer nor the dimer configuration. Figure 1.2 illustrates this point.

1.7 Future Directions

The work presented here suggests a unified framework for simulating Markovian processes taking values in countable state spaces. By defining the rules of a system on equivalence classes we separate the logic of sampling jumps from the bookkeeping details of updating the system state. This makes it possible to re-use algorithms developed for well-mixed processes in simulating interacting particle systems. The basic Direct and First Reaction methods in BioSimulator.jl already support IPS simulations. An important goal of this proposal is to extend the notion of a reaction dependency graph to IPSs. This will allow us to simulate spatial stochastic processes with relatively efficient algorithms such as the *sorting direct method* [MPC06] and *rejection stochastic simulation algorithm* [TZP15]. To our knowledge, no software package provides this level of generality. More importantly, the performance implications on spatial simulations is not characterized in the literature.

Our emphasis on generic algorithms and code modularity should attract methods developers. These two features will be in important in exploring techniques to accelerate simulation in the context of inference problems. For example, *importance sampling* is often used in statistics to decrease the variance of estimators at the cost of biasing estimates. When the bias is known, one can adjust estimates while simultaneously reducing the number of samples needed for convergence. This technique has been applied to kinetic Monte Carlo simulations, but it is not fully appreciated in the systems biology community [KM08, RGP10].

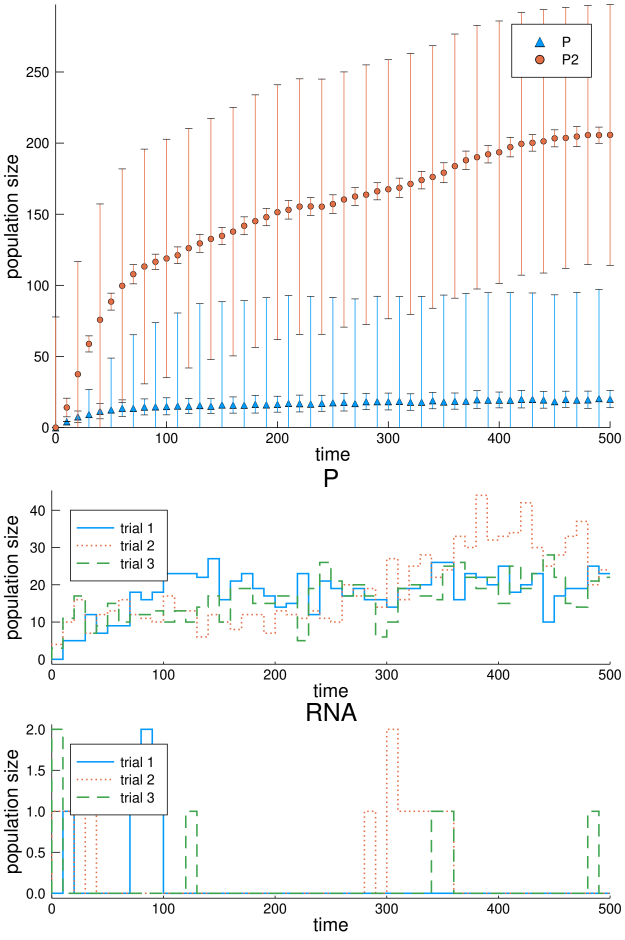


Figure 1.2: **Sample output from BioSimulator.jl.** (a) Mean trajectories for the protein and dimer in a simple auto-regulatory gene network. The bars represent 1 standard deviation away from the mean. (b) Full sample paths for the protein (top) and RNA (bottom). Note that the RNA sample path is sensitive to the number of epochs if one were using the fixed-interval option. Using a large window may fail to capture peaks in the output even though the process is simulated exactly. This subtlety is important for qualitative model assessment.

CHAPTER 2

Extensions of the Proximal Distance Method

2.1 Introduction

The generic problem of minimizing a continuous function $f(\boldsymbol{x})$ over a closed set S of $\mathbb{R}^p$ can be attacked by a combination of the penalty method and distance majorization. The classical penalty method seeks the solution of a penalized version $h_\rho(\boldsymbol{x}) = f(\boldsymbol{x}) + \rho q(\boldsymbol{x})$ of $f(\boldsymbol{x})$, where the penalty $q(\boldsymbol{x})$ is nonnegative and 0 precisely when $\boldsymbol{x} \in S$. If one follows the solution vector $\boldsymbol{x}_\rho$ as ρ tends to ∞, then in the limit one recovers the constrained solution [Bel70, Cou43]. The function

$$q(\boldsymbol{x}) \quad = \quad \frac{1}{2}\operatorname{dist}(\boldsymbol{x}, S)^2 \quad = \quad \frac{1}{2}\min_{\boldsymbol{y} \in S} \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

is one of the most fruitful penalties in this setting. Our previous research for solving this penalized minimization problem has focused on an MM (majorization-minimization) algorithm based on distance majorization [CZL14, KZL19]. In distance majorization one constructs the surrogate function

$$g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n) \quad = \quad f(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{x} - \mathcal{P}(\boldsymbol{x}_n)\|^2$$

using the Euclidean projection $\mathcal{P}(\boldsymbol{x}_n)$ of the current iterate $\boldsymbol{x}_n$ onto S. The minimum of the surrogate occurs at the proximal point

$$\boldsymbol{x}_{n+1} \quad = \quad \operatorname{prox}_{\rho^{-1} f}[\mathcal{P}(\boldsymbol{x}_n)].$$

According to the MM principle, this choice of $\boldsymbol{x}_{n+1}$ decreases $g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n)$ and hence the objective $h_\rho(\boldsymbol{x})$ as well.

We have named this iterative scheme the proximal distance algorithm [KZL19, Lan16]. It enjoys several virtues. First, it allows one to exploit the extensive body of results on proximal maps and projections. Second, it does not demand that the constraint set S be convex. If S is merely closed, then the map $\mathcal{P}(\boldsymbol{x})$ may be multivalued, and one must choose a representative element from the projection $\mathcal{P}(\boldsymbol{x}_n)$. Third, the algorithm does not require the objective function $f(\boldsymbol{x})$ to be differentiable. Fourth, the algorithm dispenses with the chore of choosing a step length. Fifth, if sparsity is desirable, then the sparsity level can be directly specified rather than implicitly determined by the tuning parameter of the lasso or other penalty.

Traditional penalty methods have been criticized for their numerical instability. This hazard is mitigated in the proximal distance algorithm by its reliance on proximal maps, which usually are highly accurate. The major defect of the proximal distance algorithm is slow convergence. This can be ameliorated by Nesterov acceleration [Nes13]. There is also the question of how fast one should send ρ to ∞. Although no optimal schedule is known, simple numerical experiments usually yield a good choice. Finally, soft constraints can be achieved by stopping the steady increase of ρ at a finite value.

2.2 Extensions

This simple version of distance majorization can be generalized in various ways. For instance, it can be expanded to multiple constraint sets. In practice, at most two constraint sets usually suffice. Another generalization is to replace the constraint $\boldsymbol{x} \in S$ by the constraint $\boldsymbol{D}\boldsymbol{x} \in S$, where $\boldsymbol{D}$ is a compatible matrix. Again, the original case $\boldsymbol{D} = \boldsymbol{I}$ is allowed. By analogy with the fused lasso of [TSR05a], we will call the matrix $\boldsymbol{D}$ a fusion matrix. This paper is devoted to the study of the general problem of minimizing a differentiable function $f(\boldsymbol{x})$ subject to r fused constraints $\boldsymbol{D}_i\boldsymbol{x} \in S_i$. We will approach this problem by extending the proximal distance method. For a fixed penalty constant ρ, the objective function and its

MM surrogate now become

$$
\begin{aligned}
h_\rho(\boldsymbol{x}) &= f(\boldsymbol{x}) + \frac{\rho}{2} \sum_{i=1}^{r} \operatorname{dist}(\boldsymbol{D}_i \boldsymbol{x}, S_i)^2 \\
g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n) &= f(\boldsymbol{x}) + \frac{\rho}{2} \sum_{i=1}^{r} \|\boldsymbol{D}_i \boldsymbol{x} - \mathcal{P}_i(\boldsymbol{D}_i \boldsymbol{x}_n)\|^2,
\end{aligned}
$$

where $\mathcal{P}_i(\boldsymbol{y})$ denotes the projection of $\boldsymbol{y}$ onto S_i. Any or all of the fusion matrices $\boldsymbol{D}_i$ can be the identity $\boldsymbol{I}$.

Fortunately, we can simplify the problem by defining S to be the Cartesian product $\prod_{i=1}^{r} S_i$ and $\boldsymbol{D}$ to be the stacked matrix

$$
\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_1 \\ \vdots \\ \boldsymbol{D}_r \end{pmatrix}.
$$

Our objective and surrogate then revert to the less complicated forms

$$
h_\rho(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{\rho}{2} \operatorname{dist}(\boldsymbol{D}\boldsymbol{x}, S)^2 \tag{2.1}
$$

$$
g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n) = f(\boldsymbol{x}) + \frac{\rho}{2} \|\boldsymbol{D}\boldsymbol{x} - \mathcal{P}(\boldsymbol{D}\boldsymbol{x}_n)\|^2, \tag{2.2}
$$

where $\mathcal{P}(\boldsymbol{x})$ is the Cartesian product of the projections $\mathcal{P}_i(\boldsymbol{x})$. Note that all closed sets S_i with simple projections, including sparsity sets, are fair game.

2.2.1 Newton's Method and Least Squares

Unfortunately, the proximal operator $\operatorname{prox}_{\rho^{-1} f}(\boldsymbol{y})$ is no longer relevant in calculating the MM update $\boldsymbol{x}_{n+1}$. When $f(\boldsymbol{x})$ is smooth, Newton's method applied to the surrogate $g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n)$ gives the update

$$
\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \Big[\boldsymbol{H}_n + \rho \boldsymbol{D}^t \boldsymbol{D}\Big]^{-1} \Big\{\nabla f(\boldsymbol{x}_n) + \rho \boldsymbol{D}^t [\boldsymbol{D}\boldsymbol{x}_n - \mathcal{P}(\boldsymbol{D}\boldsymbol{x}_n)]\Big\},
$$

where $\boldsymbol{H}_n = d^2 f(\boldsymbol{x}_n)$ or is an approximation to it. To enforce the descent property, it is often prudent to substitute a positive definite approximation $\boldsymbol{H}_n$ for $d^2 f(\boldsymbol{x}_n)$. In statistical applications, the expected information matrix is a natural substitute. It is also crucial to retain as much curvature information on $f(\boldsymbol{x})$ as possible. Newton's method has two drawbacks. First, it is necessary to compute and store $d^2 f(\boldsymbol{x}_n)$. This is mitigated in statistical applications by the substitution just mentioned. Second, there is the necessity of solving a large linear system. Fortunately, the matrix $\boldsymbol{H}_n + \rho \boldsymbol{D}^t \boldsymbol{D}$ is often well-conditioned, for example, when $\boldsymbol{D}$ has full column rank. The method of conjugate gradients can be employed to solve the linear system under ideal conditions.

To reduce the condition number of the matrix $\boldsymbol{H}_n + \rho \boldsymbol{D}^t \boldsymbol{D}$ even further, one can sometimes rephrase the Newton step as iteratively reweighted least squares. For instance, in a generalized linear model, the gradient $\nabla f(\boldsymbol{x})$ and the expected information $\boldsymbol{H}$ can be written as

$$\nabla f(\boldsymbol{x}) \;=\; -\boldsymbol{Z}^t \boldsymbol{W}^{1/2} \boldsymbol{r} \quad \text{and} \quad \boldsymbol{H} \;=\; \boldsymbol{Z}^t \boldsymbol{W} \boldsymbol{Z},$$

where $\boldsymbol{r}$ is a vector of standardized residuals, $\boldsymbol{Z}$ is a design matrix, and $\boldsymbol{W}$ is a diagonal matrix of case weights [Lan10b, NW72]. The Newton step is now equivalent to minimizing the least squares criterion

$$\begin{aligned} \frac{1}{2}\boldsymbol{x}^* J \boldsymbol{x} - \nabla f(\boldsymbol{x})^* \boldsymbol{x} &= \|\boldsymbol{W}^{1/2} \boldsymbol{Z} \boldsymbol{x} - \boldsymbol{W}^{-1/2} \nabla f(\boldsymbol{x})\|^2 \\ &= \left\| \begin{pmatrix} \boldsymbol{W}_n^{1/2} \boldsymbol{Z} \\ \sqrt{\rho} \boldsymbol{D} \end{pmatrix} \boldsymbol{x} - \begin{pmatrix} \boldsymbol{W}_n^{1/2} \boldsymbol{Z} \boldsymbol{x}_n + \boldsymbol{r}_n \\ \sqrt{\rho} \mathcal{P}(\boldsymbol{D} \boldsymbol{x}_n) \end{pmatrix} \right\|^2 . \end{aligned}$$

In this context a version of the conjugate gradient algorithm adapted to least squares is attractive [PS82]. The algorithm LSQR [PS82] and its sibling LSMR perform well when ill conditioning is an issue [FS11].

2.2.2 Proximal Distance by Steepest Descent

In high-dimensional optimization problems, gradient descent is typically employed to avoid matrix inversion. Determination of an appropriate step length is now a primary concern. In the presence of fusion constraints $\boldsymbol{D}\boldsymbol{x} \in S$ and a convex quadratic loss $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^t\boldsymbol{x}$, the gradient of the proximal distance objective at $\boldsymbol{x}_n$ amounts to

$$\boldsymbol{v}_n \quad = \quad \boldsymbol{A}_n\boldsymbol{x} + \boldsymbol{b} + \rho\boldsymbol{D}^t[\boldsymbol{D}\boldsymbol{x}_n - \mathcal{P}(\boldsymbol{D}\boldsymbol{x}_n)].$$

For the steepest descent update $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - t_n\boldsymbol{v}_n$, one can show that the optimal step length is

$$t_n \quad = \quad \frac{\|\boldsymbol{v}_n\|^2}{\boldsymbol{v}_n^t\boldsymbol{A}\boldsymbol{v}_n + \rho\|\boldsymbol{D}\boldsymbol{v}_n\|^2}.$$

This update obeys the descent property and avoids matrix inversion. Once again one can substitute a local convex quadratic approximation around $\boldsymbol{x}_n$ for $f(\boldsymbol{x})$. If the approximation majorizes $f(\boldsymbol{x})$, then the descent property is preserved. In the failure of majorization, the safeguard of step halving is trivial to implement.

In addition to Nesterov acceleration, gradient descent can be accelerated by the subspace MM technique [CIM10]. Let $\boldsymbol{G}_n$ be the matrix with k columns determined by the k most current gradients of the objective $h_\rho(\boldsymbol{x})$, including $\nabla h_\rho(\boldsymbol{x}_n)$. Generalizing our previous assumption, suppose $f(\boldsymbol{x})$ has a quadratic surrogate with constant Hessian $\boldsymbol{H}_n$ at $\boldsymbol{x}_n$. Overall we get the quadratic surrogate

$$\begin{aligned} q_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n) \quad &= \quad g_\rho(\boldsymbol{x}_n \mid \boldsymbol{x}_n) + \nabla g_\rho(\boldsymbol{x}_n \mid \boldsymbol{x}_n)^t(\boldsymbol{x} - \boldsymbol{x}_n) \\ &\quad\quad + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_n)^t(\boldsymbol{H}_n + \rho\boldsymbol{D}^t\boldsymbol{D})(\boldsymbol{x} - \boldsymbol{x}_n) \end{aligned}$$

of $g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n)$. We now seek the best linear perturbation $\boldsymbol{x}_n + \boldsymbol{G}_n\boldsymbol{\beta}$ of $\boldsymbol{x}_n$ by minimizing $q_\rho(\boldsymbol{x}_n + \boldsymbol{G}_n\boldsymbol{\beta} \mid \boldsymbol{x}_n)$ with respect to the coefficient vector $\boldsymbol{\beta}$. To achieve this end, we solve the stationarity equation

$$\boldsymbol{0} \quad = \quad \boldsymbol{G}_n^t\nabla g_\rho(\boldsymbol{x}_n \mid \boldsymbol{x}_n) + \boldsymbol{G}_n^t(\boldsymbol{H}_n + \rho\boldsymbol{D}^t\boldsymbol{D})\boldsymbol{G}_n\boldsymbol{\beta}$$

and find $\boldsymbol{\beta} = -[\boldsymbol{G}_n^t(\boldsymbol{H}_n + \rho\boldsymbol{D}^t\boldsymbol{D})\boldsymbol{G}_n]^{-1}\boldsymbol{G}_n^t\nabla g_\rho(\boldsymbol{x}_n \mid \boldsymbol{x}_n)$, where the gradient is

$$\nabla g_\rho(\boldsymbol{x}_n \mid \boldsymbol{x}_n) = \nabla h_\rho(\boldsymbol{x}_n) = \nabla f(\boldsymbol{x}_n) + \rho\boldsymbol{D}^t[\boldsymbol{D}\boldsymbol{x}_n - \mathcal{P}(\boldsymbol{D}\boldsymbol{x}_n)].$$

The indicated matrix inverse is just $k \times k$.

2.2.3 ADMM

Alternating Direction Method of Multipliers (ADMM) is a potential competitor to the proximal distance algorithms just described [HLR16]. ADMM is designed to minimize functions of the form $f(\boldsymbol{x})+g(\boldsymbol{D}\boldsymbol{x})$ subject to $\boldsymbol{x} \in C$, where C is closed and convex. Splitting variables leads to the revised objective $f(\boldsymbol{x}) + g(\boldsymbol{y})$ subject to $\boldsymbol{x} \in C$ and $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$. ADMM invokes the augmented Lagrangian

$$\mathcal{A}_\mu(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\lambda}) \quad = \quad f(\boldsymbol{x}) + g(\boldsymbol{y}) + \boldsymbol{\lambda}^t(\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y}) + \frac{\mu}{2}\|\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y}\|^2$$

with Lagrange multiplier $\boldsymbol{\lambda}$ and step length $\mu > 0$. At iteration $n+1$ of ADMM one calculates successively

$$\begin{aligned}
\boldsymbol{x}_{n+1} &= \operatorname{argmin}_{\boldsymbol{x}\in C}\left[f(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y}_n + \boldsymbol{\lambda}_n\|^2\right] && (2.3)\\
\boldsymbol{y}_{n+1} &= \operatorname{argmin}_{\boldsymbol{y}}\left[g(\boldsymbol{y}) + \frac{\mu}{2}\|\boldsymbol{D}\boldsymbol{x}_{n+1} - \boldsymbol{y} + \boldsymbol{\lambda}_n\|^2\right] && (2.4)\\
\boldsymbol{\lambda}_{n+1} &= \boldsymbol{\lambda}_n + \mu(\boldsymbol{D}\boldsymbol{x}_{n+1} - \boldsymbol{y}_{n+1}). && (2.5)
\end{aligned}$$

Update (2.3) succumbs to Newton's method when $f(\boldsymbol{x})$ is smooth and $C = \mathbb{R}^p$, and update (2.4) succumbs to a proximal map of $g(\boldsymbol{y})$. Update (2.5) of the Lagrange multiplier $\boldsymbol{\lambda}$ amounts to steepest ascent on the dual function. A standard extension to the scheme in (2.3) through (2.5) is to vary the step length μ by considering the magnitude of residuals [BPC11]. For example, letting $\boldsymbol{r}_n = \boldsymbol{D}\boldsymbol{x} - \boldsymbol{y}$ and $\boldsymbol{s}_n = \mu\boldsymbol{D}^t(\boldsymbol{y}_{n-1} - \boldsymbol{y}_n)$ denote primal and dual residuals at iteration n, we make use of the heuristic

$$\mu_{n+1} = \begin{cases} 2\ \mu_n, & \text{if } \|\boldsymbol{r}_n\|/\|\boldsymbol{s}_n\| > 10 \\ \mu_n/2, & \text{if } \|\boldsymbol{r}_n\|/\|\boldsymbol{s}_n\| < 10 \\ \mu_n, & \text{otherwise} \end{cases}$$

which (a) keeps the primal and dual residuals within an order of magnitude of each other, (b) makes ADMM less sensitive to the choice of step length, and (c) improves convergence.

Our problem conforms to the ADMM paradigm when S is equal to the Cartesian product $\prod_{i=1}^{r} S_i$ and $g(\boldsymbol{y}) = \frac{\rho}{2}\,\mathrm{dist}(\boldsymbol{y}, S)^2$. Fortunately, the $\boldsymbol{y}$ update (2.4) reduces to a simple formula [BC17]. To derive this formula, note that the proximal map $\boldsymbol{y} = \mathrm{prox}_{\alpha g}(\boldsymbol{z})$ satisfies the stationarity condition

$$\boldsymbol{0} \quad = \quad \boldsymbol{y} - \boldsymbol{z} + \alpha[\boldsymbol{y} - \mathcal{P}(\boldsymbol{y})]$$

for any $\boldsymbol{z}$, including $\boldsymbol{z} = \boldsymbol{D}\boldsymbol{x}_{n+1} + \boldsymbol{\lambda}_n$, and any α, including $\alpha = \rho/\mu$. Since the projection map $\mathcal{P}(\boldsymbol{y})$ has the constant value $\mathcal{P}(\boldsymbol{z})$ on the line segment $[\boldsymbol{z}, \mathcal{P}(\boldsymbol{z})]$, the value

$$\mathrm{prox}_{\alpha g}(\boldsymbol{z}) \quad = \quad \frac{\alpha}{1+\alpha}\mathcal{P}(\boldsymbol{z}) + \frac{1}{1+\alpha}\boldsymbol{z}$$

satisfies the stationarity condition. Because the explicit update (2.4) for $\boldsymbol{y}$ decreases the Lagrangian even when S is nonconvex, we will employ it generally.

The $\boldsymbol{x}$ update (2.3) is given by the proximal map $\mathrm{prox}_{\mu^{-1}f}(\boldsymbol{\lambda}_n - \boldsymbol{y}_n)$ when $S = \mathbb{R}^p$ and $\boldsymbol{D} = \boldsymbol{I}$. Otherwise, the update of $\boldsymbol{x}$ is more problematic. Assuming $f(\boldsymbol{x})$ is smooth and $S = \mathbb{R}^p$, Newton's method gives the approximate update

$$\boldsymbol{x}_{n+1} \quad = \quad \boldsymbol{x}_n - \Big[d^2 f(\boldsymbol{x}_n) + \mu \boldsymbol{D}^t \boldsymbol{D}\Big]^{-1}\Big[\nabla f(\boldsymbol{x}_n) + \mu \boldsymbol{D}^t(\boldsymbol{D}\boldsymbol{x}_n - \boldsymbol{y}_n + \boldsymbol{\lambda}_n)\Big].$$

Our earlier suggestion of replacing $d^2 f(\boldsymbol{x}_n)$ by a positive definite approximation also applies in this context. Let us emphasize that ADMM eliminates the need for distance majorization. Although distance majorization is convenient, it is not necessarily a tight majorization. Thus, one can hope to see gains in rates of convergence. Balanced against this positive is the fact that ADMM is often slow to converge to high accuracy.

2.3 Convergence Analysis

Let us commence by establishing the existence of a minimum point. Further constraints on $\boldsymbol{x}$ beyond those imposed in the distance penalties are ignored or rolled into the essential domain of $f(\boldsymbol{x})$ when $f(\boldsymbol{x})$ is convex. As noted earlier, we can assume a single fusion matrix $\boldsymbol{D}$ and a single closed convex constraint set S. In such setting we have the following result.

Proposition 2.3.1 *Suppose the convex function* $f(\boldsymbol{x})$ *on* $\mathbb{R}^p$ *possesses a unique minimum point* $\boldsymbol{y}$ *on the closed convex set* $T = \boldsymbol{D}^{-1}(S)$*. Then for all sufficiently large* ρ*, the objective function* $h_\rho(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{\rho}{2}\operatorname{dist}(\boldsymbol{D}\boldsymbol{x}, S)^2$ *is coercive and therefore attains its minimum value.*

Next we show that the majorization surrogate defined in (2.2) attains its minimum value for large enough ρ.

Proposition 2.3.2 *Under the conditions of Proposition 2.3.1, for sufficiently large* ρ*, every surrogate* $g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n) = f(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{D}\boldsymbol{x} - \mathcal{P}(\boldsymbol{D}\boldsymbol{x}_n)\|^2$ *is coercive and therefore attains its minimum value. If*

$$f(\boldsymbol{x}) \quad \geq \quad f(\boldsymbol{x}_n) + \boldsymbol{v}_n^t(\boldsymbol{x} - \boldsymbol{x}_n) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_n)^t \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_n)$$

for all $\boldsymbol{x}$ *and some positive semidefinite matrix* $\boldsymbol{A}$ *and subgradient* $\boldsymbol{v}_n$ *at* $\boldsymbol{x}_n$*, and if the inequality* $\boldsymbol{u}^t\boldsymbol{A}\boldsymbol{u} > 0$ *holds whenever* $\|\boldsymbol{D}\boldsymbol{u}\| = 0$ *and* $\boldsymbol{u} \neq \boldsymbol{0}$*, then for* ρ *sufficiently large,* $g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n)$ *is strongly convex and hence coercive.*

We continue to assume that $f(\boldsymbol{x})$ and S are convex and that a minimum point

$$\boldsymbol{x}_{n+1} \quad \in \quad \underset{\boldsymbol{x} \in S}{\operatorname{argmin}}\, g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n) \tag{2.6}$$

of the surrogate $g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n)$ is available. Uniqueness of $\boldsymbol{x}_{n+1}$ holds when $g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n)$ is strictly convex. The constraint set S is implicitly captured by the essential domain of $f(\boldsymbol{x})$. Our earlier research shows that moving some constraints to the essential domain of $f(\boldsymbol{x})$ is sometimes helpful [KZL19, Lan16]. In any event, in our ideal convex setting we have a first convergence result for fixed ρ.

Proposition 2.3.3 *Supposes a) that* S *is closed and convex, b) that the loss* $f(\boldsymbol{x})$ *is convex and differentiable, and c) that the constrained problem possesses a unique minimum point. For* ρ *sufficiently large, let* $\boldsymbol{z}_\rho$ *denote a minimal point of the objective* $h_\rho(\boldsymbol{x})$ *defined by equation (2.1). Then the MM iterates (2.6) satisfy*

$$0 \leq \quad h_\rho(\boldsymbol{x}_n) - h_\rho(\boldsymbol{z}_\rho) \quad \leq \quad \tfrac{\rho}{2(n+1)}\|\boldsymbol{D}(\boldsymbol{z}_\rho - \boldsymbol{x}_0)\|^2.$$

Furthermore, the iterates $h_\rho(\boldsymbol{x}_n)$ systematically decrease.

In even more restricted circumstances, one can prove linear convergence of function values in the framework of [KNS16].

Proposition 2.3.4 *Suppose that S is a closed and convex set and that the loss $f(\boldsymbol{x})$ is L-smooth and μ-strongly convex. Then the objective function $h_\rho(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{\rho}{2}\operatorname{dist}(\boldsymbol{D}\boldsymbol{x}, S)^2$ possesses a unique minimum point $\boldsymbol{z}_\rho$, and the proximal distance iterates $\boldsymbol{x}_n$ satisfy*

$$0 \;\le\; h_\rho(\boldsymbol{x}_n) - h_\rho(\boldsymbol{z}_\rho) \;\le\; \left[1 - \frac{\mu^2}{2(L+\rho\|\boldsymbol{D}\|^2)^2}\right]^n [h_\rho(\boldsymbol{x}_0) - h_\rho(\boldsymbol{z}_\rho)].$$

The convergence properties of ADMM have been well studied in the optimization literature [Bec17]. To avail ourselves of the known results, we define three functions

$$\begin{aligned} H_\rho(\boldsymbol{x}, \boldsymbol{y}) &= f(\boldsymbol{x}) + \frac{\rho}{2}\operatorname{dist}(\boldsymbol{y}, S)^2 \\ \mathcal{L}_\rho(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}) &= H_\rho(\boldsymbol{x}, \boldsymbol{y}) + \lambda^t(\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y}) \\ q(\boldsymbol{\lambda}) &= \min_{\boldsymbol{x},\boldsymbol{y}} \mathcal{L}_\rho(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}), \end{aligned}$$

the second and third being the Lagrangian and dual function. This notation leads to following result; see [Bec17] for an accessible proof.

Proposition 2.3.5 *Suppose that S is closed and convex and that the loss $f(\boldsymbol{x})$ is proper, closed, and convex with domain whose relative interior is nonempty. Also assume the dual function $q(\boldsymbol{\lambda})$ achieves its maximum value. If the objective $f(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{D}\boldsymbol{x}\|^2 + \boldsymbol{a}^t\boldsymbol{x}$ achieves its minimum value for all $\boldsymbol{a} \neq \boldsymbol{0}$, then the ADMM running averages*

$$\bar{\boldsymbol{x}}_n = \frac{1}{n}\sum_{k=1}^n \boldsymbol{x}_k \quad \text{and} \quad \bar{\boldsymbol{y}}_n = \frac{1}{n}\sum_{k=1}^n \boldsymbol{y}_k$$

satisfy

$$\begin{aligned} |H_\rho(\bar{\boldsymbol{x}}_n, \bar{\boldsymbol{y}}_n) - h_\rho(\boldsymbol{x}_\rho)| &= O\Big(\frac{\rho}{n}\Big) \\ \|D\bar{\boldsymbol{x}}_n - \boldsymbol{y}_n\| &= O\Big(\frac{1}{n}\Big). \end{aligned}$$

Note that Proposition 2.3.2 furnishes a sufficient condition under which the functions $f(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{D}\boldsymbol{x}\|^2 + \boldsymbol{a}^t\boldsymbol{x}$ achieve their minima. Linear convergence holds under stronger assumptions.

Proposition 2.3.6 *Suppose that S is closed and convex, that the loss $f(\boldsymbol{x})$ is L-smooth and μ-strongly convex, and that the map determined by $\boldsymbol{D}$ is onto. Then the ADMM iterates converge at a linear rate.*

[GB16] proved Proposition 2.3.6 by operator methods. A range of convergence rates is specified there.

2.4 Numerical Examples

This section considers five concrete examples of constrained optimization amenable to distance majorization with fusion constraints. In each case, the loss function is both strongly convex and differentiable. The specific examples that we consider are the metric projection problem, convex regression, convex clustering, image denoising with total variation, and projection of a matrix to one with a better condition number. Each example is notable for the large number of fusion constraints. Except for convex clustering, they also involve convex constraint sets. In convex clustering we encounter a sparse constraint set. Our previous publications feature many nonconvex problems [KZL19, XCL17]. In describing each problem, we reserve the symbol $\boldsymbol{D}$ for the $m \times p$ fusion matrix.

Our numerical experiments compare proximal distance algorithms to ADMM. We consider two variants of proximal distance algorithms. The first directly minimizes the majorizing surrogate (MM), while the second performs steepest descent (SD) on it. Steepest descent increases the number of iterations until convergence but avoids solving large linear systems. Moreover, steepest descent avoids explicitly allocating memory for large fusion matrices; these can be computed on the fly as needed. In addition to the aforementioned methods, we tried the subspace MM algorithm described in Section 2.2.2. Unfortunately, this method was outperformed in both time and accuracy comparisons by Nesterov accelerated MM; the MM subspace results are therefore omitted. We also omit comparisons on the choice of algorithm to solve linear systems; we found that the method of conjugate gradients sacrificed little accuracy and largely outperformed LSQR. More details of the comparisons and their implementations appear in the appendices.

2.4.1 Mathematical Descriptions

Next we provide the mathematical details for each example.

2.4.1.1 Metric Projection

Solutions to the metric projection problem restore transitivity to noisy distance data for the n nodes of a graph [BDS08, STD05]. The data are encoded in an $n \times n$ dissimilarity matrix $\boldsymbol{Y} = (y_{ij})$ with nonnegative weights in the matrix $\boldsymbol{W} = (w_{ij})$. The metric projection problem requires finding the symmetric semi-metric $\boldsymbol{X} = (x_{ij})$ minimizing

$$f(\boldsymbol{X}) \;=\; \sum_{i>j} w_{ij}(x_{ij} - y_{ij})^2$$

subject to all $\binom{n}{2}$ nonnegativity constraints $x_{ij} \geq 0$ and all $3\binom{n}{3}$ triangle inequality constraints $x_{ij} - x_{ik} - x_{kj} \leq 0$. The diagonal entries of $\boldsymbol{Y}$, $\boldsymbol{W}$, and $\boldsymbol{X}$ are zero by definition. The fusion matrix $\boldsymbol{D}$ has $m = \binom{n}{2} + 3\binom{n}{3}$ rows, and the projected value of $\boldsymbol{DX}$ must fall in the set S of symmetric matrices satisfying all pertinent constraints.

One can simplify the required projection by stacking the nonredundant entries along each successive column of $\boldsymbol{X}$ to create a vector $\boldsymbol{x}$ with $\binom{n}{2}$ entries. This captures the lower triangle of $\boldsymbol{X}$. The sparse matrix $\boldsymbol{D}$ is correspondingly redefined to be $m \times \binom{n}{2}$. These maneuvers simplify constraints to $\boldsymbol{Dx} \leq \boldsymbol{0}$, and projection involves sending each entry u of $\boldsymbol{Dx}$ to $\min\{0, u\}$. The linear system $(\boldsymbol{I} + \rho \boldsymbol{D}^t \boldsymbol{D})\boldsymbol{x} = \boldsymbol{b}$ appears in both the MM and

ADMM updates for $\boldsymbol{x}_n$. Application of the Woodbury and Sherman-Morrison formulas yield an exact solution to the linear system and allow one to forgo iterative methods.

2.4.1.2 Convex Regression

Convex regression is a nonparametric method for estimating a regression function under shape constraints. Given n responses y_i and corresponding predictors $\boldsymbol{x}_i \in \mathbb{R}^d$, the goal is to find the convex function $\psi(\boldsymbol{x})$ minimizing the sum of squares $\frac{1}{2}\sum_{i=1}^n [y_i - \psi(\boldsymbol{x}_i)]^2$. Asymptotic and finite sample properties of this convex estimator have been described in detail by [SS11]. The convex regression program can be restated as the finite dimensional problem of finding the value θ_i and subgradient $\boldsymbol{\xi}_i \in \mathbb{R}^d$ of $\psi(\boldsymbol{x})$ at each sample point $(y_i, \boldsymbol{x}_i)$. Convexity imposes the supporting hyperplane constraint $\theta_j + \boldsymbol{\xi}_j^t(\boldsymbol{x}_i - \boldsymbol{x}_j) \leq \theta_i$ for each pair $i \neq j$. Thus, the problem becomes one of minimizing $\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|^2$ subject to these $m = n(n-1)$ inequality constraints. In the proximal distance framework, we must minimize

$$h_\rho(\boldsymbol{\theta}, \Xi) \quad = \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\theta}\|^2 + \frac{\rho}{2}\operatorname{dist}(\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{B}\Xi, \mathbb{R}^m_-)^2,$$

where $\boldsymbol{D} = [\boldsymbol{A}\ \boldsymbol{B}]$ encodes the required fusion matrix.

2.4.1.3 Convex Clustering

Convex clustering of n samples based on d features can be formulated in terms of the regularized objective

$$F_\gamma(\boldsymbol{U}) \quad = \quad \frac{1}{2}\|\boldsymbol{U} - \boldsymbol{X}\|_F^2 + \gamma \sum_{i>j} w_{ij}\|\boldsymbol{u}_i - \boldsymbol{u}_j\|,$$

where $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ encodes the data, and the columns $\boldsymbol{u}_i$ of $\boldsymbol{U} \in \mathbb{R}^{d\times n}$ represent centroids assigned to each sample. The predetermined weights w_{ij} have a graphical interpretation under which similar samples have positive edge weights w_{ij} and distant samples have 0 edge weights. The edge weights are chosen by the user to guide the clustering process. In general, minimization of $F_\gamma(\boldsymbol{U})$ separates over the connected components of the graph. To allow all sample points to coalesce into a single cluster, we assume that the underlying graph is connected. The regularization parameter $\gamma > 0$ tunes the number of clusters in a nonlinear fashion and potentially captures hierarchical information. Previous work establishes that the solution path $\boldsymbol{U}(\gamma)$ varies continuously with respect to γ [CL15]. Unfortunately, there is no explicit way to determine the number of clusters entailed by a particular value of γ.

Alternatively, we can attack the problem using sparsity and distance majorization. Consider the penalized objective

$$h_\rho(\boldsymbol{U}) \;=\; \frac{1}{2}\|\boldsymbol{U} - \boldsymbol{X}\|_F^2 + \frac{\rho}{2}\operatorname{dist}(\boldsymbol{U}\boldsymbol{D}, S_\nu)^2.$$

The fusion matrix $\boldsymbol{D}$ has $m = \binom{n}{2}$ columns $w_{ij}(\boldsymbol{e}_i - \boldsymbol{e}_j)$ and serves to map the centroid matrix $\boldsymbol{U}$ to a $d \times m$ matrix $\boldsymbol{V}$ encoding the weighted differences $w_{ij}(\boldsymbol{u}_i - \boldsymbol{u}_j)$. The members of the sparsity set S_ν are $d \times m$ matrices with at most ν non-zero columns. Projection of $\boldsymbol{V}$ onto the closed set S_ν is straightforward to implement by sorting the Euclidean lengths of the columns of $\boldsymbol{V}$ and sending to $\boldsymbol{0}$ all but the ν most dominant columns. Ties are broken arbitrarily. Our sparsity-based method trades the continuous penalty parameter $\gamma > 0$ in the previous formulation for an integer sparsity index $\nu \in \{0, 1, 2, \ldots, \binom{n}{2}\}$. For example with $\nu = 0$, all differences $\boldsymbol{u}_i - \boldsymbol{u}_j$ equal $\boldsymbol{0}$, and all sample points cluster together. The other extreme $\nu = \binom{n}{2}$ assigns each point to its own cluster. The size of the matrices $\boldsymbol{D}$ and $\boldsymbol{V}$ can be reduced by discarding column pairs corresponding to 0 weights.

2.4.1.4 Total Variation Image Denoising

To approximate an image $\boldsymbol{U}$ from a noisy input $\boldsymbol{W}$ matrix, [ROF92] regularize a loss function $f(\boldsymbol{U})$ by a total variation (TV) penalty. After discretizing the problem, the least squares loss leads to the objective

$$F_\gamma(\boldsymbol{U}) \;=\; \sum_{i,j}(U_{i,j} - W_{i,j})^2 + \gamma \sum_{i,j}\sqrt{(U_{i+1,j} - U_{i,j})^2 + (U_{i,j+1} - U_{i,j})^2},$$

where $\boldsymbol{U}, \boldsymbol{W} \in \mathbb{R}^{n \times p}$ are rectangular monochromatic images and γ controls the strength of regularization. The anisotropic norm

$$\mathrm{TV}_1(\boldsymbol{U}) \quad = \quad \sum_{i,j} |U_{i+1,j} - U_{i,j}| + |U_{i,j+1} - U_{i,j}| \quad = \quad \|\boldsymbol{D}_n \boldsymbol{U}\|_1 + \|\boldsymbol{U}\boldsymbol{D}_p^t\|_1$$

is often preferred because it induces sparsity in the differences. Here $\boldsymbol{D}_p$ is the forward difference operator on p data points. Stacking the columns of $\boldsymbol{U}$ into a vector $\boldsymbol{u} = \mathrm{vec}(\boldsymbol{U})$ allows one to identify a fusion matrix $\boldsymbol{D}$ and write $\mathrm{TV}_1(\boldsymbol{U})$ more compactly as $\mathrm{TV}_1(\boldsymbol{u}) = \|\boldsymbol{D}\boldsymbol{u}\|_1$. In this context we reformulate the denoising problem as minimizing $f(\boldsymbol{U})$ subject to the set constraint $\|\boldsymbol{D}\boldsymbol{u}\|_1 \leq \gamma$. This revised formulation directly quantifies the quality of a solution and brings into play fast pivot-based algorithms for projecting onto multiples of the ℓ_1 unit ball [Con16].

2.4.1.5 Projection of a Matrix to a Good Condition Number

Consider an $m \times p$ matrix $\boldsymbol{M}$ with $m \geq p$ and full singular value decomposition $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t$. The condition number of $\boldsymbol{M}$ is the ratio $\sigma_{\max}/\sigma_{min}$ of the largest to the smallest singular value of $\boldsymbol{M}$. We denote the diagonal of $\boldsymbol{\Sigma}$ as $\boldsymbol{\sigma}$. Owing to the von Neumann-Fan inequality, the closest matrix $\boldsymbol{N}$ to $\boldsymbol{M}$ in the Frobenius norm has singular value decomposition $\boldsymbol{N} = \boldsymbol{U}\boldsymbol{X}\boldsymbol{V}^t$, where the diagonal $\boldsymbol{x}$ of $\boldsymbol{X}$ satisfies inequalities pertinent to a decent condition number [BL10]. Suppose $c \geq 1$ is the maximum condition number. Then every pair (x_i, x_j) satisfies $x_i - cx_j \leq 0$. Note that $x_i - cx_i > 0$ if and only if $x_i < 0$. Thus, nonnegativity of the entries of $\boldsymbol{x}$ is enforced. The proximal distance approach to the condition number projection problem invokes the objective and majorization

$$\begin{aligned}
h_\rho(\boldsymbol{x}) \quad &= \quad \frac{1}{2}\|\boldsymbol{\sigma} - \boldsymbol{x}\|^2 + \frac{\rho}{2}\sum_{(i,j)} \mathrm{dist}(x_i - cx_j, \mathbb{R}_-)^2 \\
&= \quad \frac{1}{2}\|\boldsymbol{\sigma} - \boldsymbol{x}\|^2 + \frac{\rho}{2}\sum_{(i,j)} (x_i - cx_j)_+^2 \\
&\leq \quad \frac{1}{2}\|\boldsymbol{\sigma} - \boldsymbol{x}\|^2 + \frac{\rho}{2}\sum_{(i,j)} (x_i - cx_j - q_{nij})^2
\end{aligned}$$

at iteration n, where $q_{nij} = \min\{x_{ni} - cx_{nj}, 0\}$. We can write the majorization more concisely as

$$h_\rho(\boldsymbol{x}) \;\leq\; \frac{1}{2}\|\boldsymbol{A}_\rho \boldsymbol{x} - \boldsymbol{r}_n\|^2, \quad \boldsymbol{A}_\rho \;=\; \begin{pmatrix} \boldsymbol{I}_p \\ \sqrt{\rho}\boldsymbol{D} \end{pmatrix}, \quad \boldsymbol{r}_n \;=\; \begin{pmatrix} \boldsymbol{\sigma} \\ \sqrt{\rho}\,\mathrm{vec}\,\boldsymbol{Q}_n \end{pmatrix},$$

where $\mathrm{vec}\,\boldsymbol{Q}_n$ stacks the columns of $\boldsymbol{Q}_n = (q_{nij})$ and the $p^2 \times p$ fusion matrix $\boldsymbol{D}$ satisfies $(\boldsymbol{D}\boldsymbol{x})_k = x_i - cx_j$ for each component k. The minimum of the surrogate occurs at $\boldsymbol{x}_{n+1} = (\boldsymbol{A}_\rho^t \boldsymbol{A}_\rho)^{-1}\boldsymbol{A}_\rho^t \boldsymbol{r}_n$. This linear system can be solved exactly.

2.4.2 Numerical Results

We now explain example by example the implementation details behind our efforts to benchmark the three methods (MM, SD, and ADMM). The two proximal distance algorithm are subjected to Nesterov acceleration. Each method is initialized at the solution of the corresponding unconstrained problem. Performance is assessed in terms of speed in seconds or milliseconds, number of iterations until convergence, the converged value of the loss $f(\boldsymbol{x})$, and the converged distance to the constraint set $\mathrm{dist}(\boldsymbol{D}\boldsymbol{x}, S)$. For some problems, other performance metrics are highlighted.

In each example, our program is allotted an iteration budget to achieve convergence as defined both by the relative change in the loss

$$|f(\boldsymbol{x}_{n+1}) - f(\boldsymbol{x}_n)| \;\leq\; \epsilon_1[f(\boldsymbol{x}_n) + 1],$$

and the magnitude $\mathrm{dist}(\boldsymbol{D}\boldsymbol{x}, S) \leq \epsilon_2$ of the distance to the constraint set. We select annealing schedules from the family of geometric progressions $\rho(n) = ar^{\lfloor n/b \rfloor}$ with initial value $a = 1$, multiplier r, and modulus b. Specific choices for these parameters are described in each example.

	Time (s)			Loss $\times 10^{-3}$			Distance $\times 10^{3}$			Iterations		
n	MM	SD	ADMM	MM	SD	ADMM	MM	SD	ADMM	MM	SD	ADMM
32	0.1732	**0.1182**	0.2985	**0.5732**	**0.5732**	**0.5732**	0.9911	0.9934	**0.991**	**2005**	**2005**	**2005**
64	1.563	**1.218**	3.675	2.336	2.336	**2.335**	**0.9911**	0.9953	0.9963	2149	2149	**2129**
128	13.17	**9.396**	39.83	**9.171**	**9.171**	**9.171**	0.9826	0.9861	**0.9786**	**2124**	**2124**	**2124**
256	127	**97.01**	351.3	37.44	37.44	**37.43**	0.9976	0.975	**0.9771**	**2183**	2184	2184

Table 2.1: **Metric projection experiments.** Columns are scaled as indicated.

2.4.2.1 Metric projection

In our comparisons, we use input matrices $\boldsymbol{Y} \in \mathbb{R}^{n \times n}$ whose i.i.d. entries y_{ij} are drawn uniformly from the interval $[0, 10]$ and set weights $w_{ij} = 1$. Each algorithm is allotted 3000 iterations to achieve a relative change in the loss of $\epsilon_1 = 10^{-6}$ and the distance to feasibility of $\epsilon_1 = 10^{-3}$. The annealing schedule is set to $\rho(n) = \min\{10^6, 1.1^{\lfloor n/20 \rfloor}\}$ for the proximal distance methods. Table 2.1 summarizes the performance of the three algorithms. Best values appear in boldface. The three algorithms take a comparable number of iterations to converge and deliver solutions of comparable quality. Interestingly, ADMM appears to have a slight edge in achieving a smaller distance penalty with fewer iterations. These positives are outweighed by the large cost in terms of compute time with SD clearly being the fastest algorithm as the problem size increases.

2.4.2.2 Convex regression

In convex regression one is given n responses y_i and corresponding predictor vectors $\boldsymbol{x}_i \in \mathbb{R}^d$. The goal is to find the convex function $\psi(\boldsymbol{x})$ minimizing the sum of squares $\frac{1}{2}\sum_{i=1}^{n}[y_i - \psi(\boldsymbol{x}_i)]^2$. Recall the problem reduces to estimating the value θ_i and subgradients $\boldsymbol{\xi}_j$ at each point $\boldsymbol{x}_i$. In our numerical examples the y_i are independent Gaussian deviates with means $\psi(\boldsymbol{x}_i)$ and common variance $\sigma^2 = 0.1$. The predictors are i.i.d. deviates sampled from the uniform distribution on $[-1, 1]^d$. We choose the simple convex function $\psi(\boldsymbol{x}_i) = \|\boldsymbol{x}_i\|^2$ for our benchmarks for ease in interpretation; the interested reader may consult the work of [MCI19] for a detailed account of the applicability of the technique in general. Each algorithm is allotted 3000 iterations to converge with $\epsilon_1 = 10^{-6}$, $\epsilon_2 = 10^{-3}$, and $\rho(n) = \min\{10^6, 1.1^{\lfloor n/20 \rfloor}\}$.

Table 2.2 shows that although ADMM is a distant third in speed, in the large-scale problems it strongly outperforms both MM and SD on the metrics of loss and mean squared error (MSE). It has a poorer distance to feasibility. As for MM and SD, they are comparable in accuracy, but SD is much faster. It is possible that the annealing schedule in large-scale problems is ill adapted to the convergence properties of MM and SD, but we are reluctant to generalize from such a small sample.

		Time (s)			Loss $\times 10^3$			Distance $\times 10^3$			MSE $\times 10^3$		
d	n	MM	SD	ADMM	MM	SD	ADMM	MM	SD	ADMM	MM	SD	ADMM
1	50	0.1104	**0.07588**	0.1527	**193.1**	197.1	207.9	**0.9799**	0.9996	0.9818	1.814	**1.593**	2.327
2	50	0.1558	**0.02733**	0.7126	**56.21**	66.28	57.02	0.5813	0.7095	**0.2646**	5.961	**5.587**	5.885
10	50	0.3415	**0.01845**	1.546	**0.00206**	0.0704	0.03454	**0**	**0**	0.07506	10.61	**10.54**	10.58
20	50	1.047	**0.04032**	3.643	0.0005515	**0.0002422**	0.01308	**0**	**0**	0.02146	9.69	**9.681**	9.702
1	100	0.4035	**0.1835**	1.583	452.1	474.5	**441.5**	**0.6618**	0.9691	0.7512	1.004	**0.4323**	1.046
2	100	0.7724	**0.152**	3.233	416.3	468.5	**350.1**	0.8624	**0.7612**	0.8727	**1.626**	2.065	1.977
10	100	3.238	**0.2226**	15.44	1.789	11.02	**0.08498**	0.9708	0.6192	**0.09576**	10.34	10.62	**10.29**
20	100	5.848	**0.2972**	23.63	**0.000179**	0.2437	0.03805	0.06488	0.5855	**0.003289**	7.692	7.691	**7.687**
1	200	2.699	**1.764**	7.081	1218	1321	**1134**	0.9534	**0.5567**	1.725	0.8839	1.773	**0.2313**
2	200	4.581	**1.295**	13.58	1114	1544	**792.4**	1.352	**0.9655**	3.292	3.985	8.092	**1.721**
10	200	18.07	**1.711**	110.3	276.9	343.9	**0.2766**	0.8843	0.9989	**0.05209**	12.21	13	**9.261**
20	200	69.93	**3.028**	289	159.8	468.9	**0.1646**	0.5258	0.5957	**0.04089**	11.61	15.37	**9.741**
1	400	18.71	**10.35**	40.43	2417	3333	**1818**	**0.9301**	0.9534	2.818	3.624	8.373	**0.5421**
2	400	28.98	**12.38**	74.98	3992	6506	**1625**	1	**0.05938**	7.162	12.28	24.67	**1.449**
10	400	125.8	**15.38**	734.6	4179	4209	**205.4**	**0.9679**	0.9783	1.908	23.91	23.33	**9.596**
20	400	288.2	**25.28**	1917	3291	3257	**24.34**	0.918	**0.7363**	1.754	22.91	23.16	**9.809**

Table 2.2: **Convex regression experiments.** Columns are scaled as indicated. The mean squared error (MSE) is based on the ground truth $\boldsymbol{y}_i$ and estimates $\boldsymbol{\theta}_i$.

2.4.2.3 Convex clustering

To evaluate the performance of the different methods on convex clustering, we consider a mixture of simulated data and discriminant analysis data from the UCI Machine Learning Repository. The simulated data in `gaussian300` consists of 3 Gaussian clusters generated from bivariate normal distributions with means $\boldsymbol{\mu} = (0.0, 0.0)^t$, $(2.0, 2.0)^t$, and $(1.8, 0.5)^t$, standard deviation $\sigma = 0.1$, and class sizes $n_1 = 150, n_2 = 50, n_3 = 100$. This easy dataset is included to validate Algorithm 1 described later as a reasonable solution path heuristic. The data in `iris` and `zoo` are representative of clustering with purely continuous or purely discrete data, respectively. In these two datasets samples with same class label form a cluster. Finally, the simulated data `spiral500` is a classic example that thwarts k-means clustering. Each algorithm is allotted 3000 iterations to converge with $\epsilon_1 = 10^{-6}$, $\epsilon_2 = 10^{-4}$, and $\rho(n) = \min\{10^6, 1.2^{\lfloor n/20 \rfloor}\}$.

				Time (s)		Distance $\times 10^3$		ARI	
dataset	d	n	k	MM	SD	MM	SD	MM	SD
zoo	16	101	7	67.19	**43.26**	**0.9192**	0.92	**0.8324**	0.7474
iris	4	150	3	52.49	**52.09**	**15.3**	**15.3**	**0.5895**	**0.5895**
gaussian300	2	300	3	119.9	**91.05**	**0.9989**	**0.9989**	**1**	**1**
spiral500	2	500	2	148.4	**65.34**	**0.834**	0.8359	**0.1642**	**0.1642**

Table 2.3: **Convex clustering experiments.** The adjusted Rand index (ARI) measures the closeness of reconstructed clusters to underlying true clusters. Columns are scaled as indicated. Times reflect the total time spent generating candidate cluster assignments. The Distance and ARI values correspond to the optimal clustering on the basis of maximal ARI.

Because the number of clusters is usually unknown, we implement the search heuristic outlined in Algorithm 1. The idea behind the heuristic is to gradually coerce clustering without exploring the full range of sparsity levels ν. Our procedure generates a list of candidate clusters that can be evaluated by various measures of similarity [VEB10]. ADMM is not remotely competitive on these examples given its extremely long compute times;

these times are only exacerbated by the search heuristic. The findings reported in Table 2.3 indicate about the same accuracy for MM and SD as measured by loss and distance to feasibility. The adjusted Rand index (ARI) provides a reasonable measure of the distance to the ground truth in our examples. Both MM and SD achieve similar ARI values on each dataset. Notably, the combination of the proximal distance algorithms and the overall search heuristic (Algorithm 1) yields nearly perfect clusters in the `gaussian300` example. To its disadvantage, the search heuristic is greedy and generally requires tuning.

Algorithm 1 Search Heuristic

```
procedure CVXCLUSTERPATH($\boldsymbol{X}$, $s_0$, $s_{\text{step}}$)
    $\boldsymbol{U} \leftarrow \boldsymbol{X}$                                   ▷ Initialize centroid assignments.
    $s \leftarrow s_0$                                      ▷ Initialize sparsity level in $[0, 1]$.
    $\nu_{\max} \leftarrow \binom{n}{2}$                           ▷ Determine upper bound from samples $n$.
    while $s < 1$ do
        $\nu \leftarrow \text{round}((1 - s) \times \nu_{\max})$         ▷ Set parameter for sparse projection.
        $\boldsymbol{U} \leftarrow \text{argmin}\, h_\rho(\boldsymbol{U})$                 ▷ Minimize with choice of $\nu$.
        $s_{\text{proposal}} \leftarrow \text{count}(\boldsymbol{U})/\nu_{\max}$
                                     ▷ Propose new level based on satisfied constraints.
        if $s_{\text{proposal}} > s$ then
            $s \leftarrow s_{\text{proposal}}$                   ▷ Accept proposal if it increases sparsity.
        else
            $s \leftarrow s + s_{\text{step}}$                     ▷ Otherwise move by a fixed amount.
        end if
    end while
end procedure
```

image	width	height	Time (s)	Loss	Distance $\times 10^3$	MSE	PSNR
cameraman	512	512	743.3	4016	9.955	0.002727	25.64
peppers_gray	512	512	888.3	4138	9.631	0.002208	26.56

Table 2.4: **Image denoising experiments.** Results are for algorithm SD only. The mean squared error (MSE) and peak signal-to-noise ratio (PSNR) assess the quality of the reconstructed image relative to the ground truth. Columns are scaled as indicated. Times reflect the total time spent generating candidate images. Loss, Distance, MSE, and PSNR correspond to the optimal image on the basis of minimal MSE.

2.4.2.4 Total Variation Image Denoising

To evaluate our denoising algorithm, we consider two standard test images, `cameraman` and `peppers_gray`. White noise with $\sigma = 0.2$ is applied to an image and then reconstructed using a variant of Algorithm 1. Only SD is tested with a maximum of 5000 iterations and convergence thresholds $\epsilon_1 = 10^{-6}$ and $\epsilon_2 = 10^{-2}$. ADMM and MM are too slow to merit consideration. A gentle schedule $\rho(n) = \{10^6, 1.075^{\lfloor n/20 \rfloor}\}$ performs best for a broad range of sparsity levels. We adapt the search heuristic (Algorithm 1) to this example by replacing the parameter $\nu_{\max}$ with the total variation of the input image, $\mathrm{TV}_1(\boldsymbol{W})$. This provides a device that generates a solution path of images with varying levels of noise. Table 2.4 reports the quality of the images in terms of the Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR). Timings reflect the total time spent generating solutions, starting from a 50% reduction in the total variation of the input image. Figure 2.1 depicts the original and reconstructed images along the solution path.

2.4.2.5 Projection of a Matrix to a Good Condition Number

We generate base matrices $\boldsymbol{M} \in \mathbb{R}^{p \times p}$ as random correlation matrices using MatrixDepot.jl [ZH16], which relies on Davies' and Higham's refinement [DH00] of the Bendel-Mickey algorithm [BM78]. Our simulations generate matrices with condition numbers $\mathrm{c}(\boldsymbol{M})$ in the

Figure 2.1: **Sample images along the solution path of the search heuristic.** Images are arranged from left to right as follows: noisy input, first candidate image (50% reduction of total variation), best candidate by MSE, and candidate generated after the optimal image.

set $\{7.02, 107, 690\}$. Our subsequent analyses target condition number decreases in the percentage set $\{20, 40, 60, 80\}$. Each algorithm is allotted 5000 iterations to converge with $\epsilon_1 = 10^{-3}$, $\epsilon_2 = 10^{-2}$, and $\rho(n) = \min\{10^6, 1.1^{\lfloor n/20 \rfloor}\}$. Table 2.5 summarizes the performance of the three algorithms. Both MM and SD attain smaller losses than ADMM. On the other hand, the table suggests that ADMM has a slight edge in enforcing adherence to the constraint set. Notably, all three methods preserve much of the sign structure of the original correlation matrix.

2.4.3 Hybrid Algorithms

Motivated by ADMM's tendency to generate high quality solutions, we implemented a hybrid algorithm by combining ADMM with the speed of SD. The method is split into two phases. In Phase 1 we obtain the approximate solution $\boldsymbol{x}_{\text{sol}} = \operatorname{argmin} h_\rho(\boldsymbol{x})$ using SD. This is followed by projection of $\boldsymbol{x}_{\text{sol}}$ onto the constraint set in Phase 2. Because the projection lacks a closed-

			Time (ms)			Loss $\times 10^3$			Distance $\times 10^3$			c($\boldsymbol{X}$)		
p	c($\boldsymbol{M}$)	%	MM	SD	ADMM	MM	SD	ADMM	MM	SD	ADMM	MM	SD	ADMM
10	7.02	20	0.0364	**0.0322**	0.073	**3.84**	3.86	3.88	9.41	7.81	**7.32**	**5.64**	**5.64**	**5.64**
100	107	20	0.447	**0.441**	0.818	**0.0198**	**0.0198**	0.0283	4.67	4.67	**0**	86.1	86.1	**82.8**
1000	690	20	208	213	**152**	**0.000497**	**0.000497**	0.000624	7.64	7.64	**0**	554	554	**5**41
10	7.02	40	0.174	**0.173**	0.273	**28.1**	**28.1**	**28.1**	**9.61**	9.64	9.38	**4.23**	**4.23**	**4.23**
100	107	40	0.483	**0.474**	0.864	**0.15**	**0.15**	0.186	**0**	**0**	**0**	63.9	63.9	**61.1**
1000	690	40	270	278	**217**	**0.00438**	**0.00438**	0.00599	8.64	8.64	**0**	415	415	**413**
10	7.02	60	0.415	**0.383**	0.485	172	172	**170**	**2.46**	3.25	9.28	**2.81**	**2.81**	2.82
100	107	60	0.741	**0.739**	0.962	**0.88**	**0.88**	1.07	0.979	0.98	**0**	43	43	**41.7**
1000	690	60	269	280	**215**	**0.0369**	**0.0369**	0.0474	6.67	6.67	**0**	277	277	**274**
10	7.02	80	**0.407**	**0.407**	0.673	**1090**	**1090**	**1090**	9.27	**9.21**	9.26	**1.41**	**1.41**	**1.41**
100	107	80	0.712	**0.698**	4.76	12.7	12.7	**12.1**	0.438	**0.379**	9.45	**21.5**	**21.5**	**21.5**
1000	690	80	**270**	280	284	**0.469**	**0.469**	0.575	6.69	6.69	**0**	**138**	**138**	**138**

Table 2.5: **Condition number experiments.** Here c($\boldsymbol{M}$) is the condition number of the input matrix, % is the target percentage decrease in the condition number, and c($\boldsymbol{X}$) is the condition number of the result for a given algorithm.

form solutions, it must be computed by an iterative procedure such as ADMM. According to the classical penalty method, the projection is well approximated by solving the penalized problem

$$\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_{\text{sol}}\|^2 + \frac{\rho}{2}\,\text{dist}(\boldsymbol{D}\boldsymbol{x}, C)^2$$

for large ρ. In practice, we carry over to Phase 2 the penalty coefficient ρ achieved at the end of Phase 1, provided it is large enough. In our experience, this hybrid SD-ADMM procedure makes considerable progress in improving the loss and reducing the distance to feasibility. The alternative of refining the objective by initializing ADMM with the final value attained by SD is much less effective. Presumably this happens because ADMM requires a good dual variable (multiplier) to work well. By its nature SD operates entirely on the primal problem.

Table 2.6 summarizes our findings for the metric projection problem using the annealing schedule $\rho(n) = \min\{10^6, 1.1^{\lfloor n/20 \rfloor}\}$ and strict convergence parameters $\epsilon_1 = 10^{-3}$ and $\epsilon_2 = 10^{-6}$. In our experiments Phase 1 (SD) and Phase 2 (ADMM) are allotted 2000 and 1000

n	Time (s)	Loss $\times 10^{-3}$	Distance $\times 10^{3}$	Iterations
32	0.1232	0.5611	0	2009
64	1.137	2.303	0	2013
128	9.172	9.089	0	2014
256	91.97	37.28	0	2011

Table 2.6: **Metric projection using a hybrid SD/ADMM algorithm.** Columns are scaled as indicated. Iterations in excess of 2000 reflect the number of steps taken by ADMM to enforce feasibility.

d	n	Time (s)	MSE $\times 10^{3}$	Loss $\times 10^{3}$	Distance $\times 10^{3}$	Iterations
1	50	0.1042	1.593	196.7	1.093	3000
1	100	0.4763	0.4253	474.7	0.65	3000
1	200	2.486	1.777	1273	0.4957	3000
1	400	16.93	8.48	3352	0.4433	3000
2	50	0.05995	5.588	66.3	0	2036
2	100	0.6175	2.066	450.9	0.2435	3000
2	200	3.224	8.11	1542	0.432	3000
2	400	24.53	24.51	6471	0.4734	3000
10	50	0.01892	10.54	0.07002	0	201
10	100	1.264	10.59	9.308	0	2881
10	200	6.632	12.96	339.8	5.824×10^{-5}	3000
10	400	42.91	23.27	4192	2.353×10^{-5}	3000
20	50	0.1367	9.681	0.0002378	0	229
20	100	1.558	9.588	0.046	0	2558
20	200	7.438	30.84	461.4	0	2431
20	400	70.46	23.08	3240	6.889×10^{-5}	3000

Table 2.7: **Convex regression using a hybrid SD/ADMM algorithm.** Columns are scaled as indicated. Iterations in excess of 2000 reflect the number of steps taken by ADMM.

iterations, respectively. Comparing Table 2.1 with 2.6 reveals that the hybrid algorithm improves the loss and produces a feasible point with only a moderate increase in computing time.

However, Table 2.7 reports results on convex regression that counters the idea that the hybrid approach is strictly superior. Specifically, improvements to feasibility in Table 2.7 are less dramatic across problem sizes compared to SD in Table 2.2 and there is no little to no improvement to MSE compared to ADMM. These results are based on identical settings as in Table 2.2 with $\rho(n) = \min\{10^6, 1.5^{\lfloor n/100 \rfloor}\}$, $\epsilon_1 = 10^{-6}$, and $\epsilon_2 = 10^{-3}$.

2.5 Discussion

Let us recapitulate the main findings of our numerical experiments. Tables 2.1 through 2.5 show a consistent pattern of superior speed by the steepest descent (SD) versions of proximal distance algorithms. This is hardly surprising since unlike ADMM and MM, SD avoids solving a linear system at each iteration. Interestingly, the speed gap between the three algorithms is much narrower when the linear system can be solved exactly. This phenomenon is apparent in the condition number example summarized in Table 2.5. In our examples MM is usually faster than ADMM, which we attribute to the number of operations required per iteration. Specifically, ADMM requires more matrix-vector multiplications involving the fusion matrix $\boldsymbol{D}$. In fairness on convex regression, ADMM does a noticeably better job of minimizing loss and MSE. The hybrid algorithm combining SD followed by ADMM corrects the poor constraint satisfaction of SD but retains most of its speed. Tables 2.6 and 2.7 document this tendency.

Let us remind readers of some advantages and disadvantages of the proximal distance method. First, fusion constraints fit naturally in the proximal distance framework. Second, proximal distances enjoy the descent property. Third, there is a nearly optimal step size for gradient descent when second-order information is available on the loss. The main disadvantages of the proximal distance methods are (a) the overall slow convergence due to the lost of curvature information on the distance penalty and (b) the need for a reasonable annealing schedule. In practice, a little experimentation can yield a reasonable schedule for an entire class of problems. Fourth, proximal distance algorithms are competitive if not superior to ADMM on many problems. Fifth, proximal distance algorithms like iterative hard thresholding rely on set projection and are therefore helpful in dealing with hard sparsity constraints. Many methods are only capable of dealing with soft constraints imposed by the lasso and other convex penalties. To their detriment soft constraints entail severe parameter shrinkage and often lead to an excess of false positives in model selection.

We readily acknowledge that other algorithms may perform better than MM and proximal distance algorithms on specific problems. The triangle fixing algorithm for metric projection is a case in point [BDS08]. This objection obscures the generic utility of the proximal distance principle. ADMM can certainly be beat on many specific problems, but nobody seriously suggests that it be rejected across the board. Optimization, particularly constrained optimization, is a fragmented subject, with no clear winner across problem domains. Generic methods serve as workhorses, benchmarks, and backstops.

In closing, we would like to draw the reader's attention to some generalizations of the MM principle and connections to other well-studied algorithm classes. For instance, a linear fusion constraint $\boldsymbol{D}\boldsymbol{x} \in S$ can in principle by replaced by a nonlinear fusion constraint $M(\boldsymbol{x}) \in S$. The objective and majorizer are then

$$\begin{aligned} h_\rho(\boldsymbol{x}) &= f(\boldsymbol{x}) + \frac{\rho}{2}\operatorname{dist}[M(\boldsymbol{x}), S]^2 \\ g(\boldsymbol{x} \mid \boldsymbol{x}_n) &= f(\boldsymbol{x}) + \frac{\rho}{2}\|M(\boldsymbol{x}) - \mathcal{P}_S[M(\boldsymbol{x}_n)]\|^2. \end{aligned}$$

The objective has gradient $\boldsymbol{g} = \nabla f(\boldsymbol{x}) + \rho dM(\boldsymbol{x})^t\{M(\boldsymbol{x}) - \mathcal{P}_S[M(\boldsymbol{x})]\}$. The second differential of the majorizer is approximately $d^2f(\boldsymbol{x}) + \rho dM(\boldsymbol{x})^t dM(\boldsymbol{x})$ for $M(\boldsymbol{x})$ close to $\mathcal{P}_S[M(\boldsymbol{x})]$. Thus, gradient descent can be implemented with step size

$$\gamma = \frac{\|\boldsymbol{g}_n\|^2}{\boldsymbol{g}_n^t d^2 f(\boldsymbol{x}_n)\boldsymbol{g}_n + \rho\|dM(\boldsymbol{x}_n)\boldsymbol{g}_n\|^2},$$

assuming the approximate second differential $d^2f(\boldsymbol{x}_n)$ is positive definite.

Algebraic penalties such as $\|g(\boldsymbol{x})\|^2$ reduce to distance penalties with constraint set $\{\boldsymbol{0}\}$. The corresponding projection operator sends any vector $\boldsymbol{y}$ to $\boldsymbol{0}$, and $\|g(\boldsymbol{x})\|^2 = \operatorname{dist}[g(\boldsymbol{x}), \{\boldsymbol{0}\}]^2$. This observation is pertinent to constrained least squares with $g(\boldsymbol{x}) =$

$\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}$ [GV96]. The proximal distance surrogate can be expressed as

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \frac{\rho}{2}\|\boldsymbol{d} - \boldsymbol{C}\boldsymbol{x}\|^2 \;=\; \frac{1}{2}\left\| \begin{pmatrix} \boldsymbol{y} \\ \sqrt{\rho}\boldsymbol{d} \end{pmatrix} - \begin{bmatrix} \boldsymbol{A} \\ \sqrt{\rho}\boldsymbol{C} \end{bmatrix} \boldsymbol{x} \right\|^2$$

and minimized by standard least squares algorithm. No annealing is necessary. Inequality constraints $g(\boldsymbol{x}) \leq \boldsymbol{0}$ behave somewhat differently. The proximal distance majorization $\operatorname{dist}[g(\boldsymbol{x}), \mathbb{R}^m_-]^2 \leq \|g(\boldsymbol{x}) - \mathcal{P}_{\mathbb{R}^m_-}[g(\boldsymbol{x}_n)]\|^2$ is not the same as the Beltrami penalty $g(\boldsymbol{x})_+^2$ [Bel70]. However, the standard majorization [Lan16] $g(\boldsymbol{x})_+^2 \leq \|g(\boldsymbol{x}) - \mathcal{P}_{\mathbb{R}^m_-}[g(\boldsymbol{x}_n)]\|^2$ brings them back into alignment.

ADMM can be motivated by the MM principle. The optimal pair $(\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{\lambda}$ furnishes a stationary point of the Lagrangian. Because the Lagrangian is linear in $\boldsymbol{\lambda}$, its maximum for fixed $(\boldsymbol{x}, \boldsymbol{y})$ is ∞. To correct this defect, we add a viscosity minorization to the Lagrangian. This produces the modified Lagrangian

$$\mathcal{L}_\mu(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}) \;=\; f(\boldsymbol{x}) + g(\boldsymbol{y}) + \mu\boldsymbol{\lambda}^t(\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y}) + \frac{\mu}{2}\|\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y}\|^2 - \frac{\alpha}{2}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_n\|^2.$$

The penalty term has no impact on the $\boldsymbol{x}$ and $\boldsymbol{y}$ updates. However, the MM update for $\boldsymbol{\lambda}$ is determined by the stationarity condition

$$\boldsymbol{0} \;=\; \mu(\boldsymbol{D}\boldsymbol{x}_{n+1} - \boldsymbol{y}_{n+1}) - \alpha(\boldsymbol{\lambda} - \boldsymbol{\lambda}_n),$$

so that

$$\boldsymbol{\lambda}_{n+1} \;=\; \boldsymbol{\lambda}_n + \frac{\mu}{\alpha}(\boldsymbol{D}\boldsymbol{x}_{n+1} - \boldsymbol{y}_{n+1}).$$

The choice $\alpha = 1$ gives the standard ADMM update. Thus, the ADMM algorithm alternates decreasing and increasing the Lagrangian in a search for the saddlepoint represented by the optimal trio $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda})$.

2.6 Proofs

In this section we provide proofs for the convergence results discussed in Section 2.3.

2.6.1 Proposition 2.3.1

Proof: Without loss of generality we can translate the coordinates so that $\boldsymbol{y} = \mathbf{0}$. Let B be the unit sphere $\{\boldsymbol{x} : \|\boldsymbol{x}\| = 1\}$. Our first aim is to show that $h_\rho(\boldsymbol{x}) > f(\mathbf{0})$ throughout B. Consider the set $B \cap T$, which is possibly empty. On this set the infimum b of $f(\boldsymbol{x})$ is attained, so $b > f(\mathbf{0})$ by assumption. The set $B \setminus T$ will be divided into two regions, a narrow zone adjacent to T and the remainder. Now let us show that there exists a $\delta > 0$ such that $h_\rho(\boldsymbol{x}) \geq f(\boldsymbol{x}) \geq f(\mathbf{0}) + \delta$ for all $\boldsymbol{x} \in B$ with $\operatorname{dist}(\boldsymbol{D}\boldsymbol{x}, S) \leq \delta$. If this is not so, then there exists a sequence $\boldsymbol{x}_n \in B$ with $f(\boldsymbol{x}_n) < f(\mathbf{0}) + \frac{1}{n}$ and $\operatorname{dist}(\boldsymbol{D}\boldsymbol{x}_n, S) \leq \frac{1}{n}$. By compactness, some subsequence of $\boldsymbol{x}_n$ converges to $\boldsymbol{z} \in B \cap T$ with $f(\boldsymbol{z}) \leq f(\mathbf{0})$, contradicting the uniqueness of $\boldsymbol{y}$. Finally, let $a = \min_{\boldsymbol{x} \in B} f(\boldsymbol{x})$. To deal with the remaining region take ρ large enough so that $a + \frac{\rho}{2}\delta^2 > f(\mathbf{0})$. For such ρ, $h_\rho(\boldsymbol{x}) > f(\mathbf{0})$ everywhere on B. It follows that on the unit ball $\{\boldsymbol{x} : \|\boldsymbol{x}\| \leq 1\}$, $h_\rho(\boldsymbol{x})$ is minimized at an interior point. Because $h_\rho(\boldsymbol{x})$ is convex, a local minimum is necessarily a global minimum.

To show that the objective $h_\rho(\boldsymbol{x})$ is coercive, it suffices to show that it is coercive along every ray $\{t\boldsymbol{v} : t \geq 0,\ \|\boldsymbol{v}\| = 1\}$ [Lan16]. The convex function $r(t) = h_\rho(t\boldsymbol{v})$ satisfies $r(t) \geq r(1) + r'_+(1)(t-1)$. Because $r(0) < r(1)$, the point 1 is on the upward slope of $r(t)$, and the one-sided derivative $r'_+(1) > 0$. Coerciveness follows from this observation. $\square$

2.6.2 Proposition 2.3.2

Proof: The first assertion follows from the bound $g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n) \geq h_\rho(\boldsymbol{x})$. To prove the second assertion, we note that it suffices prove the existence of some constant $\rho > 0$ such that the matrix $\boldsymbol{A} + \rho\boldsymbol{D}^t\boldsymbol{D}$ is positive definite [Deb52]. If no choice of ρ renders $\boldsymbol{A} + \rho\boldsymbol{D}^t\boldsymbol{D}$ positive definite, then there is a sequence of unit vectors $\boldsymbol{u}_m$ and a sequence of scalars ρ_m tending to ∞ such that

$$\boldsymbol{u}_m^t\boldsymbol{A}\boldsymbol{u}_m + \rho_m\boldsymbol{u}_m^t\boldsymbol{D}^t\boldsymbol{D}\boldsymbol{u}_m \quad \leq \quad 0. \tag{2.7}$$

By passing to a subsequence if needed, we may assume that the sequence $\boldsymbol{u}_m$ converges to a unit vector $\boldsymbol{u}$. On the one hand, because $\boldsymbol{D}^t\boldsymbol{D}$ is positive semidefinite, inequality (2.7) compels the conclusions $\boldsymbol{u}_m^t\boldsymbol{A}\boldsymbol{u}_m \le 0$, which must carry over to the limit. On the other hand, dividing inequality (2.7) by ρ_m and taking limits imply $\boldsymbol{u}^t\boldsymbol{D}^t\boldsymbol{D}\boldsymbol{u} \le 0$ and therefore $\|\boldsymbol{D}\boldsymbol{u}\| = 0$. Because the limit vector $\boldsymbol{u}$ violates the condition $\boldsymbol{u}^t\boldsymbol{A}\boldsymbol{u} > 0$, the required $\rho > 0$ exists. ☐

2.6.3 Proposition 2.3.3

Proof: Systematic decrease of the iterates $h_\rho(\boldsymbol{x}_n)$ is a consequence of the MM principle. The existence of $\boldsymbol{z}_\rho$ follows from Proposition 2.3.1. To prove the stated bound, first observe that the function $g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n) - \frac{\rho}{2}\|\boldsymbol{D}\boldsymbol{x}\|^2$ is convex, being the sum of the convex function $f(\boldsymbol{x})$ and a linear function. Because $\nabla g_\rho(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n)^t(\boldsymbol{x}-\boldsymbol{x}_{n+1}) \ge \boldsymbol{0}$ for any $\boldsymbol{x}$ in C, the supporting hyperplane inequality implies that

$$\begin{aligned} g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n) - \frac{\rho}{2}\|\boldsymbol{D}\boldsymbol{x}\|^2 &\ge g_\rho(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) - \frac{\rho}{2}\|\boldsymbol{D}\boldsymbol{x}_{n+1}\|^2 \\ &\quad -\rho\boldsymbol{x}_{n+1}^t\boldsymbol{D}^t\boldsymbol{D}(\boldsymbol{x}-\boldsymbol{x}_{n+1}), \end{aligned}$$

or equivalently

$$g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_n) \ge g_\rho(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) + \frac{\rho}{2}\|\boldsymbol{D}(\boldsymbol{x}-\boldsymbol{x}_{n+1})\|^2. \tag{2.8}$$

Now note that the difference

$$d(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \mathcal{P}(\boldsymbol{y})\|^2 - \frac{1}{2}\|\boldsymbol{x} - \mathcal{P}(\boldsymbol{x})\|^2$$

has gradient

$$\nabla d(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{P}(\boldsymbol{x}) - \mathcal{P}(\boldsymbol{y}).$$

Because $\mathcal{P}(\boldsymbol{x})$ is non-expansive, the gradient $\nabla d(\boldsymbol{x} \mid \boldsymbol{y})$ is Lipschitz with constant 1. The tangency conditions $d(\boldsymbol{y} \mid \boldsymbol{y}) = 0$ and $\nabla d(\boldsymbol{y} \mid \boldsymbol{y}) = \boldsymbol{0}$ therefore yield

$$\begin{aligned} d(\boldsymbol{x} \mid \boldsymbol{y}) &\leq d(\boldsymbol{y} \mid \boldsymbol{y}) + \nabla d(\boldsymbol{y} \mid \boldsymbol{y})^t(\boldsymbol{x} - \boldsymbol{y}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \\ &= \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \end{aligned} \tag{2.9}$$

for all $\boldsymbol{x}$. At a minimum $\boldsymbol{z}_\rho$ of $h_\rho(\boldsymbol{x})$, combining inequalities (2.8) and (2.9) gives

$$\begin{aligned} & h_\rho(\boldsymbol{x}_{n+1}) + \frac{\rho}{2}\|\boldsymbol{D}(\boldsymbol{z}_\rho - \boldsymbol{x}_{n+1})\|^2 \\ \leq\ & g_\rho(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n) + \frac{\rho}{2}\|\boldsymbol{D}(\boldsymbol{z}_\rho - \boldsymbol{x}_{n+1})\|^2 \\ \leq\ & g_\rho(\boldsymbol{z}_\rho \mid \boldsymbol{x}_n) \\ =\ & h_\rho(\boldsymbol{z}_\rho) - \frac{\rho}{2}\|\boldsymbol{D}\boldsymbol{z}_\rho - \mathcal{P}(\boldsymbol{D}\boldsymbol{z}_\rho)\|^2 + \frac{\rho}{2}\|\boldsymbol{D}\boldsymbol{z}_\rho - \mathcal{P}(\boldsymbol{D}\boldsymbol{x}_n)\|^2 \\ =\ & h_\rho(\boldsymbol{z}_\rho) + \rho d(\boldsymbol{D}\boldsymbol{z}_\rho \mid \boldsymbol{D}\boldsymbol{x}_n) \\ \leq\ & h_\rho(\boldsymbol{z}_\rho) + \frac{\rho}{2}\|\boldsymbol{D}\boldsymbol{z}_\rho - \boldsymbol{D}\boldsymbol{x}_n\|^2. \end{aligned}$$

Adding the result

$$h_\rho(\boldsymbol{x}_{n+1}) - h_\rho(\boldsymbol{z}_\rho) \leq \frac{\rho}{2}\Big[\|\boldsymbol{D}(\boldsymbol{z}_\rho - \boldsymbol{x}_n)\|^2 - \|\boldsymbol{D}(\boldsymbol{z}_\rho - \boldsymbol{x}_{n+1})\|^2\Big]$$

over n and invoking the descent property $h_\rho(\boldsymbol{x}_{n+1}) \leq h_\rho(\boldsymbol{x}_n)$, telescoping produces the desired error bound

$$\begin{aligned} h_\rho(\boldsymbol{x}_{n+1}) - h_\rho(\boldsymbol{z}_\rho) &\leq \frac{\rho}{2(n+1)}\Big[\|\boldsymbol{D}(\boldsymbol{z}_\rho - \boldsymbol{x}_0)\|^2 - \|\boldsymbol{D}(\boldsymbol{z}_\rho - \boldsymbol{x}_{n+1})\|^2\Big] \\ &\leq \frac{\rho}{2(n+1)}\|\boldsymbol{D}(\boldsymbol{z}_\rho - \boldsymbol{x}_0)\|^2. \end{aligned}$$

This is precisely the asserted bound. □

2.6.4 Proposition 2.3.4

Proof: The existence and uniqueness of $\boldsymbol{z}_\rho$ are obvious. The remainder of the proof hinges on the facts that $h_\rho(\boldsymbol{x})$ is μ-strongly convex and the surrogate $g_\rho(\boldsymbol{x} \mid \boldsymbol{w})$ is $L+\rho\|\boldsymbol{D}\|^2$-smooth for all $\boldsymbol{w}$. The latter assertion follows from

$$\nabla g_\rho(\boldsymbol{x} \mid \boldsymbol{w}) - \nabla g_\rho(\boldsymbol{y} \mid \boldsymbol{w}) \;=\; \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) + \rho \boldsymbol{D}^t \boldsymbol{D}(\boldsymbol{x} - \boldsymbol{y}).$$

These facts together with $\nabla g_\rho(\boldsymbol{y} \mid \boldsymbol{y}) = \boldsymbol{0}$ imply

$$\begin{aligned} h_\rho(\boldsymbol{x}) - h_\rho(\boldsymbol{y}) &\le g_\rho(\boldsymbol{x} \mid \boldsymbol{y}) - g_\rho(\boldsymbol{y} \mid \boldsymbol{y}) \\ &\le \nabla g_\rho(\boldsymbol{y} \mid \boldsymbol{y})^t(\boldsymbol{x} - \boldsymbol{y}) + \frac{L + \rho\|\boldsymbol{D}\|^2}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 \qquad (2.10) \\ &= \frac{L + \rho\|\boldsymbol{D}\|^2}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2. \end{aligned}$$

The strong convexity condition

$$0 \;\ge\; h_\rho(\boldsymbol{z}_\rho) - h_\rho(\boldsymbol{x}) \;\ge\; \nabla h_\rho(\boldsymbol{x})^t(\boldsymbol{z}_\rho - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{z}_\rho - \boldsymbol{x}\|^2$$

entails

$$\|\nabla h_\rho(\boldsymbol{x})\| \cdot \|\boldsymbol{z}_\rho - \boldsymbol{x}\| \;\ge\; -\nabla h_\rho(\boldsymbol{x})^t(\boldsymbol{z}_\rho - \boldsymbol{x}) \;\ge\; \frac{\mu}{2}\|\boldsymbol{z}_\rho - \boldsymbol{x}\|^2.$$

It follows that $\|\nabla h_\rho(\boldsymbol{x})\| \ge \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{z}_\rho\|$. This last inequality and inequality (2.10) produce the Polyak-Łojasiewicz bound

$$\frac{1}{2}\|\nabla h_\rho(\boldsymbol{x})\|^2 \;\ge\; \frac{\mu^2}{2(L + \rho\|\boldsymbol{D}\|^2)}[h_\rho(\boldsymbol{x}) - h_\rho(\boldsymbol{z}_\rho)].$$

Taking $c = L + \rho\|\boldsymbol{D}\|^2$ and

$$\boldsymbol{x} = \boldsymbol{x}_k - c^{-1}\nabla g_\rho(\boldsymbol{x}_k \mid \boldsymbol{x}_k) \;=\; \boldsymbol{x}_k - c^{-1}\nabla h_\rho(\boldsymbol{x}_k),$$

the Polyak-Łojasiewicz bound gives

$$\begin{aligned} h_\rho(\boldsymbol{x}_{k+1}) - h_\rho(\boldsymbol{x}_k) &\le g_\rho(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k) - g_\rho(\boldsymbol{x}_k \mid \boldsymbol{x}_k) \\ &\le g_\rho(\boldsymbol{x} \mid \boldsymbol{x}_k) - g_\rho(\boldsymbol{x}_k \mid \boldsymbol{x}_k) \\ &\le -c^{-1}\nabla g_\rho(\boldsymbol{x}_k \mid \boldsymbol{x}_k)^t \nabla h_\rho(\boldsymbol{x}_k) + \frac{c}{2}\|c^{-1}\nabla h_\rho(\boldsymbol{x}_k)\|^2 \\ &= -\frac{1}{2c}\|\nabla h_\rho(\boldsymbol{x}_k)\|^2 \\ &\le -\frac{\mu^2}{2c^2}[h_\rho(\boldsymbol{x}_k) - h_\rho(\boldsymbol{z}_\rho)]. \end{aligned}$$

Rearranging this inequality yields

$$h_\rho(\boldsymbol{x}_{k+1}) - h_\rho(\boldsymbol{z}_\rho) \quad \leq \quad \left[1 - \frac{\mu^2}{2c^2}\right][h_\rho(\boldsymbol{x}_k) - h_\rho(\boldsymbol{z}_\rho))],$$

which can be iterated to give the stated bound. □

CHAPTER 3

Examining School Reopening Strategies During a Pandemic

3.1 Introduction

Reopening K-12 schools is a topic of intense discussion. Because transmission of SARS-CoV-2 occurs through respiratory droplets, reopening policies must adequately reduce crowded environments at school to protect children, teachers, staff, and ultimately communities. Unfortunately, many factors work to the detriment of ostensibly reasonable strategies, including extended hours for teachers, challenges in transporting children to and from school, and reduced quality of educational experience. Although U.S. school closures in March 2020 reduced COVID-19 cases in states with low cumulative incidence, education researchers worry about lagging educational development of children once schools reopen [ASR20, KT20, KST20]. A predictable, regular attendance policy is crucial in balancing social burden with maintaining steady educational progress.

As school systems, professional organizations, and governments have proposed different reopening strategies to reduce infection risks to students, teachers, school staff, and faculty, it is helpful to quantify ramifications of different plans [SSS20]. Here we explore a simple, interpretable mathematical model that compares infection rates under various reopening scenarios. We compare consequences of (1) reopening at full capacity, (2) allowing half of all children to return to in-person schooling while the other half continues with remote learning (parallel cohorts), and (3) alternating sessions in which different student cohorts attend school every other or every third week (rotating cohorts). Our goal is to provide insight into epidemiological consequences of reopening strategies and to quantify their consequences. In particular, we explore implications of the reclosing guidelines announced by Governor Gavin Newsom for California schools [Off20].

3.2 Methods

3.2.1 Compartmental Model

Our approach uses a deterministic Susceptible-Exposed-Infected-Removed (SEIR) model stratified by age group and cohort. We assume that infecteds may or may not present with symptoms and that the removed pool accounts for individuals with negligible contribution to infection spread, including individuals that have either recovered with full immunity or died. Given that natural immunity may persist over several months [SGM20, GWL21, SCB21, DMK21] and that our simulations span a period of 6 months, we make the plausible assumption that individuals do not return to the susceptible pool once infected. For simplicity, the simulation scope is limited to two age groups, children in K-12 education spread over 1 to 3 child cohorts and adults over 18 years. Births are ignored because our simulations operate on relatively short time scales. Although mortality certainly represent an important metric for public health concerns, we do not model deaths explicitly. This simplification avoids introducing additional model parameters. In theory, one might approximate deaths by adjusting our predictions for the number of removed individuals by community-specific estimates for death rates. Model assumptions are further elaborated in our discussion of transmission rates and other model parameters.

In our differential equations model the functions $S(t)$, $E(t)$, $I(t)$, and $R(t)$ denote the fraction of susceptible, exposed, infected, and removed individuals, respectively, in an overall population at time t. Each compartment is stratified by age class (1 for children, 2 for adults) and cohort membership so that I_{1k} refers to infected children in cohort k. With this notation in mind, we propose the following model for the force of infection acting on susceptible individuals in class j and cohort k, denoted $\lambda_{jk}(t)$, as

$$\begin{aligned}
\lambda_{jk}(t) &= \sum_{\text{cohort } \ell} \sum_{\text{age class } i} \begin{pmatrix} \text{interaction} \\ \text{between} \\ \text{cohorts } k, \ell \end{pmatrix} \times \begin{pmatrix} \text{transmission} \\ \text{from age} \\ \text{group } i \to j \end{pmatrix} \times \begin{pmatrix} \text{fraction of} \\ \text{infecteds in} \\ \text{age group } i, \\ \text{and cohort } \ell \end{pmatrix} \\
&= \sum_{\ell} \sum_{i} \alpha_{k\ell} \times \beta_{ij} \times I_{i\ell}(t).
\end{aligned}$$

For pairs of cohorts $k \neq \ell$, the extremes $\alpha_{k\ell} = 0$ and $\alpha_{k\ell} = 1$ reflect complete separation and mixing between two cohorts, respectively. Values in between these limits may be interpreted as decreased interaction due to physical or social distancing. Weak cohort interactions are fixed at $\alpha_{k\ell} = 0.05$ in all of our simulations. The transmission rates β_{ij} may be asymmetric to capture heterogeneity in transmission due to different contact patterns, susceptibility, or infectiousness. Lastly, the parameters σ_j and γ_j for age class j represent rates at which exposed individuals become infectious (latency) and infecteds recover from the contagious stage (infectiousness), respectively. Specifically, we take $1/\gamma_j$ as the average number of days an individual in class j is contagious based on a time-homogeneous Markovian model; an analogous interpretation holds for the latency parameters.

The ordinary differential equations (ODEs) describing cohort k are given by

$$\begin{array}{ll}
\text{Children} & \text{Adults} \\
\dfrac{\mathrm{d}S_{1k}}{\mathrm{d}t} = -\lambda_{1k} S_{1k} & \dfrac{\mathrm{d}S_{2k}}{\mathrm{d}t} = -\lambda_{2k} S_{2k} \\
\dfrac{\mathrm{d}E_{1k}}{\mathrm{d}t} = \lambda_{1k} S_{1k} - \sigma_1 E_{1k} & \dfrac{\mathrm{d}E_{2k}}{\mathrm{d}t} = \lambda_{2k} S_{2k} - \sigma_2 E_{2k} \\
\dfrac{\mathrm{d}I_{1k}}{\mathrm{d}t} = \sigma_1 E_{1k} - \gamma_1 I_{1k} & \dfrac{\mathrm{d}I_{2k}}{\mathrm{d}t} = \sigma_2 I_{2k} - \gamma_2 I_{2k} \\
\dfrac{\mathrm{d}R_{1k}}{\mathrm{d}t} = \gamma_1 I_{1k} & \dfrac{\mathrm{d}R_{2k}}{\mathrm{d}t} = \gamma_2 I_{2k},
\end{array}$$

where the left and right columns correspond to SEIR compartments for children and adults, respectively. Fig. 3.1 summarizes the high level features of our mathematical model. All numerical simulations are carried out in the Julia programming language using tools from the SciML ecosystem [RN17, RN19, Sci20a, Sci20b].

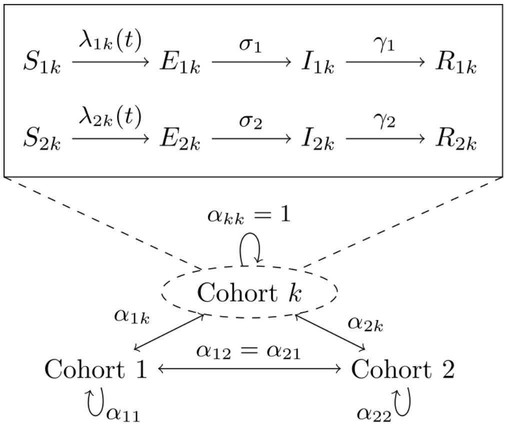


Figure 3.1: **Overview of SEIR compartmental model.** The main compartments are denoted by $S(t)$, $E(t)$, $I(t)$, and $R(t)$ for susceptible, exposed, infected, and removed, respectively. Compartments are stratified by age class (1 – children, 2 – adults) and membership to cohort k. The coefficients $\alpha_{k\ell} \in [0, 1]$ account for the strength of interaction between cohorts k and ℓ.

3.2.2 Basic Reproductive Number

We characterize the basic reproductive number $\mathcal{R}_0$ indicative of growth potential of an infectious disease. Specifically, $\mathcal{R}_0$ quantifies the expected number of secondary infections due to a single infected within a completely susceptible population. The threshold $\mathcal{R}_0$ value of 1 marks the boundary between explosive growth ($\mathcal{R}_0 > 1$) and decline of an epidemic to extinction ($\mathcal{R}_0 < 1$). We derive $\mathcal{R}_0$ for our stratified SEIR model using the next generation method outlined by Diekmann, Heesterbeek, and Roberts [DHR10]. Near a disease-free equilibrium point, it is reasonable to linearize dynamics by taking the initial proportion of susceptibles, $S_{jk}(0)$, approximately equal to its maximal value defined by the population's demography, q_{jk}. Thus, the transmission and transition operators $\boldsymbol{T}$ and $\boldsymbol{\Sigma}$ are given by the matrices

$$\boldsymbol{T} = \begin{bmatrix} 0 & \alpha_{11}\beta_{11}q_{11} & 0 & \alpha_{11}\beta_{12}q_{11} \\ 0 & & 0 & 0 \\ 0 & \alpha_{11}\beta_{21}q_{21} & 0 & \alpha_{11}\beta_{22}q_{21} \\ 0 & & 0 & 0 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} -\sigma_1 & 0 & 0 & 0 \\ \sigma_1 & -\gamma_1 & 0 & 0 \\ 0 & 0 & -\sigma_2 & 0 \\ 0 & 0 & \sigma_2 & -\gamma_2 \end{bmatrix},$$

based on the infectious subsystem defined by $\boldsymbol{x} = [E_{11}, I_{11}, E_{21}, I_{21}]^\top$ for a single cohort. Together, these linear operators recover the linearized subsystem $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = (\boldsymbol{T} + \boldsymbol{\Sigma})\boldsymbol{x}$. Following Diekmann et al. [DHR10], $\mathcal{R}_0$ is taken as the spectral radius of the next generation matrix with large domain, $-\boldsymbol{T}\boldsymbol{\Sigma}^{-1}$. In the case of multiple cohorts, the structures of $\boldsymbol{T}$ and $\boldsymbol{\Sigma}$ as given are repeated in a tiled fashion, with the appropriate changes in indices for $\alpha_{k\ell}$ and q_{jk}.

3.2.3 Simulation Studies on Prevalence Thresholds

We consider the effect of a stopping rule on cumulative prevalence. Inspired by California's guidelines urging schools to revert to remote learning whenever the infections within a school reach 5% in 2-week period (3), we define the stopping time $t_{\text{threshold}}$ as the first time that detected school cases reach the specified threshold. Formally, the stopping time is given by

$$t_{\text{threshold}} = \inf\left\{ t \in T : \sum_{s \in \text{window}(t)} C(s) \geq 5\% \right\},$$

where T is a set of testing times and the sum is taken over a sliding 14-day window up to time t. The quantity $C(s) = \text{sensitivity} \times I_{1k}(s)/q$ represents detected cases adjusted for population size q. Detection necessarily depends on a particular test's sensitivity and is based on testing at the beginning of a school day, after which infected individuals in the active cohort are immediately isolated and placed in the removed state $(I(t) \to R(t))$. The isolation rule applies only to the cohort at school, while the sensitivity factor in the rule captures imprecision in testing and reporting.

Note that $t_{\text{threshold}} = \infty$ if the threshold is never reached over the time span of our simulations. Furthermore, our simulation results involving $t_{\text{threshold}}$ represent a lower bound because case isolation is taken to be instantaneous. In practice, segregation of affected pupils is delayed. Our model does not explicitly account for adult staff at school. Our simplifying assumption is justified by our focus on qualitative behavior and the fact that students typically outnumber teachers and ancillary staff. For example, the average class sizes for public elementary and public secondary schools are estimated to be 21.2 and 26.8 students, respectively, for the 2011–2012 academic year [Nat19]. At a 20:1 student to staff ratio, a school with 1000 students would need 53 cases in a 14-day period to meet the closure criterion of 50 set in our simulations.

3.2.4 Modelling Transmission Between Age Classes

In spite of less severe disease and lower case-fatality rates than adults, children may be just as prone to SARS-CoV-2 infections as adults [ZC20]. Children's symptoms range from fever, rhinitis, cough, and GI symptoms, to a Kawasaki-like disease termed Multisystem Inflammatory Syndrome in Children (MIS-C) [VMG20, Cen20, DKC20]. However, because children's symptoms are typically less severe and of shorter duration than those of adults [Lud20], the likelihood of pediatric infection escaping symptom-based monitoring, such as temperature screening, is higher than that of adults. This reality increases pre-symptomatic and asymptomatic transmission [PZW20, JMV20]. Thus, detecting transmission in children is difficult; quantifying it is all the more challenging [Lop20, PCP20, LPC20].

Contact tracing data from Singapore suggest that per contact transmission between children, particularly in educational settings, is low compared with adult-adult transmission [YKN20]. Yet the number of contacts between children is expected to be significantly higher compared to other age groups[ZLL20, Sza20]. Changes in contact structure will necessarily change estimates of transmission rates. For example, Li et al. provide transmission rate estimates for Wuhan prior to (1.12 per day) and following travel restrictions (0.52 per day) [LPC20]. An additional source of heterogeneity in transmission is the potentially reduced susceptibility of children compared to adults [DKL20]. The review by Viner et al.

summarizes much of the early literature on this topic and suggests that the susceptibility of adolescents may be similar to that of adults [VMB21]. Infectiousness of different age groups is not as well characterized. Each source of heterogeneity poses a challenge to developing a parsimonious mathematical model.

Rather than reconciling transmission rate estimates across populations based on different scientific models, we vary transmission rates between and within age classes to underscore the influence of modelling assumptions on epidemiological consequences and to calibrate the range of effects given existing evidence. To this end, our transmission rates β_{ij} are designed to separate the magnitude of transmission from the effects of different age class interactions. Scale is determined by baseline transmission rate, β_0 and is interpreted as a characteristic of a population. The baseline transmission rate is then used to define each β_{ij} based on the formula $\beta_{ij} = \beta_0 \times f_{ij}$ with weights $f_{ij} \in [0, 1]$ capturing the contribution of each $i \to j$ interaction to the aggregate transmission rate β_0. For our model with only 2 age classes, imposing the constraint

$$f_{11} + f_{12} + f_{21} + f_{22} = 1,$$

allows us to explore the effect of transmissibility assumptions at a fixed scale while retaining the complexity of contact matrices, susceptibility, and infectiousness.

There are a few special cases to point out. The case

$$f_{11} = f_{12} = f_{21} = f_{22} = 1/4,$$

assumes that child-child, child-adult, adult-child, and adult-adult interactions are indistinguishable and therefore that the two age groups are equivalent on the basis of transmission. In the absence of cohort structures, the force of infection on class j becomes

$$\lambda_j = \beta_0 \sum_i f_{ij} I_i = \beta_0 I,$$

effectively collapsing our model to the basic SEIR equations under which β_0 is the transmission rate of a homogeneous population. The case $f_{11} < f_{22}$ reflects lower susceptibility in children compared to adults, differences in contact structure, or both. Assuming the nature of contacts between groups is symmetric, the scenario $f_{12} > f_{21}$ suggests that child-adult interactions contribute more to transmission than adult-child interactions due to differences in susceptibility or infectiousness. Our definition of transmission rates sidesteps the complexity in modelling SARS-CoV-2 transmission and affords our model greater flexibility and interpretability at the expense of parameter identifiability.

3.2.5 Cohort Structure and Increased Child-Child Contact

A well-timed cyclic attendance strategy tuned to the latent period of SARS-CoV-2 may curtail secondary infections [KBM20]. Assuming a latent period of 3-4 days, a weekly rotation schedule synchronizes with peak infectiousness. To compare with full-time and online-only instruction, we investigate consequences in reopening at 50% and 33% capacity with rotating cohorts. Our simulations therefore model transmission between children using period rates that cycle between high and low contact values. Namely, we take $t \mapsto c \times \beta_{11}$ on school days and $t \mapsto \beta_{11}$ otherwise, where c is a multiplier reflecting increased contacts in children. This function is phased between cohorts to reflect school rotation. In summary, children in rotating cohorts attend school for 5 consecutive days and then rotate with the next cohort at the beginning of the following week. With two cohorts children attend school every other week; for three cohorts they attend every third week (rotating cohort strategy). A trend in the U.S. is to allow families to opt for remote learning in lieu of in-person instruction during the SARS-CoV-2 pandemic. We model this situation by dividing our virtual school community into two cohorts of equal size, one which attends school and thus experiences an elevated transmission rate while a second group opts for a remote learning option (parallel cohort strategy).

3.2.6 Choices for Other Model Parameters

In contrast to factors contributing to transmission rates, latent, infectious, and incubation periods for SARS-CoV-2 are better characterized in the literature. Lauer et al. estimate a median incubation period of approximately 5 days [LGB20]. Li et al. infer latency and infectious periods of 3.69 and 3.47 days, respectively [LPC20]. The review by Bar-On reports median latent and infectious periods of 3 and 4 days, respectively [BFP20]. Other studies report serial intervals and incubation periods consistent with these estimates for latency and infectiousness [DXW20, HLW20]. Unfortunately, the literature on similar epidemiological inferences in children is sparse. One observational study suggests children may have incubation periods similar to those of adults [ZBH20].

Because our simulations model school reopening, the proportion of infected individuals will influence prevalence and especially time to school closures. A periodic joint report from the American Academy of Pediatrics and Children's Hospital Association indicates that children account for 12.9% (range: 8%-20%) of COVID-19 cases across US states and territories as of February 4, 2021 [Ame21].

Table 3.1 summarizes our choices and lists references pertinent to each choice, where applicable.

3.3 Results

3.3.1 Cohorts Reduce Basic Reproduction Number Under Various Transmission Modalities

We first examine the impact of separating children into rotating cohorts on the basic reproduction number $\mathcal{R}_0$ of the model. Unfortunately, this quantity necessarily depends on poorly characterized transmission rates and varies with different contact patterns and human behaviors. Thus, we use our parameterization $\beta_{ij} = \beta_0 f_{ij}$ to identify dominant terms f_{ij} con-

Parameter	Description	Range/Estimate
β_0	Bulk transmission rate for population	1.2 day^{-1}, 1.5 day^{-1}
f_{11}	Weight for child to child transmission.	0–1; 0.1
f_{12}	Weight for child to adult transmission.	0–1; 0.25
f_{21}	Weight for adult to child transmission.	0–1; 0.15
f_{22}	Weight for adult to adult transmission.	0–1; 0.5
$1/\sigma_1$	Average child latency period.	3 days
$1/\sigma_2$	Average adult latency period.	3 days
$1/\gamma_1$	Average child infectious period.	4 days
$1/\gamma_2$	Average adult infectious period.	4 days
α_{kk}	Strength of interactions within a cohort k.	1
$\alpha_{k\ell}$	Strength of interactions between cohorts k and ℓ.	0.05
$I(0)$	Proportion of infecteds at reopening (incidence).	0 – 10%
I_{1k}	Proportion of infected children at reopening.	0 – 10%
c	Multiplier modelling increased child-child contact.	1,2,10

Table 3.1: **Summary of model parameters with ranges and estimates.** The range for transmission between adults suggested by Li et al. [LPC20] calibrates the bulk rate. Latency and infectious period estimates are based on Li et al. and the summary by Bar-On et al. [LPC20, BFP20]. The initial proportion of infected individuals is equally distributed across cohorts.

tributing to $\mathcal{R}_0$ under varying cohort numbers but with β_0 fixed. In particular, we consider 3 interesting cases: (1) adult-child and child-adult transmission are symmetric, (2) child-child transmission is weak, and (3) adult-adult transmission is weak. Fig. 3.2 summarizes the results of our analysis. In each of these cases, we find that splitting a school community into 2 or 3 rotating cohorts substantially reduces R_0 under a wide range of parameter values. For example, in the regime with symmetric between-class transmission and weak child-child transmission, moving from full capacity to 2 cohorts shifts $\mathcal{R}_0$ from about 3 to about 1.5 (Fig. 3.2A–B, right corners). Moving further to 3 cohorts brings the reproduction number below 1 in the same regime (Fig. 3.2C). Relaxing the symmetry assumption, we find the pattern recapitulated under both assumptions of weak child-child transmission (Fig. 3.2D–F) and weak adult-adult transmission (Fig. 3.2G–I). Further, this analysis suggests that child-adult and adult-adult transmission can play dominant roles in the short-term dynamics of our model under the plausible scenario of weak child-child transmission (Fig. 3.2D, right and top corners). The influence of β_{22} is not surprising because our virtual population's demography is skewed toward adults (78%). However, our results demonstrate that child-adult transmission should be weighed carefully in reopening decisions because it less characterized and poses a potent risk, especially to school teachers. We focus our attention on the asymmetric case with weak child-child transmission for the remainder of the study. Specifically, we set $f_{11} = 0.1$, $f_{12} = 0.25$, $f_{21} = 0.15$, and $f_{22} = 0.5$ to model this scenario, and take $\beta_0 = 1.2$ to simulate under $R_0 \approx 3$. This choice does not reflect a belief about conditions of the pandemic in any particular population; it is merely intended to demonstrate effects of mitigation strategies within our modelling framework.

3.3.2 Reopening Under Prevalence-Informed Criteria

Identifying conditions under which schools can be safely reopened is paramount to proposing public health policy for containing the epidemic. In particular, reopening schools only to quickly close down after a few days of instruction is costly both in resources and its negative health effect. Here we investigate the influence of initial conditions and elevated child-child transmission on the stopping time $t_{\text{threshold}}$ under an ideal scenario with a 100% sensitive test. Fig. 3.3 reports values for $t_{\text{threshold}}$ after varying child-child transmission in active school cohorts by a factor of $c = 1$, $c = 2$, and $c = 10$. Reopening schools under high infection burden leads to smaller values for $t_{\text{threshold}}$, as expected. Interestingly, these results suggest that multiple cohorts have a desired effect of delaying school closures beyond the time span of 26 weeks (6 months) in our simulations. For example, assuming 0.1% prevalence at reopening leads to school closure after 6–7 weeks under a single cohort whereas multiple cohorts or the hybrid approach have $t_{\text{threshold}} > 26$ weeks. The behavior of the stopping time is insensitive to the contact multiplier c. However, there is a sharp transition from $t_{\text{threshold}} > 26$ weeks to $t_{\text{threshold}} \approx 4$ weeks under multiple cohorts as prevalence at reopening increases.

Next, we investigate the influence of test sensitivity in our simulated monitoring program and closure criteria on period prevalence, taken as the sum of infecteds and removed individuals. We compare predictions of our model over 26 weeks (6 months) when (1) no action is taken (Fig. 3.4A–B), (2) the monitoring program employs a perfectly sensitive test without delays in reporting (Fig. 3.4C–D), and (3) the monitoring program employs a rapid but less sensitive test (Fig. 3.4E–F). Our simulations with a single cohort indicate that a 5% percent threshold policy can shift period prevalence in children from 55% to 45% over the simulated 26-week period (Fig. 3.4A–C). Compared to this ideal scenario, an imperfect test with 50% detection leads to a slightly later stopping time owing to infections spread by undetected cases and greater overall pediatric infections (Fig. 3.4E). The effect is less pronounced in the adult population due to high adult-adult transmission. Crucially, reopening with a surveillance program may provide approximately 2 weeks of continuous instruction. In our model, infections after closing are driven by a lack of interventions outside of school; testing and isolation in this context can curtail this growth. Our results support the importance of testing and complete school closure in preventing a major disease outbreak after reopening.

We repeat the same simulation study with the hybrid parallel cohort policy. Fig. 3.5 reports the same indices recorded under the same parameter values as in the single cohort policy. Reducing the force of infection through the community's contact network successfully decreases period prevalence, sustained contact between children notwithstanding (Fig. 3.5A–D). The stopping rule for the in-person cohort is not triggered even when detection is imperfect (Fig. 3.5E–F). Infections are generally higher in the in-person cohort compared to the remote cohort for both children and adults.

3.3.3 Mitigating Transmission Between Children

Although face masks can reduce the spread of SARS-CoV-2 by 40% in adults [MKR20], risks of mask wearing by elementary school children include impaired learning, speech development, social development, and facial recognition [Ame20a, Ame20b]. It is also unclear whether children can consistently wear masks. An October 2020 survey of middle school and high school students, communicated by the CDC, underscores this point with mask wearing varying from approximately 65% in classrooms and hallways to 25% in outdoor settings within school boundaries [Cen21].

We explore the impacts of varying degrees of protection conferred by combined risk reduction strategies, such as mask wearing, desk shields, handwashing, vigilant surface cleaning, improved ventilation, and outdoor instruction. Combined impacts of these strategies are modeled as 20%, 40%, 60%, and 80% reductions in the transmission rates β_{11} and β_{22} relative to reference values. Specifically, we take $\beta_{11} = 0.12$, $\beta_{12} = 0.3$, $\beta_{21} = 0.18$, and $\beta_{22} = 0.6$ as natural rates and apply a 40% reduction factor to adults by setting $\beta_{21} = 0.072$ and $\beta_{22} = 0.24$. This implies $R_0 \approx 1.7$ prior to reopening. Increased contact is modeled by taking $c = 10$ so that $\beta_{11} = 1.2$, which corresponds to $R_0 \approx 2.2$ under the full capacity reopening scenario. This represents an extreme that illustrates effects in a poor situation.

Fig. 3.6 compares prevalence trajectories for interventions directly targeting transmission under a single or two rotating child cohorts. With a single cohort and no mitigation in children, our choices lead to approximately 8%, 24%, and 28% infected children after 4, 13, and 26 weeks following reopening, respectively (Fig. 3.6A, blue line). However, with measures that lead to an 80% reduction in transmission, infections at 4, 13, and 26 weeks are 5%, 11%, and 13%, respectively (Fig. 3.6A, purple line). Targeting transmission rates in children also reduces infections in adults to a similar degree (Fig. 3.6B). Much stricter adherence to transmission mitigation measures is required for low infection levels when there is a single cohort (Fig. 3.6A–B) than when there are two cohorts (Fig. 3.6C–D). A combination of both types of interventions ultimately results in even fewer infections.

3.4 Concluding Remarks

Our analysis identifies child-adult transmission as a potential risk to reopening schools even under the plausible assumption of weak child-child transmission relative to adult-adult transmission (Fig 3.2D–F). Moreover, our simulation studies highlight the profound impact of reducing cohort size with parallel or rotating cohorts under a range of transmission rates and reproduction numbers. For example, during a 6-month time span, reopening schools in a population with 0.1% infections with 2 cohorts avoids triggering a prevalence closure decision rule based on a 5% pediatric infection threshold. This, allows schools to stay open longer compared to reopening at 100% capacity without cohort separation (Fig. 3.3). Simultaneous adherence to transmission mitigation measures and multiple separated cohorts can keep cases low, for example under 3% (Fig. 3.6C–D). Our work also underscores the importance of tracking infections and setting a threshold for reverting to remote learning. In the absence of any intervention to in-person instruction, the proportion of school safe from infection stays just above 40% at equilibrium (Fig. 3.4B, blue). This compares with keeping the susceptible proportion above 60% under the combination of a rapid testing program, a stopping rule, and a single cohort (Fig. 3.4B, green and orange).

There are several limitations to our modelling that could be addressed in future studies. Finer age stratification is required to predict outcomes in specific communities and can be implemented within our modelling framework. For example, high school students may wear masks and practice physical distancing more reliably than elementary school children, and may also have transmission rates closer to those of adults [VMB21]. Second, we assume equal transmission rates among all adults and omit explicit interactions between students and teachers within a classroom, which are critical in implementing backup protocols that allow switches to remote learning. Network-based models are better suited to accounting for classroom and household structures in a population, as well as shifting contact patterns [BAX20, Bri20, JKY19] Third, our model treats school communities in isolation. Schools in urban settings have diverse commuting patterns and face potential for importing cases from outside adjacent neighborhoods. Fourth, our conclusions about reproduction numbers, period prevalence, stopping times, and impact of various mitigation strategies should be understood as offering policy guidance rather than precise quantitative predictions. Our ODEs are suited to fitting prevalence data rather than incidence data which poses a challenge to predictive capabilities. Lastly, our models omit the stochastic nature of infections in small populations. Although these caveats limit the quantitative accuracy of our predictions, we contend that our qualitative conclusions are correct.

We find that measures reducing class density by rotating cohorts between in-person and remote schooling are likely to have greater impact in reducing the spread of SARS-CoV-2 than policies such as mask wearing, handwashing, and physical distancing in the classroom. Nevertheless, the latter policies combined with a reduction in class density are still quite effective in reducing effective transmission. From the perspective of mathematical epidemiology, this is to be expected as separating a contact graph into disconnected pieces ultimately limits the proliferative potential of an infectious disease. Surprisingly, parallel cohorts are as effective as rotating cohorts in case reduction, while requiring less coordination and work schedule adjustment for parents. Educating children under either cohort strategy should be a priority in school re-openings. Benefits of switching to remote learning when infections climb to an unacceptable level benefit from rapid testing, even if imperfect. Our rapid testing predictions are consistent with a recent study [LWL20] on the influence of viral kinetics, test sensitivity, test frequency, and sample-to-answer reporting time in surveillance protocols, which also demonstrates that test efficacy is a secondary concern given the dangers of the pandemic.

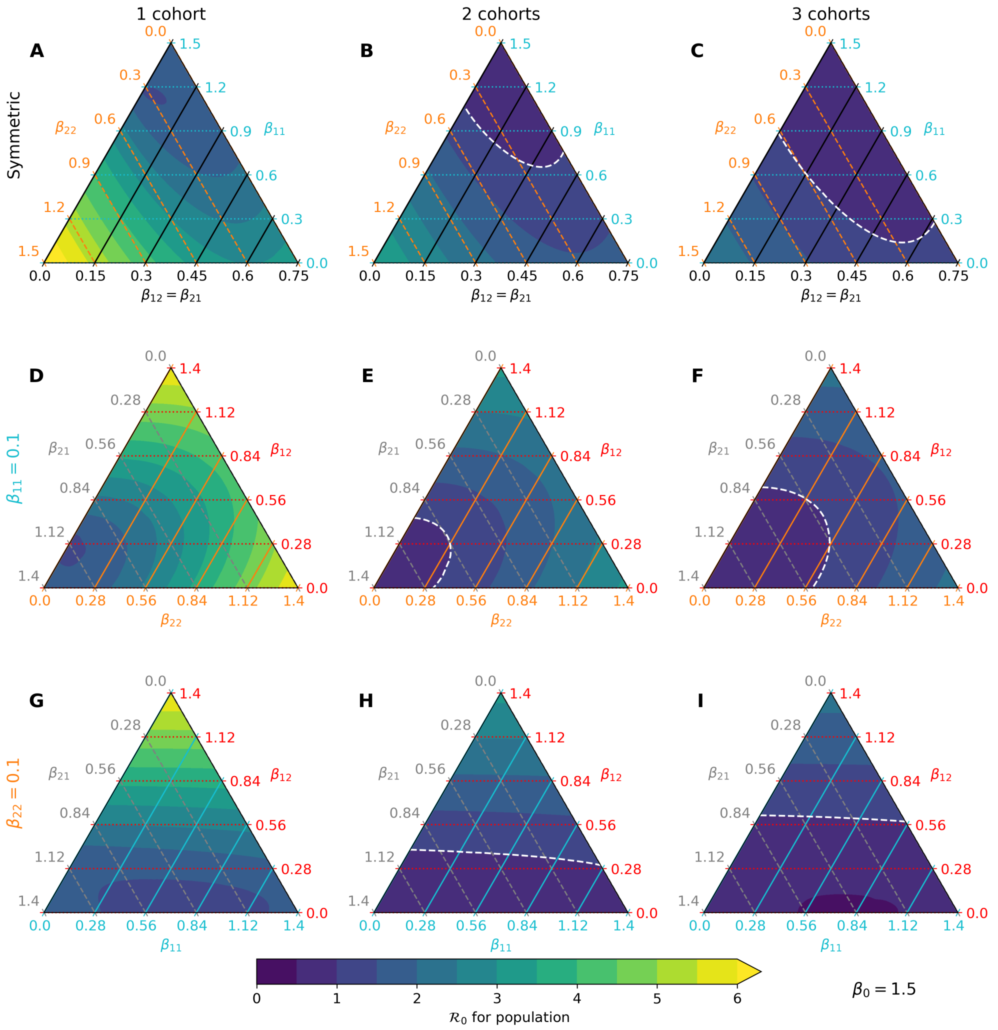


Figure 3.2: **Predicted $\mathcal{R}_0$ under various transmission-cohort scenarios.** The color gradient changes from purple to blue to reflect R_0 shifting from < 1 to > 1 in each ternary plot, with the white line denoting the boundary. Yellow is used to represent $R_0 > 6$. (A-C) Assuming child-adult and adult-child transmission rates are identical (black axis), movement along the blue axis indicates that child-child transmission has a weak effect on R_0 at a fixed scale for β_0. (D-F) Fixing child-child transmission to be weak ($\beta_{11} = 0.1$) relative to other interactions, both child-adult and adult-adult transmission play dominant roles in increasing R_0. (G-I) Fixing adult-adult transmission to be weak ($\beta_{22} = 0.1$), only child-adult transmission plays a dominant role in increasing R_0.

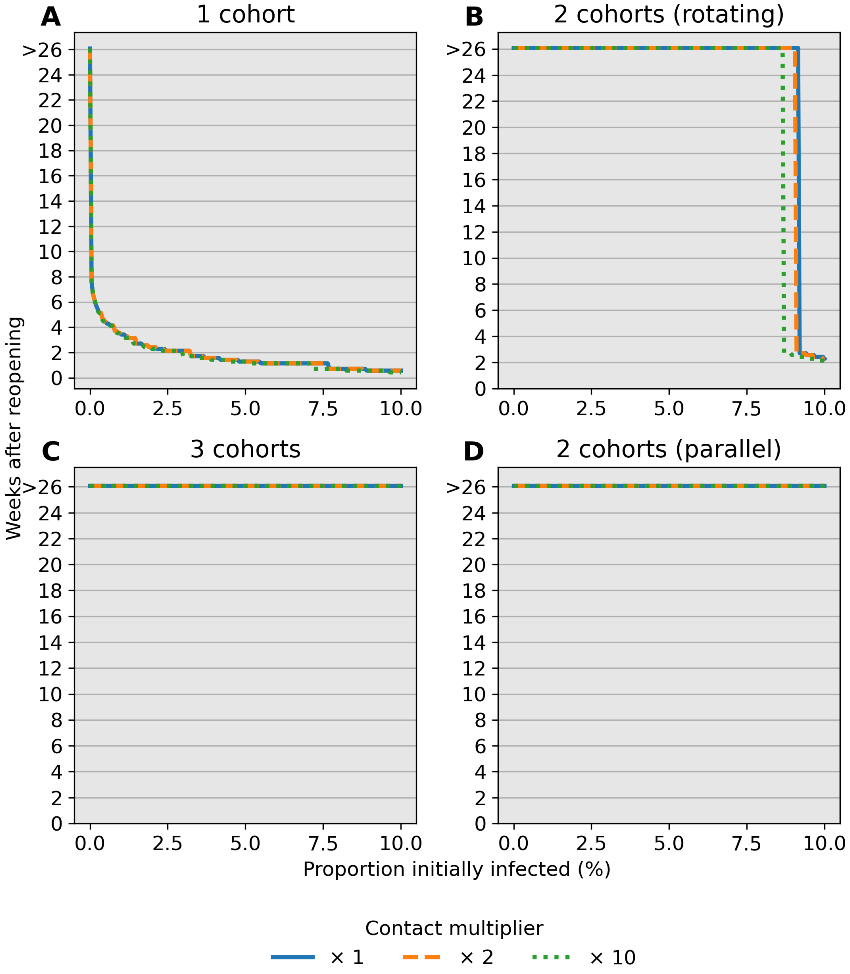


Figure 3.3: **Number of weeks to reach a 5% stopping threshold in a community.** Each scenario assumes a 100% sensitive test. The stopping time $t_{\text{threshold}}$ (y-axis) is simulated under varying prevalence conditions at reopening (x-axis). The contact multiplier for child-child transmission is also varied from (A) $c = 1$ to (B) $c = 2$ and (C) $c = 10$ and has little influence on stopping times. Multiple cohorts are effective at prolonging school operations while staying below a 5% prevalence threshold over a 14-day window. Note that only detected cases in children contribute to the decision rule.

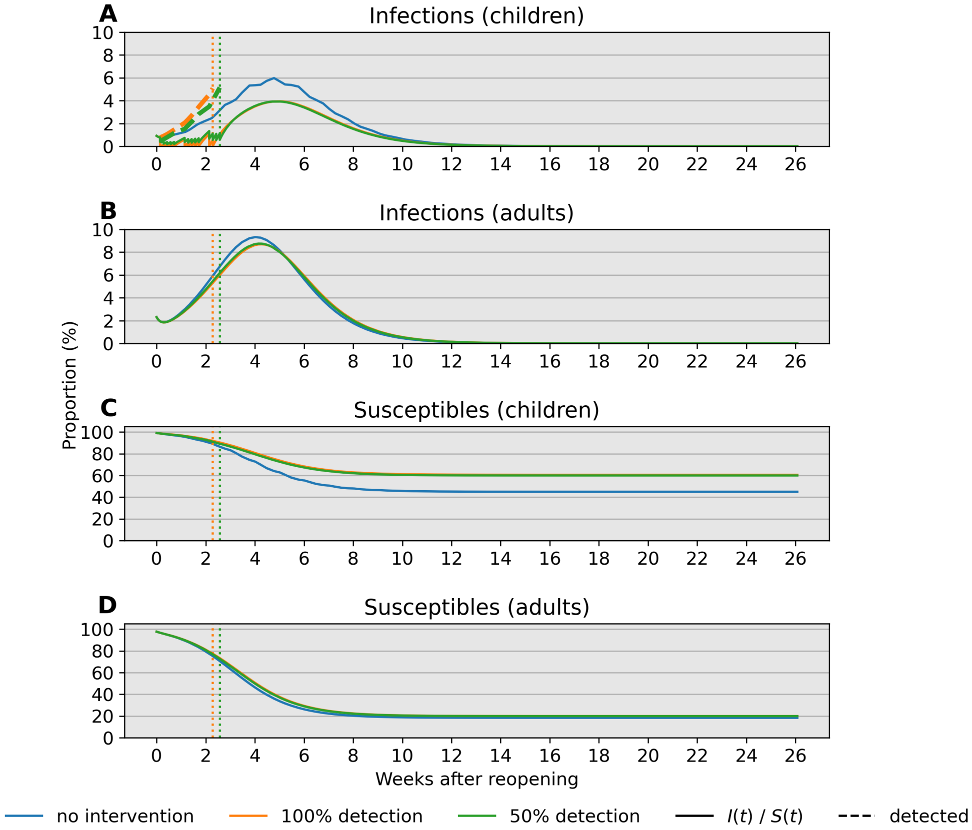


Figure 3.4: **Comparison of infections and susceptibles under different test sensitivities in both children and adults.** Simulations are based on parameter values $f_{11} = 0.1$, $f_{12} = 0.25$, $f_{21} = 0.15$, and $f_{22} = 0.5$ with bulk transmission rate $\beta_0 = 1.2$. Reopening takes place at a 2% prevalence level (2000 infections per 100,000). The decision criterion over a 14-day sliding window is highlighted in a dotted line. Blue, orange, and green lines correspond to scenarios without intervention, with a 100% sensitive test, and a 50% sensitive test, respectively. (A) The 14-day prevalence criteria hits the 5% threshold after just over 2 weeks in the two testing scenarios. (B) Prevalence in adults peaks after about 4 weeks independent of test sensitivity in children. (C) Testing is effective in keeping most children safe from infection regardless of test sensitivity. (D) Testing in children has little impact on keeping adults free from infection under these conditions.

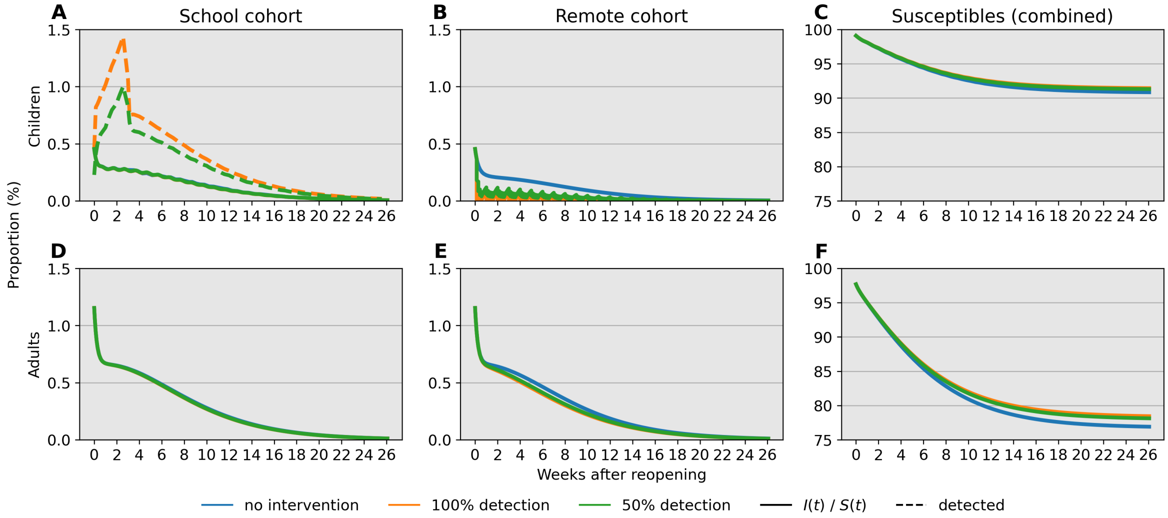


Figure 3.5: **Comparison of cumulative under the parallel cohort approach.** (A-C) The 14-day prevalence criteria increases over the first 4 weeks, but point prevalence consistently trends downward due to cohort structure. Over 90% of children are kept safe from infection under the conditions of this simulation. (D-F) The combination of testing in children and cohort separation prevents a high level of infection in adults.

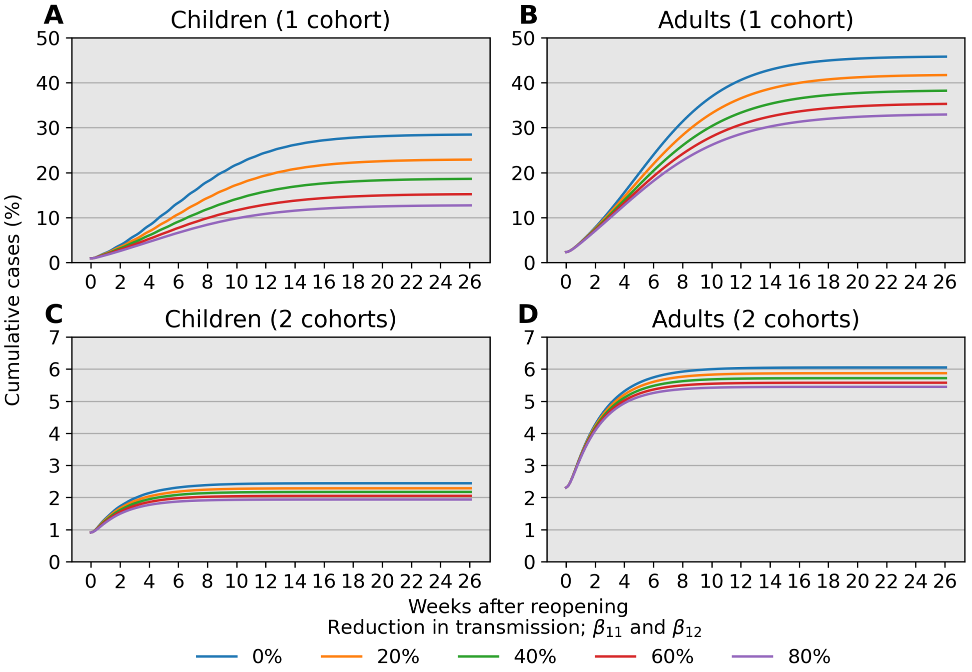


Figure 3.6: **Cumulative prevalence trajectories under risk reduction strategies for children while at school.** For child-child transmission, we set $\beta_{11} = 0.1 \times (1-r)$ outside of school and $\beta_{11} = (1-r) \times c \times 0.1$ during school, where r is a reduction factor due to effective risk reduction strategies and $c = 10$ accounts for increased contact between children. (A–B) Mitigation that reduces transmission between children can lead to a substantial reduction in infections for both children and adults, provided the mitigation effects are large. (C–D) The impact of risk reduction strategies persists when children are separated into 2 rotating cohorts but does not demand as strict an adherence to be effective. An 80% reduction in pediatric transmission has a weaker effect compared to separating children into 2 rotating cohorts as the latter strategy result in fewer than 5% pediatric infection over 26 weeks (6 months).

CHAPTER 4

Improving Hierarchical Linear Models via the Proximal Distance Method and Hard Thresholding

4.1 Introduction

Regression with hierarchical linear models, also known as multilevel or mixed models, generalizes the notion of regression to nested data. A canonical example involves test scores of students across different classrooms. Characteristics about both individual students and classrooms may influence test scores. Given that students are naturally grouped into classrooms, one might reasonably expect classroom effects, such as a teacher's experience level, to influence effects at the level of students, such as student participation, which in turn affect test performance. In the realm of genetics, there is interest in integrating sequencing data, methylation patterns, transcriptomic expression, and other functional annotations to better understand phenotypic variation in complex traits [SZH13]. Pharmacodynamics and pharmacokinetics modelling consider variable drug absorption and biological responses in study participants. Applications of hierarchical linear models abound in the biological sciences due to the natural occurrence of multiple scales.

Before delving into our methodology, let us first review the theory for multiple regression for a continuous response as a motivating example. Denote by $y_i \sim \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i$ a simple linear model. In this setting one fits the p model parameters $(\beta_1, \beta_2, \dots \beta_p)^\top$ to capture the effects of each predictor $(x_{1i}, x_{2i}, \dots, x_{pi})^\top$ on the response y_i for the particular sample i. Under the appropriate regularity conditions and additional assumptions that $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2$, the best linear unbiased estimator for $\boldsymbol{\beta}$ is obtained by considering n samples and solving a least-squares problem, $\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$, yielding the solution

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}.$$

This solution exists only when the $n \times p$ data matrix $\boldsymbol{X}$ has full column rank. One remedy to this problem is to consider the Moore-Penrose pseudoinverse, $(\boldsymbol{X}^\top \boldsymbol{X})^+$, which guarantees an ℓ_2-minimal solution. In fact, $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is the pseudoinverse $\boldsymbol{X}^+$. The solution generated by this approach is equivalent to the solution produced by solving the ridge regression problem,

$$\frac{1}{2}\left[\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\beta}\|^2\right]$$

in the limit $\lambda \to 0$.

In practice, large data sets may have far more putative predictors than samples so that $n \ll p$. Two issues may arise in this setting. First, computing a pseudoinverse requires computing a singular value decomposition of $\boldsymbol{X}$, which is typically more expensive than a QR decomposition used to solve $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$. Second, it may be desirable to obtain sparse solutions for $\boldsymbol{\beta}$ to recover a more parsimonious model. Specifically, a model with few nonzero components β_j will better generalize to out-of-sample predictions, have nicer statistical properties, and reflect a stronger scientific model overall.

Unlike ridge regression [HK70], the general framework of Tikhonov regularization [TA77] can penalize individual parameters differently to emphasize a particular structure in solutions by solving

$$\min_{\boldsymbol{\beta}} \; \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \|\boldsymbol{D}\boldsymbol{\beta}\|_2^2,$$

with a suitably chosen Tikhonov matrix $\boldsymbol{D}$ (ridge regression takes $\boldsymbol{D} = \lambda \boldsymbol{I}$). In practice, it is not always clear what structure to choose for $\boldsymbol{D}$. For example, to perform subset selection one would need to choose the Tikhonov matrix in a manner that emphasizes a particular sparsity pattern in solutions, but the sparsity pattern itself is an unknown property of one's model. Moreover, it is known that the special case of ridge regression will necessarily introduce shrinkage in solution and therefore dilute estimates for effect sizes. The lasso developed by Tibshirani [Tib96, TSR05b]addresses these problems by imposing an ℓ_1 penalty; that is, one solves

$$\min_{\boldsymbol{\beta}} \ \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \|\boldsymbol{D}\boldsymbol{\beta}\|_1.$$

In the basic case with $\boldsymbol{D} = \lambda \boldsymbol{I}$, one controls the degree of regularization on the magnitude of each effect by the free parameter λ. Namely, increasing λ sends more model parameters towards 0. Efficient algorithms for this problem and some of its variants exist. While effective, one unfortunate drawback of the lasso method is that it enforces sparsity indirectly. More critically, taking λ too small will drive up the false positive rate as in ridge regression whereas taking λ too large will introduce false negatives. Cross-validation is a useful technique to avoid this issue.

Our review thus far motivates the need for regularization methods from a mostly theoretical perspective. The interesting case of the lasso makes a connection to practice by addressing the important issue of parsimony in building models. However, the lasso as originally developed mostly applies either in cases of complete ignorance about the effects of predictors or when there is no clear consensus about which predictors are most meaningful. It is at this point that we make our return to praxis and motivate hierarchical linear models in this context.

Specifically, we consider a situation in which an additional set of data provides information about the original predictors, thus establishing a hierarchy in the set of independent variables in the regression. One example is in human genetics in which one studies a particular trait by integrating information from Single Nucleotide Polymorphisms (SNPs) and functional annotations. In this setting, the trait of interest (y_i) is observed and quantified for each of n individuals. The genetic information of each individual is also recorded through sequencing to identify SNPs ($\boldsymbol{x}_i$). In a Genome-Wide Association Study (GWAS), one then regresses the trait of interest on each SNP individually while possibly controlling for additional non-genetic predictors. This produces an estimate the effect of a SNP on the trait of interest. Multiple hypothesis tests are then used to ascertain the statistical significance of each SNP individually. Unfortunately, multiple testing requires adjusting p-values (e.g. through the Benjamini-Hochberg procedure) in order to protect against an inflated false discovery rate. Furthermore, it is not clear how to integrate prior information from related studies that establish compelling empirical evidence for the role of each SNP j in signaling pathways, expression patterns, regulatory functions, and other biologic information that is typically overlapping ($\boldsymbol{z}_j$). In summary, the typical GWAS study considers only univariate linear models of the form $y_i \sim x_{ij}\beta_j + \epsilon_i$ and therefore cannot perform any subset selection or incorporate prior information $\boldsymbol{z}_j$. Comparatively, a hierarchical modelling approach might tackle this problem as follows:

$$\begin{aligned} y_i &\sim \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, &\qquad &\text{for each person } i = 1, 2, \ldots, n. \\ \beta_j &\sim \boldsymbol{z}_j\boldsymbol{\alpha} + \gamma_j, &\qquad &\text{for each SNP } j = 1, 2, \ldots, p. \end{aligned}$$

Here the model parameters $(\alpha_1, \alpha_2, \ldots, \alpha_d)^\top$ account for the contributions of d functional annotations in a particular predictor SNP j. Thus, an ideal situation is to fit both SNP effects and functional annotation effects for a particular trait.

4.2 A Proximal Distance Approach

Recent work on iterative hard-thresholding (IHT) applies multiple regression to a GWAS-like context for continuous and discrete traits based on both linear and generalized linear models [CKG20]. This approach relies on fast iterative updates based on gradient descent that directly enforce sparsity via projections to sparsity sets. However, the projected gradient descent updates will not generalize to our particular hierarchical regression problem. Another related work is implemented in the software package `xrnet` [WL19]. This approach is based on a soft-thresholding coordinate descent algorithm extending the work of Friedman et al. to the exact hierarchical problem considered above [FHT10]. While fast, the `xrnet` approach faces the same issues as the lasso in terms of tuning a hyperparameter to minimize false positives and false negatives.

Here we outline an alternative approach based on the proximal distance method. Formally, we seek solutions to the penalized least-squares problem

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{1}{2}\left[\|\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y}\|_2^2+\lambda\|\boldsymbol{Z}\boldsymbol{\alpha}-\boldsymbol{\beta}\|_2^2+\rho\,\mathrm{dist}(\boldsymbol{\alpha},S_k)^2\right].$$

The first term is the usual least-squares term corresponding to the internal data. The second term accounts for external data that regularizes the main effect estimates, $\boldsymbol{\beta}$, and is modulated by the hyperparameter λ. The third term quantifies distance from a sparsity set with k non-zero elements, applied to the external effects. Recalling our genetics example from before, the motivation for enforcing sparsity only on external effects is due to the fact that a SNP's functional annotations may overlap and may only apply in certain contexts (that is, for the particular trait under analysis). In theory, one might extend the objective above to enforce sparsity on $\boldsymbol{\beta}$ directly but this direction is not discussed here.

The key to producing the MM update is to write the loss as the unified least-squares criterion

$$f(\boldsymbol{\alpha},\boldsymbol{\beta}) \;=\; \frac{1}{2}\left\|\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{0}\end{pmatrix}-\begin{bmatrix}\boldsymbol{0} & \boldsymbol{X}\\ \sqrt{\lambda}\boldsymbol{Z} & -\sqrt{\lambda}\boldsymbol{I}\end{bmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}\right\|^2.$$

The surrogate then can be expressed as

$$h_\rho(\boldsymbol{\alpha},\boldsymbol{\beta}\mid\boldsymbol{\alpha}_n,\boldsymbol{\beta}_n) \;\leq\; \frac{1}{2}\left\|\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{0}\\ \sqrt{\rho}\mathcal{P}(\boldsymbol{\alpha}_n)\end{pmatrix}-\begin{bmatrix}\boldsymbol{0} & \boldsymbol{X}\\ \sqrt{\lambda}\boldsymbol{Z} & -\sqrt{\lambda}\boldsymbol{I}\\ \sqrt{\rho}\boldsymbol{I} & \boldsymbol{0}\end{bmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}\right\|^2.$$

Thus, the MM update reduces to the solution of a least squares problem.

4.2.1 Direct Solution

The normal equations for our least-squares problem read as

$$\begin{bmatrix}\lambda\boldsymbol{Z}^\top\boldsymbol{Z}+\rho\boldsymbol{I} & -\lambda\boldsymbol{Z}^\top\\ -\lambda\boldsymbol{Z} & \boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}\end{bmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} \;=\; \begin{pmatrix}\rho\mathcal{P}(\boldsymbol{\alpha}_n)\\ \boldsymbol{X}^\top\boldsymbol{y}\end{pmatrix}.$$

Solving the first and second blocks for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively, substitution, and rearrangement yields explicit updates

$$\begin{aligned}\boldsymbol{\alpha}_{n+1} &= \boldsymbol{M}_2^{-1}\left[\rho\mathcal{P}(\boldsymbol{\alpha}_n) + \lambda\boldsymbol{Z}^\top\boldsymbol{M}_1^{-1}\boldsymbol{X}^\top\boldsymbol{y}\right] \\ \boldsymbol{\beta}_{n+1} &= \boldsymbol{M}_1^{-1}\left[\boldsymbol{X}^\top\boldsymbol{y} + \lambda\boldsymbol{Z}\boldsymbol{\alpha}_{n+1}\right],\end{aligned}$$

where $\boldsymbol{M}_1 = \boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_{p\times p}$ and $\boldsymbol{M}_2 = \lambda\boldsymbol{Z}^\top\boldsymbol{Z} - \lambda^2\boldsymbol{Z}^\top\boldsymbol{M}_1^{-1}\boldsymbol{Z} + \rho\boldsymbol{I}_{d\times d}$. This direct approach requires inverting two matrices, one of which depends on the penalty coefficient ρ. Fortunately, high-dimensional problems with tall or wide matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ can make use of Woodbury's matrix identity to reduce computational complexity. For example in the case $p \gg n \gg d$ pertinent to genetic analyses, computing $\boldsymbol{X}^\top\boldsymbol{X}$ is prohibitive whereas $\boldsymbol{X}\boldsymbol{X}^\top$ is considerably smaller. Further, one can avoid additional storage requirements by computing a $\boldsymbol{LU}$ factorization for $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ and work instead with matrix-vector operations.

4.2.2 A Steepest Descent Alternative

Motivated by the satisfactory performance of steepest descent methods via proximal distance, we consider the same approach here. For a fixed viscosity ν and surrogate h_ρ anchored at (α_n, β_n), the optimal step η for a steepest descent method is given by

$$\eta = \frac{\|\nabla h_\rho\|^2}{\nabla h_\rho^\top\boldsymbol{A}\nabla h_\rho + \nu\|\nabla h_\rho\|^2},$$

where $\boldsymbol{A}$ is the matrix identified previously in the unified least squares criterion.

4.2.3 Alternative Parameterization

The substitution $\boldsymbol{\gamma} = \boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\alpha}$ transforms the original model to

$$f(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{1}{2}\left\|\boldsymbol{y} - \begin{pmatrix}\boldsymbol{X}\boldsymbol{Z} & \boldsymbol{X}\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\gamma}\end{pmatrix}\right\|^2 + \frac{\lambda}{2}\|\boldsymbol{\gamma}\|^2.$$

We can then apply the same principles to derive an explicit iterative map from the normal equations or a steepest descent method. While the problems $\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} f(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\min_{\boldsymbol{\alpha},\boldsymbol{\gamma}} f(\boldsymbol{\alpha}, \boldsymbol{\gamma})$

are equivalent under a sparsity constraint, the choice of parameterization may converge to different local optima because the penalized objective is not convex. The solution sets between the two versions of the problem may in fact be the same, but one may be preferred depending on how the parameter landscape changes under the change of variables.

4.3 Preliminary results

Here we compare our proximal distance methods that enforce sparsity directly against the ℓ_1-regularized solutions produced by `xrnet`. Throughout these experiments we simulate an instance of the problem with size (n, p, d) and sparsity level k_0 according to the recipe

$$\begin{aligned}
X_{ij} &\sim \text{Uniform}\{0, 1, 2\} \\
Z_{jk} &\sim \text{Normal}(0, 1) \\
\alpha_k &= 1 \text{ or } \alpha_k = 0 \quad \text{such that } \#\{\alpha_k \neq 0\} = k_0 \\
\beta_j &\sim \boldsymbol{Z}_{j,\cdot}\boldsymbol{\alpha} + \text{Normal}(0, 1) \\
y_i &\sim \boldsymbol{X}_{i,\cdot}\boldsymbol{\beta} + \text{Normal}(0, 1).
\end{aligned}$$

This produces $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{Z} \in \mathbb{R}^{p \times d}$, $\boldsymbol{y} \in \mathbb{R}^{n}$, $\boldsymbol{\beta} \in \mathbb{R}^{p}$, and $\boldsymbol{\alpha} \in \mathbb{R}^{d}$. We fix $n = 1000$, $d = 50$, and $k_0 = 10$ while varying the number of predictors $p \in \{2000, 4000, 6000, 8000, 10000\}$. Our proximal distance algorithm is allotted 200 iterations for solving a problem for fixed ρ with the penalty coefficient scaling by a factor of 1.2. Further, our method varies $k \in \{15, 10, 5\}$ and $\lambda \in \{1, 10, 10^2, 10^3\}$. In Figure 4.1 we report performance of our steepest descent method against `xrnet` on the basis of mean squared error (MSE) with respect to ground truth. Figure 4.2 reports results on the basis of false positives and false negatives, respectively. Figure 4.3 compares our proximal distance method against `xrnet`. These preliminary results suggest that our sparsity-based algorithms avoid a high false positive rate while preserving a low false negative rate, which is a notable improvement over `xrnet`. Further work is needed in implementing a fast cross-validation procedure using our algorithms.

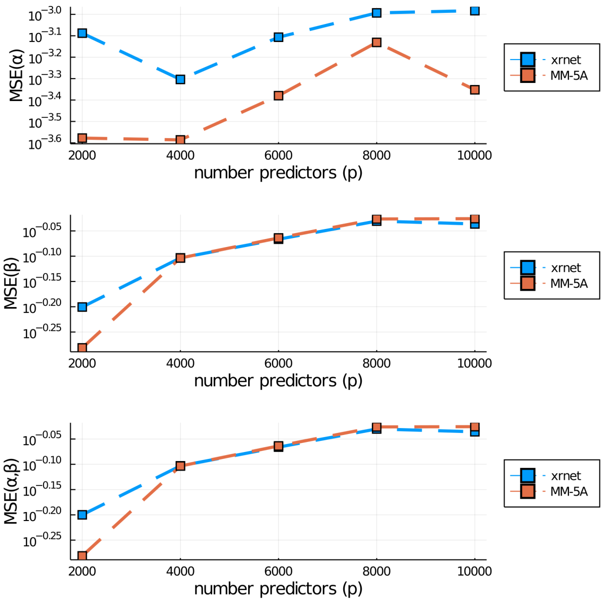


Figure 4.1: **Comparison based on mean squared error.** MSE optimal solutions with respect to ground truth on the basis of $\boldsymbol{\alpha}$ only (top), $\boldsymbol{\beta}$ only (middle), and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ jointly (bottom). Our MM algorithms outperform `xrnet` in estimating external effects $\boldsymbol{\alpha}$.

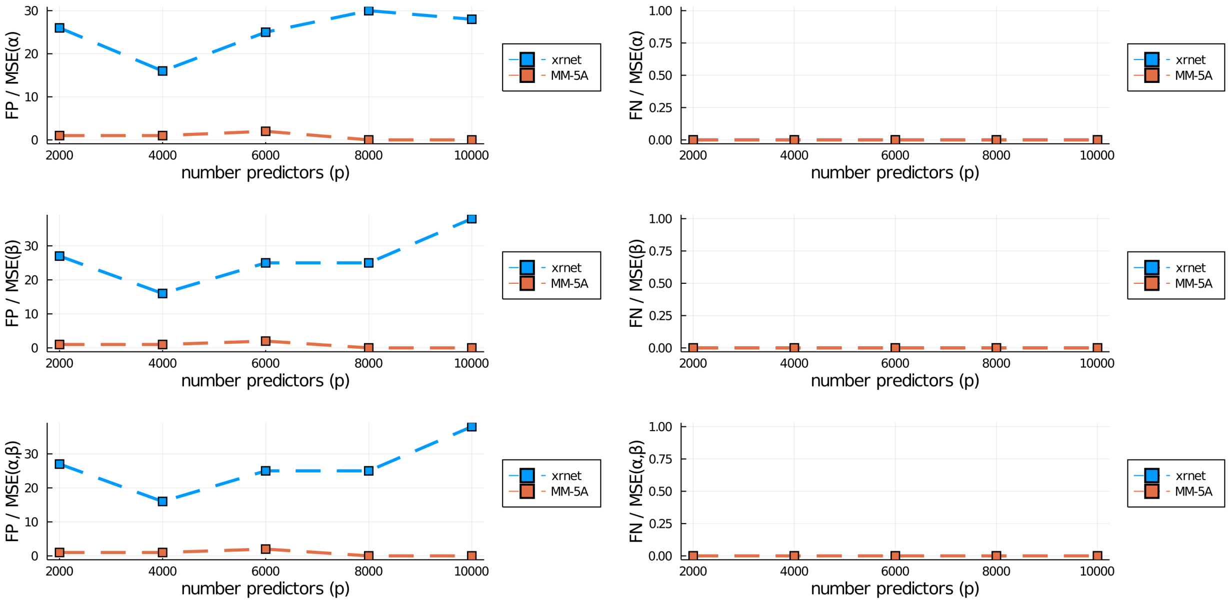


Figure 4.2: **Comparison based on false positives and false negatives.** The number of false positives is based on solutions that minimize MSE with respect to $\boldsymbol{\alpha}$ only (top), $\boldsymbol{\beta}$ only (middle), and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ jointly (bottom). Our MM algorithms are successful in controlling false discoveries and show an improvement compared to `xrnet`.

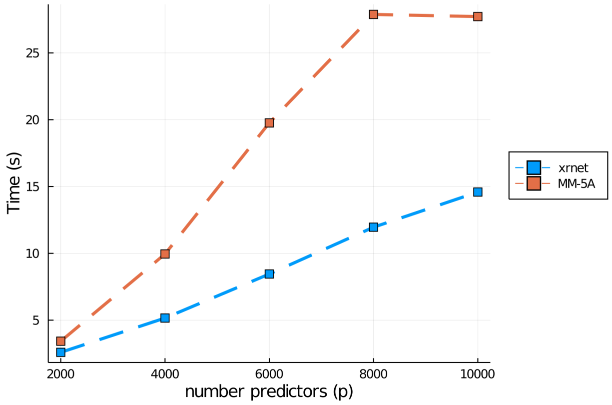


Figure 4.3: **Timing comparisons with `xrnet`.** Our proximal distance method has a higher cost in terms of compute time. Importantly, the `xrnet` software results reflect the time needed to compute solutions over a wide range of values for the hyperparameters λ_1 and λ_2.

REFERENCES

[Ame20a] American Academy of Pediatrics. "COVID-19 Planning Considerations: Guidance for School Re-Entry." http://services.aap.org/en/pages/2019-novel-coronavirus-covid-19-infections/clinical-guidance/covid-19-planning-considerations-return-to-in-person-education-in-schools/, June 2020.

[Ame20b] American Federation of Teachers. "Pediatricians, Educators and Superintendents Urge a Safe Return to School This Fall." https://www.aft.org/press-release/pediatricians-educators-and-superintendents-urge-safe-return-school-fall, July 2020.

[Ame21] American Academy of Pediatrics. "Children and COVID-19 State Data Report." https://downloads.aap.org/AAP/PDF/AAP%20and%20CHA%20-%20Children%20and%20COVID-19%20State%20Data%20Report%202.4.21%20FINAL.pdf, February 2021.

[ASR20] Katherine A. Auger, Samir S. Shah, Troy Richardson, David Hartley, Matthew Hall, Amanda Warniment, Kristen Timmons, Dianna Bosse, Sarah A. Ferris, Patrick W. Brady, Amanda C. Schondelmeyer, and Joanna E. Thomson. "Association Between Statewide School Closure and COVID-19 Incidence and Mortality in the US." *JAMA*, July 2020.

[BAX20] Fan Bu, Allison E. Aiello, Jason Xu, and Alexander Volfovsky. "Likelihood-Based Inference for Partially Observed Epidemics on Dynamic Networks." *Journal of the American Statistical Association*, pp. 1–17, August 2020.

[BC17] Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd Edition*, volume 408. Springer, 2017.

[BDS08] Justin Brickell, Inderjit S. Dhillon, Suvrit Sra, and Joel A. Tropp. "The Metric Nearness Problem." *SIAM Journal on Matrix Analysis and Applications*, **30**(1):375–396, January 2008.

[Bec17] Amir Beck. *First-Order Methods in Optimization*, volume 25. SIAM, 2017.

[Bel70] Edward J Beltrami. *An Algorithmic Approach to Nonlinear Analysis and Optimization.* Academic Press, 1970.

[BFP20] Yinon M Bar-On, Avi Flamholz, Rob Phillips, and Ron Milo. "SARS-CoV-2 (COVID-19) by the Numbers." *eLife*, **9**:e57309, March 2020.

[BKL75] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz. "A New Algorithm for Monte Carlo Simulation of Ising Spin Systems." *Journal of Computational Physics*, **17**(1):10–18, January 1975.

[BL10] Jonathan Borwein and Adrian S Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples.* Springer Science & Business Media, 2010.

[BM78] Robert B. Bendel and M. Ray Mickey. "Population Correlation Matrices for Sampling Experiments." *Communications in Statistics - Simulation and Computation*, **7**(2):163–182, January 1978.

[BPC11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers." *Foundations and Trends in Machine Learning*, **3**(1):1–122, 2011.

[Bri20] Tom Britton. "Epidemic Models on Social Networks—With Inference." *Statistica Neerlandica*, **74**(3):222–241, 2020.

[Cen20] Centers for Disease Control and Prevention. "Multisystem Inflammatory Syndrome in Children (MIS-C) Associated with Coronavirus Disease 2019 (COVID-19)." https://emergency.cdc.gov/han/2020/han00432.asp, May 2020.

[Cen21] Centers for Disease Control and Prevention. "COVID-19 Stats: Percentage of Middle and High School Students Aged 13–21 Years Attending In-Person Classes Who Reported Observing Fellow Students Wearing a Mask All the Time, by School Setting and Activity — United States, October 2020." *MMWR. Morbidity and Mortality Weekly Report*, **70**, 2021.

[CGP05] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. "Avoiding Negative Populations in Explicit Poisson Tau-Leaping." *The Journal of Chemical Physics*, **123**(5):054104, August 2005.

[CGP06] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. "Efficient Step Size Selection for the Tau-Leaping Simulation Method." *The Journal of Chemical Physics*, **124**(4):044109, January 2006.

[CIM10] Emilie Chouzenoux, Jérôme Idier, and Saïd Moussaoui. "A Majorize–Minimize Strategy for Subspace Optimization Applied to Image Restoration." *IEEE Transactions on Image Processing*, **20**(6):1517–1528, 2010.

[CKG20] Benjamin B Chu, Kevin L Keys, Christopher A German, Hua Zhou, Jin J Zhou, Eric M Sobel, Janet S Sinsheimer, and Kenneth Lange. "Iterative Hard Thresholding in Genome-Wide Association Studies: Generalized Linear Models, Prior Weights, and Double Sparsity." *GigaScience*, **9**(giaa044), June 2020.

[CL15] Eric C. Chi and Kenneth Lange. "Splitting Methods for Convex Clustering." *Journal of Computational and Graphical Statistics*, **24**(4):994–1013, October 2015.

[CLP04] Yang Cao, Hong Li, and Linda Petzold. “Efficient Formulation of the Stochastic Simulation Algorithm for Chemically Reacting Systems.” *The Journal of Chemical Physics*, **121**(9):4059–4067, August 2004.

[Con16] Laurent Condat. “Fast Projection onto the Simplex and the $_1$ Ball.” *Mathematical Programming*, **158**(1-2):575–585, July 2016.

[Cou43] Richard Courant. *Variational Methods for the Solution of Problems of Equilibrium and Vibrations*. Verlag Nicht Ermittelbar, 1943.

[CS12] Forrest W. Crawford and Marc A. Suchard. “Transition Probabilities for General Birth–Death Processes with Applications in Ecology, Genetics, and Evolution.” *Journal of Mathematical Biology*, **65**(3):553–580, September 2012.

[CSL16] Forrest W. Crawford, Timothy C. Stutz, and Kenneth Lange. “Coupling Bounds for Approximating Birth–Death Processes by Truncation.” *Statistics & Probability Letters*, **109**:30–38, February 2016.

[CZL14] Eric C Chi, Hua Zhou, and Kenneth Lange. “Distance Majorization and Its Applications.” *Mathematical Programming*, **146**(1-2):409–436, 2014.

[Deb52] Gerard Debreu. “Definite and Semidefinite Quadratic Forms.” *Econometrica: Journal of the Econometric Society*, pp. 295–300, 1952.

[DH00] Philip I. Davies and Nicholas J. Higham. “Numerically Stable Generation of Correlation Matrices and Their Factors.” *BIT Numerical Mathematics*, **40**(4):640–651, December 2000.

[DHR10] O. Diekmann, J. A. P. Heesterbeek, and M. G. Roberts. “The Construction of Next-Generation Matrices for Compartmental Epidemic Models.” *Journal of The Royal Society Interface*, **7**(47):873–885, June 2010.

[DKC20] Elizabeth M. Dufort, Emilia H. Koumans, Eric J. Chow, Elizabeth M. Rosenthal, Alison Muse, Jemma Rowlands, Meredith A. Barranco, Angela M. Maxted, Eli S. Rosenberg, Delia Easton, Tomoko Udo, Jessica Kumar, Wendy Pulver, Lou Smith, Brad Hutton, Debra Blog, and Howard Zucker. “Multisystem Inflammatory Syndrome in Children in New York State.” *New England Journal of Medicine*, June 2020.

[DKL20] Nicholas G. Davies, Petra Klepac, Yang Liu, Kiesha Prem, Mark Jit, and Rosalind M. Eggo. “Age-Dependent Effects in the Transmission and Control of COVID-19 Epidemics.” *Nature Medicine*, pp. 1–7, June 2020.

[DMK21] Jennifer M. Dan, Jose Mateus, Yu Kato, Kathryn M. Hastie, Esther Dawen Yu, Caterina E. Faliti, Alba Grifoni, Sydney I. Ramirez, Sonya Haupt, April Frazier, Catherine Nakao, Vamseedhar Rayaprolu, Stephen A. Rawlings, Bjoern Peters, Florian Krammer, Viviana Simon, Erica Ollmann Saphire, Davey M. Smith, Daniela Weiskopf, Alessandro Sette, and Shane Crotty. “Immunological Memory to SARS-CoV-2 Assessed for up to 8 Months after Infection.” *Science*, **371**(6529), February 2021.

[DXW20] Zhanwei Du, Xiaoke Xu, Ye Wu, Lin Wang, Benjamin J. Cowling, and Lauren Ancel Meyers. “Serial Interval of COVID-19 among Publicly Reported Confirmed Cases.” *Emerging Infectious Diseases*, **26**(6):1341–1343, 2020.

[ELS02] Michael B. Elowitz, Arnold J. Levine, Eric D. Siggia, and Peter S. Swain. “Stochastic Gene Expression in a Single Cell.” *Science*, **297**(5584):1183–1186, 2002.

[FHT10] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software*, **33**(1):1–22, February 2010.

[FS11] David Chin-Lung Fong and Michael Saunders. “LSMR: An Iterative Algorithm for Sparse Least-Squares Problems.” *SIAM Journal on Scientific Computing*, **33**(5):2950–2971, 2011.

[GB00] Michael A. Gibson and Jehoshua Bruck. “Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels.” *The Journal of Physical Chemistry A*, **104**(9):1876–1889, March 2000.

[GB16] Pontus Giselsson and Stephen Boyd. “Linear Convergence and Metric Selection for Douglas-Rachford Splitting and ADMM.” *IEEE Transactions on Automatic Control*, **62**(2):532–544, 2016.

[Gil76] Daniel T Gillespie. “A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions.” *Journal of Computational Physics*, **22**(4):403–434, 1976.

[Gil77] Daniel T. Gillespie. “Exact Stochastic Simulation of Coupled Chemical Reactions.” *The Journal of Physical Chemistry*, **81**(25):2340–2361, 1977.

[Gil01] Daniel T. Gillespie. “Approximate Accelerated Stochastic Simulation of Chemically Reacting Systems.” *Journal of Chemical Physics*, **115**(4), 2001.

[GM18] Abhishekh Gupta and Pedro Mendes. “An Overview of Network-Based and -Free Approaches for Stochastic Simulation of Biochemical Systems.” *Computation*, **6**(1):9, March 2018.

[GP03] Daniel T. Gillespie and Linda R. Petzold. “Improved Leap-Size Selection for Accelerated Stochastic Simulation.” *Journal of Chemical Physics*, **19**:8229–8234, 2003.

[GV96] HG Golub and Charles F Van Loan. *Matrix Computations, 3 Edition.* Johns Hopkins University Press, 1996.

[GWL21] Christian Gaebler, Zijun Wang, Julio C. C. Lorenzi, Frauke Muecksch, Shlomo Finkin, Minami Tokuyama, Alice Cho, Mila Jankovic, Dennis Schaefer-Babajew, Thiago Y. Oliveira, Melissa Cipolla, Charlotte Viant, Christopher O. Barnes, Yaron Bram, Gaëlle Breton, Thomas Hägglöf, Pilar Mendoza, Arlene Hurley, Martina Turroja, Kristie Gordon, Katrina G. Millard, Victor Ramos, Fabian Schmidt, Yiska Weisblum, Divya Jha, Michael Tankelevich, Gustavo Martinez-Delgado, Jim Yee, Roshni Patel, Juan Dizon, Cecille Unson-O'Brien, Irina Shimeliovich, Davide F. Robbiani, Zhen Zhao, Anna Gazumyan, Robert E. Schwartz, Theodora Hatziioannou, Pamela J. Bjorkman, Saurabh Mehandru, Paul D. Bieniasz, Marina Caskey, and Michel C. Nussenzweig. "Evolution of Antibody Immunity to SARS-CoV-2." *Nature*, pp. 1–6, January 2021.

[Hig08] D. Higham. "Modeling and Simulating Chemical Reactions." *SIAM Review*, **50**(2):347–368, January 2008.

[HK70] Arthur E. Hoerl and Robert W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, **12**(1):55–67, February 1970.

[HLR16] Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. "Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems." *SIAM Journal on Optimization*, **26**(1):337–364, 2016.

[HLW20] Xi He, Eric H. Y. Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y. Wong, Yujuan Guan, Xinghua Tan, Xiaoneng Mo, Yanqing Chen, Baolin Liao, Weilie Chen, Fengyu Hu, Qing Zhang, Mingqiu Zhong, Yanrong Wu, Lingzhai Zhao, Fuchun Zhang, Benjamin J. Cowling, Fang Li, and Gabriel M. Leung. "Temporal Dynamics in Viral Shedding and Transmissibility of COVID-19." *Nature Medicine*, **26**(5):672–675, May 2020.

[JKY19] Yufeng Jiang, Remy Kassem, Grayson York, Matthew Junge, and Rick Durrett. "SIR Epidemics on Evolving Graphs." *arXiv:1901.06568 [math]*, January 2019.

[JMV20] Terry C. Jones, Barbara Mühlemann, Talitha Veith, Guido Biele, Marta Zuchowski, Jörg Hoffmann, Angela Stein, Anke Edelmann, Victor Max Corman, and Christian Drosten. "An Analysis of SARS-CoV-2 Viral Load by Patient Age." *medRxiv*, p. 2020.06.08.20125484, June 2020.

[KBM20] Omer Karin, Yinon M. Bar-On, Tomer Milo, Itay Katzir, Avi Mayo, Yael Korem, Boaz Dudovich, Eran Yashiv, Amos J. Zehavi, Nadav Davidovich, Ron Milo, and Uri Alon. "Adaptive Cyclic Exit Strategies from Lockdown to Suppress COVID-19 and Allow Economic Activity." Preprint, Epidemiology, April 2020.

[KM08] Hiroyuki Kuwahara and Ivan Mura. "An Efficient and Exact Stochastic Simulation Method to Analyze Rare Events in Biochemical Systems." *The Journal of Chemical Physics*, **129**(16):165101, October 2008.

[KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. "Linear Convergence of Gradient and Proximal-Gradient Methods under the Polyak-Lojasiewicz Condition." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

[KST20] Megan Kuhfeld, James Soland, Beth Tarasawa, Angela Johnson, Erik Ruzek, and Jing Liu. "Projecting the Potential Impact of COVID-19 School Closures on Academic Achievement." *Educational Researcher*, **49**(8):549–565, November 2020.

[KT20] Megan Kuhfeld and Beth Tarasawa. "The COVID-19 Slide: What Summer Learning Loss Can Tell Us about the Potential Impact of School Closures on Student Academic Achievement." *Northwest Evaluation Association*, 2020.

[KZL19] Kevin L Keys, Hua Zhou, and Kenneth Lange. "Proximal Distance Algorithms: Theory and Practice." *Journal of Machine Learning Research*, **20**(66):1–38, 2019.

[Lan10a] Kenneth Lange. *Applied Probability.* Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2010.

[Lan10b] Kenneth Lange. *Numerical Analysis for Statisticians.* Springer Science & Business Media, 2010.

[Lan16] Kenneth Lange. *MM Optimization Algorithms*, volume 147. SIAM, 2016.

[LGB20] Stephen A. Lauer, Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. "The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application." *Annals of Internal Medicine*, **172**(9):577–582, March 2020.

[Lig99] Thomas M. Liggett. *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes.* Grundlehren Der Mathematischen Wissenschaften. Springer-Verlag, Berlin Heidelberg, 1999.

[Lig05] Thomas M. Liggett. *Interacting Particle Systems.* Classics in Mathematics. Springer-Verlag, Berlin Heidelberg, 2005.

[Lop20] Adriana S. Lopez. "Transmission Dynamics of COVID-19 Outbreaks Associated with Child Care Facilities — Salt Lake City, Utah, April–July 2020." *MMWR. Morbidity and Mortality Weekly Report*, **69**, 2020.

[LPC20] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. "Substantial Undocumented Infection Facilitates the Rapid Dissemination of Novel Coronavirus (SARS-CoV-2)." *Science*, **368**(6490):489–493, May 2020.

[LSK18] Alfonso Landeros, Timothy Stutz, Kevin L. Keys, Alexander Alekseyenko, Janet S. Sinsheimer, Kenneth Lange, and Mary E. Sehl. "BioSimulator.Jl: Stochastic Simulation in Julia." *Computer Methods and Programs in Biomedicine*, **167**:23–35, December 2018.

[Lud20] Jonas F. Ludvigsson. "Systematic Review of COVID-19 in Children Shows Milder Cases and a Better Prognosis than Adults." *Acta Paediatrica*, **109**(6):1088–1095, 2020.

[LWL20] Daniel B Larremore, Bryan Wilder, Evan Lester, Soraya Shehata, James M Burke, James A Hay, Milind Tambe, Michael J Mina, and Roy Parker. "Test Sensitivity Is Secondary to Frequency and Turnaround Time for COVID-19 Surveillance." *medRxiv*, p. 2020.06.22.20136309, January 2020.

[MCI19] Rahul Mazumder, Arkopal Choudhury, Garud Iyengar, and Bodhisattva Sen. "A Computational Framework for Multivariate Convex Regression and Its Variants." *Journal of the American Statistical Association*, **114**(525):318–331, January 2019.

[MK06] Brian Munsky and Mustafa Khammash. "The Finite State Projection Algorithm for the Solution of the Chemical Master Equation." *The Journal of Chemical Physics*, **124**(4):044104, January 2006.

[MKR20] Timo Mitze, Reinhold Kosfeld, Johannes Rode, and Klaus Wälde. "Face Masks Considerably Reduce COVID-19 Cases in Germany: A Synthetic Control Method Approach." *Bonn, Germany: IZA Institute of Labour Econoimics*, 2020.

[MPC06] James M. McCollum, Gregory D. Peterson, Chris D. Cox, Michael L. Simpson, and Nagiza F. Samatova. "The Sorting Direct Method for Stochastic Simulation of Biochemical Systems with Varying Reaction Execution Behavior." *Computational Biology and Chemistry*, **30**(1):39–49, February 2006.

[MPT17] Luca Marchetti, Corrado Priami, and Vo Hong Thanh. *Simulation Algorithms for Computational Systems Biology.* Texts in Theoretical Computer Science. An EATCS Series. Springer International Publishing, 2017.

[MT00] George Marsaglia and Wai Wan Tsang. "The Ziggurat Method for Generating Random Variables." *Journal of Statistical Software*, **5**(1):1–7, October 2000.

[MW11] Lesley T. MacNeil and Albertha J. M. Walhout. "Gene Regulatory Networks and the Role of Robustness and Stochasticity in the Control of Gene Expression." *Genome Research*, **21**(5):645–657, May 2011.

[Nat19] National Center for Education Statistics. “Digest of Education Statistics.” Technical report, U.S. Department of Education, December 2019.

[Nes13] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

[Nor98] J. R. Norris. *Markov Chains*. Cambridge University Press, July 1998.

[NW72] John Ashworth Nelder and Robert WM Wedderburn. “Generalized Linear Models.” *Journal of the Royal Statistical Society: Series A (General)*, **135**(3):370–384, 1972.

[Off20] Office of Governor Gavin Newsom. “Governor Gavin Newsom Lays Out Pandemic Plan for Learning and Safe Schools.” https://www.gov.ca.gov/2020/07/17/governor-gavin-newsom-lays-out-pandemic-plan-for-learning-and-safe-schools/, July 2020.

[PCP20] Young Joon Park, Young June Choe, Ok Park, Shin Young Park, Young-Man Kim, Jieun Kim, Sanghui Kweon, Yeonhee Woo, Jin Gwack, Seong Sun Kim, Jin Lee, Junghee Hyun, Boyeong Ryu, Yoon Suk Jang, Hwami Kim, Seung Hwan Shin, Seonju Yi, Sangeun Lee, Hee Kyoung Kim, Hyeyoung Lee, Yeowon Jin, Eunmi Park, Seung Woo Choi, Miyoung Kim, Jeongsuk Song, Si Won Choi, Dongwook Kim, Byoung-Hak Jeon, Hyosoon Yoo, and Eun Kyeong Jeong. “Contact Tracing during Coronavirus Disease Outbreak, South Korea, 2020 - Volume 26, Number 10—October 2020 - Emerging Infectious Diseases Journal - CDC.” *Emerging Infectious Diseases*, 2020.

[PS82] Christopher C Paige and Michael A Saunders. “Algorithm 583: LSQR: Sparse Linear Equations and Least Squares Problems.” *ACM Transactions on Mathematical Software (TOMS)*, **8**(2):195–209, 1982.

[PZW20] A. David Paltiel, Amy Zheng, and Rochelle P. Walensky. “Assessment of SARS-CoV-2 Screening Strategies to Permit the Safe Reopening of College Campuses in the United States.” *JAMA Network Open*, **3**(7):e2016818–e2016818, July 2020.

[RGP10] Min K. Roh, Dan T. Gillespie, and Linda R. Petzold. “State-Dependent Biasing Method for Importance Sampling in the Weighted Stochastic Simulation Algorithm.” *The Journal of Chemical Physics*, **133**(17):174106, November 2010.

[RN17] Christopher Rackauckas and Qing Nie. “DifferentialEquations.Jl – A Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia.” *Journal of Open Research Software*, **5**(1):15, May 2017.

[RN19] Christopher Rackauckas and Qing Nie. “Confederated Modular Differential Equation APIs for Accelerated Algorithm Development and Benchmarking.” *Advances in Engineering Software*, **132**:1–6, June 2019.

[ROF92] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. "Nonlinear Total Variation Based Noise Removal Algorithms." *Physica D: Nonlinear Phenomena*, **60**(1-4):259–268, November 1992.

[SAL09] Mary Sehl, Alexander V. Alekseyenko, and Kenneth L. Lange. "Accurate Stochastic Simulation via the Step Anticipation Tau-Leaping (SAL) Algorithm." *Journal of Computational Biology*, **16**(9):1195–1208, September 2009.

[SCB21] Aurélien Sokal, Pascal Chappert, Giovanna Barba-Spaeth, Anais Roeser, Slim Fourati, Imane Azzaoui, Alexis Vandenberghe, Ignacio Fernandez, Annalisa Meola, Magali Bouvier-Alias, Etienne Crickx, Asma Beldi-Ferchiou, Sophie Hue, Laetitia Languille, Marc Michel, Samia Baloul, France Noizat-Pirenne, Marine Luka, Jérôme Mégret, Mickaël Ménager, Jean-Michel Pawlotsky, Simon Fillatreau, Felix A. Rey, Jean-Claude Weill, Claude-Agnès Reynaud, and Matthieu Mahévas. "Maturation and Persistence of the Anti-SARS-CoV-2 Memory B Cell Response." *Cell*, **0**(0), February 2021.

[Sci20a] SciML. "DifferentialEquations.Jl." Checkout v6.15.0, July 2020.

[Sci20b] SciML. "OrdinaryDiffEq.Jl." Checkout v5.41.0, June 2020.

[SFE11] Michael W. Sneddon, James R. Faeder, and Thierry Emonet. "Efficient Modeling, Simulation and Coarse-Graining of Biological Complexity with NFsim." *Nature Methods*, **8**(2):177–183, February 2011.

[SGM20] Jeffrey Seow, Carl Graham, Blair Merrick, Sam Acors, Suzanne Pickering, Kathryn J. A. Steel, Oliver Hemmings, Aoife O'Byrne, Neophytos Kouphou, Rui Pedro Galao, Gilberto Betancor, Harry D. Wilson, Adrian W. Signell, Helena Winstone, Claire Kerridge, Isabella Huettner, Jose M. Jimenez-Guardeño, Maria Jose Lista, Nigel Temperton, Luke B. Snell, Karen Bisnauthsing, Amelia Moore, Adrian Green, Lauren Martinez, Brielle Stokes, Johanna Honey, Alba Izquierdo-Barras, Gill Arbane, Amita Patel, Mark Kia Ik Tan, Lorcan O'Connell, Geraldine O'Hara, Eithne MacMahon, Sam Douthwaite, Gaia Nebbia, Rahul Batra, Rocio Martinez-Nunez, Manu Shankar-Hari, Jonathan D. Edgeworth, Stuart J. D. Neil, Michael H. Malim, and Katie J. Doores. "Longitudinal Observation and Decline of Neutralizing Antibody Responses in the Three Months Following SARS-CoV-2 Infection in Humans." *Nature Microbiology*, **5**(12):1598–1607, December 2020.

[SML19] Ryan Suderman, Eshan D. Mitra, Yen Ting Lin, Keesha E. Erickson, Song Feng, and William S. Hlavacek. "Generalizing Gillespie's Direct Method to Enable Network-Free Simulations." *Bulletin of Mathematical Biology*, **81**(8):2822–2848, August 2019.

[SO15] Kevin R. Sanft and Hans G. Othmer. “Constant-Complexity Stochastic Simulation Algorithm with Optimal Binning.” *The Journal of Chemical Physics*, **143**(7):074108, August 2015.

[SS11] Emilio Seijo and Bodhisattva Sen. “Nonparametric Least Squares Estimation of a Multivariate Convex Regression Function.” *The Annals of Statistics*, **39**(3):1633–1657, June 2011.

[SSS20] Aziz Sheikh, Asiyah Sheikh, Zakariya Sheikh, and Sangeeta Dhami. “Reopening Schools after the COVID-19 Lockdown.” *Journal of Global Health*, **10**(1):010376, June 2020.

[STD05] Suvrit Sra, Joel Tropp, and Inderjit S Dhillon. “Triangle Fixing Algorithms for the Metric Nearness Problem.” In *Advances in Neural Information Processing Systems*, pp. 361–368, 2005.

[Sza20] Christine M. Szablewski. “SARS-CoV-2 Transmission and Infection Among Attendees of an Overnight Camp — Georgia, June 2020.” *MMWR. Morbidity and Mortality Weekly Report*, **69**, 2020.

[SZH13] Jianping Sun, Yingye Zheng, and Li Hsu. “A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies.” *Genetic Epidemiology*, **37**(4):334–344, 2013.

[SZS11] Mary Sehl, Hua Zhou, Janet S. Sinsheimer, and Kenneth L. Lange. “Extinction Models for Cancer Stem Cell Therapy.” *Mathematical Biosciences*, **234**(2):132–146, December 2011.

[TA77] Andrei Nikolaevich Tikhonov and Vasilii Yakovlevich Arsenin. *Solutions of Ill-Posed Problems*. Wiley, New York, 1977.

[Tib96] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1):267–288, 1996.

[TSR05a] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. “Sparsity and Smoothness via the Fused Lasso.” *Journal of the Royal Statistical Society: Series B*, **67**(1):91–108, 2005.

[TSR05b] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. “Sparsity and Smoothness via the Fused Lasso.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(1):91–108, February 2005.

[TZP15] Vo Hong Thanh, Roberto Zunino, and Corrado Priami. “On the Rejection-Based Algorithm for Simulation and Analysis of Large-Scale Reaction Networks.” *The Journal of Chemical Physics*, **142**(24):244106, June 2015.

[VEB10] Nguyen Xuan Vinh, Julien Epps, and James Bailey. "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance." *Journal of Machine Learning Research*, **11**(95):2837–2854, 2010.

[VMB21] Russell M. Viner, Oliver T. Mytton, Chris Bonell, G. J. Melendez-Torres, Joseph Ward, Lee Hudson, Claire Waddington, James Thomas, Simon Russell, Fiona van der Klis, Archana Koirala, Shamez Ladhani, Jasmina Panovska-Griffiths, Nicholas G. Davies, Robert Booy, and Rosalind M. Eggo. "Susceptibility to SARS-CoV-2 Infection Among Children and Adolescents Compared With Adults: A Systematic Review and Meta-Analysis." *JAMA Pediatrics*, **175**(2):143, February 2021.

[VMG20] Lucio Verdoni, Angelo Mazza, Annalisa Gervasoni, Laura Martelli, Maurizio Ruggeri, Matteo Ciuffreda, Ezio Bonanomi, and Lorenzo D'Antiga. "An Outbreak of Severe Kawasaki-like Disease at the Italian Epicentre of the SARS-CoV-2 Epidemic: An Observational Cohort Study." *The Lancet*, **395**(10239):1771–1778, June 2020.

[Wil06] Darren J. Wilkinson. *Stochastic Modeling for Systems Biology.* Chapman & Hall/CRC Press, Boca Raton, FL, first edition, 2006.

[WL19] Garrett M. Weaver and Juan Pablo Lewinger. "Xrnet: Hierarchical Regularized Regression to Incorporate External Data." *Journal of Open Source Software*, **4**(44):1761, December 2019.

[XCL17] Jason Xu, Eric Chi, and Kenneth Lange. "Generalized Linear Model Regression under Distance-to-Set Penalties." In *Advances in Neural Information Processing Systems*, pp. 1385–1395, 2017.

[YKN20] Chee Fu Yung, Kai-qian Kam, Karen Donceras Nadua, Chia Yin Chong, Natalie Woon Hui Tan, Jiahui Li, Khai Pin Lee, Yoke Hwee Chan, Koh Cheng Thoon, and Kee Chong Ng. "Novel Coronavirus 2019 Transmission Risk in Educational Settings." *Clinical Infectious Diseases*, June 2020.

[ZBH20] Yanshan Zhu, Conor J Bloxham, Katina D Hulme, Jane E Sinclair, Zhen Wei Marcus Tong, Lauren E Steele, Ellesandra C Noye, Jiahai Lu, Keng Yih Chew, Janessa Pickering, Charles Gilks, Asha C Bowen, and Kirsty R Short. "Children Are Unlikely to Have Been the Primary Source of Household SARS-CoV-2 Infections." *medRxiv*, p. 2020.03.26.20044826, January 2020.

[ZC20] Petra Zimmermann and Nigel Curtis. "Coronavirus Infections in Children Including COVID-19: An Overview of the Epidemiology, Clinical Features, Diagnosis, Treatment and Prevention Options in Children." *The Pediatric Infectious Disease Journal*, **39**(5):355–368, May 2020.

[ZH16] Weijian Zhang and Nicholas J. Higham. "Matrix Depot: An Extensible Test Matrix Collection for Julia." *PeerJ Computer Science*, **2**:e58, April 2016.

[ZLL20] Juanjuan Zhang, Maria Litvinova, Yuxia Liang, Yan Wang, Wei Wang, Shanlu Zhao, Qianhui Wu, Stefano Merler, Cécile Viboud, Alessandro Vespignani, Marco Ajelli, and Hongjie Yu. "Changes in Contact Patterns Shape the Dynamics of the COVID-19 Outbreak in China." *Science*, **368**(6498):1481–1486, June 2020.