©Copyright 2020
Jize Zhang

Nonconvex Optimization Methods with Applications to Portfolio Selection and Hybrid Systems

Jize Zhang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Aleksandr Aravkin, Chair

Samuel Burden

Archis Ghate

Tim Leung

Program Authorized to Offer Degree:
Applied Mathematics

University of Washington

Abstract

Nonconvex Optimization Methods with Applications to
Portfolio Selection and Hybrid Systems

Jize Zhang

Chair of the Supervisory Committee:
Associate Professor Aleksandr Aravkin
Department of Applied Mathematics

This thesis focuses on formulating selection problems using continuous optimization, and solving them by specialized algorithms. Problems involving selection, i.e., selecting "best" candidate(s) out of a given set, occur frequently in various applications, and can be formulated as nonconvex optimization problems. We focus on two applications, portfolio optimization in finance and hybrid systems inference in control theory. We show that using techniques including recently developed relaxations for nonconvex functions, we are able to formulate these problems as structured nonconvex problems and develop efficient algorithms with standard convergence guarantees. The effectiveness of the algorithms is demonstrated in detail for both applications with numerical experiments.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

The author wishes to express sincere gratitude to her adviser Sasha, who is always considerate and kindly supported her throughout the program, on but not limited to research and career path; and to her collaborators, for pushing through the project with her despite setbacks and at times of frustration.

Chapter 1

INTRODUCTION

Convex analysis and optimization techniques are ubiquitous in the mathematical sciences, providing a standard toolbox to answer a broad range of scientific questions. If the question of interest can be formulated as a convex program, there are numerous advantages, including theoretical guarantees, fast algorithms with known worst-case rates, and packages to solve the problem. However scientists, academics, and practitioners often ask questions that cannot be formulated and solved using convex models. An important class of such problems involves some kind of selection, either as the main problem of interest or an aspect of the problem. Examples include partitioning data into clusters (unsupervised learning), selecting the best subset of discrete items according to a given criteria, classifying data into inliers and outliers while fitting a model, and selecting data generating mechanisms from a set of candidates during inference. These problems, and ones similar to them, can rarely be addressed by convex methods.

This thesis is concerned with developing methods for nonconvex optimization problems that involves selection. The overarching scheme is to use relaxations and specialized techniques such as variable projection to transform the original problem into a more tractable problem class. The resulting class of *convex composite* problems has been shown to possess nice properties and allows efficient algorithms with standard convergence guarantees. Details on the convex composite class and these algorithms are gathered in the preliminary chapter. The specific problems we consider are motivated by collaborative work with scientists in domain applications, particularly in portfolio optimization and dynamic systems inference.

Portfolio selection is a classic problem in finance. The goal is to construct an optimal portfolio with respect to a given criterion. For instance, classic Markowitz portfolio analysis

seeks to get the best expected returns at a given level of portfolio variability, or equivalently minimize variability but achieve a given level of expected return. Conditional variance-at-risk (CVaR) portfolio optimization, on the other hand, aims at constructing portfolios with low risk of high loss. There have been many extensions and refinements by both practitioners and academics on classic portfolio selection approach, in particular on Markowitz. Among them, one type of constraints arises from industry practice. For many portfolio managers, it is desirable or even imperative to limit the number of assets in a portfolio and/or impose limits on the proportion of the portfolio devoted to any particular asset or asset class. These constraints can be driven by both the portfolio mandates set by clients (investors), or pragmatic reasons such as transaction costs, minimum lot sizes, and execution efficiency. While classic Markowitz and CVaR portfolio optimization can both be formulated as convex programs, finding sparse portfolios (those comprising fewer assets), or portfolios with specific cardinality constraints, becomes a selection problem that is nonconvex, and is one of the problems for which we develop specific techniques and analysis.

State space models and techniques for fitting them form the bedrock of systems theory, starting from Kalman filtering, and extending through to time series analysis, systems and control, and a wide range of applications, from weather to robotics. The classic problem in dynamic inference seeks to find the "best"' state estimate given known dynamics and noisy measurements. Similar to portfolio optimization, the classic version of the problem is a convex quadratic program. In some applications, however, the dynamics of a system may not be completely known. For instance, one type of systems, known as hybrid dynamical systems, switch between dynamic regimes at time- or state-triggered events. In other words, different dynamics govern a system at different times depending on the system's states, thus identifying the active dynamic model becomes another part of inference and requires selection. While selection in portfolio optimization is static, selection in this case is time-varying, hence efficiency is even more important. We develop a new efficient algorithm for such systems that is applicable under potentially nonlinear dynamics. In addition, we take into account instantaneous displacement that may happen at switches. Our approach is by

viewing it as a counterpart of measurement outliers. When we think about measurement outliers, we essentially select which observation residuals can be left out of the analysis; when we consider high-fidelity tracking of instantaneous changes, we select which innovation residuals we can leave alone. This broader view puts outlier detection and tracking sudden changes squarely in the scope of the thesis.

The paper is organized as follows. We start by introducing preliminaries relevant for building up models and algorithms in the two applications in Chapter 2. We then consider methods, models, and algorithms for portfolio selection in Chapter 3, and methods, models and algorithms for hybrid systems inference in Chapter 4. For each application we describe model formulation, discuss algorithm and convergence, and provide numerical results that illustrate impact within the application.

Chapter 2

PRELIMINARIES

In this section we introduce relevant preliminaries for this paper, including: a) concepts from general convex/variational analysis; b) (non)convex optimization algorithms; c) overview of Kalman smoothing; d) and overview of portfolio optimization. The first two provide foundations for designing an algorithm, whereas the last two lay out framework for building an model, both of which are important for solving problems from an optimization perspective, since we would want a model to reflect the nature of the application problem relatively well but also can be solved efficiently with implementable methods.

2.1 Concepts from convex and variational analysis

In this section we first clarify what we mean by "nonconvex problems" in this paper, since the scope of nonconvexity is immensely large. We then present concepts from convex/variational analysis that are most relevant for this paper. We point readers to well-known books in convex/variational analysis such as [1] and [2] for reference.

2.1.1 Varieties of Nonconvex functions

Compared with convex functions, the set of nonconvex functions is vastly larger. The term "nonconvex" encompasses a spectrum of widely ranging function classes. At one end of the spectrum are well structured function classes such as additive convex composite functions [3], while at the other end lies almost completely general classes such as those studied by [4] and [5]. As one would expect, for more structured the problem classes, there are more applicable methods with better convergence guarantees.

In this thesis "nonconvex functions" refers to structured nonconvex problems. In particular, we look at problems that fall into either additive composite form or general composite form. We also introduce the class of weakly convex problems, which includes composite function class and has been proven to exhibit certain nice properties resembling convex functions.

Definition 1 (Additive Convex Composite)**.** *The problem class*

$$\min_x c(x) + g(x)$$

where c is smooth (but not necessarily convex) and g is convex (but not necessarily smooth).

Definition 2 (Convex Composite)**.** *The problem class*

$$\min_x h(c(x))$$

where c is smooth (but not necessarily convex) and h is usually convex (but not necessarily smooth). The exact regularity conditions differ in literatures and are detailed in section 2.2.2.

Definition 3 (Weakly Convex)**.** *A function f is weakly ρ-convex if $x \mapsto f(x) + \frac{\rho}{2}\|x\|^2$ is convex. In particular, composite function where h is convex and Lipschitz and c is smooth with Lipschitz Jacobian is weakly convex.*

We refer readers to [6] and [7] on discussion of composite function classes and weakly convexity.

2.1.2 Smoothness

For optimization algorithms, two common regularity conditions required are Lipschitz continuity and β-smoothness of the objective function.

Definition 4 (Lipschitz continuity)**.** *We say a function $F : E_1 \to E_2$ is L-Lipschitz if*

$$\|F(z) - F(x)\| \leq L\|z - x\|$$

for all $z, x \in E_1$.

Definition 5 (β-smoothness)**.** *We say a function $F : E_1 \to E_2$ is β-smooth if $F \in C^1$ and ∇F is β-Lipschitz.*

Lemma 1. *A β-smooth function $F : E_1 \to E_2$ satisfies the inequality*

$$\|F(z) - F(x) - \nabla F(x)(z - x)\| \leq \frac{\beta}{2}\|z - x\|^2$$

for all $x, z \in E_1$.

2.1.3 Derivatives

The algorithms we consider in this paper fall under "derivative-based optimization" (as opposed to derivative-free optimization). For smooth functions derivatives refer to gradients; for nonsmooth functions we present below concepts that are generalization of gradients.

Definition 6 (Subgradient and subdifferential for convex functions)**.** *Given a convex function f, a vector v is called a subgradient of f at a point x if the inequality*

$$f(z) \geq f(x) + \langle v, z - x\rangle \quad \forall z.$$

The set of all subgradients at x is called the subdifferential.

Definition 7 (Frechet subdifferential)**.** *The Frechet subdifferential of a nonconvex function f at x is*

$$\hat{\partial} f(x) = \{v : \liminf_{z \neq x \; z \to x} \frac{f(z) - f(x) - \langle v, z - x\rangle}{\|z - x\|} \geq 0\}.$$

Definition 8 (Limiting subdifferential)**.** *The limiting subdifferential of a nonconvex function f at x is defined via a closure process*

$$\partial f(x) = \{v : \exists x^j \to x, f(x^j) \to f(x) \text{ and } v^j \in \hat{\partial} f(x^j) \to v \text{ as } j \to \infty\}.$$

2.1.4 Stationary points

For convex problems, the goal of optimization is to find a global minimum; for nonconvex problems however, this is too ambitious. Finding the global minimum for nonconvex functions is intractable: as pointed out in [8] and [9], the number of function and derivatives

evaluations to find a global minimum for nonconvex functions scales exponentially with respect to input dimension. The more tractable goal in nonconvex optimization is to find a stationary point, and this is often good enough in practice, particularly since one can use multiple random initializations to get the best result across several such points, given a large computational budget. For both applications discussed below, we show only convergence to stationary points.

Definition 9 (Stationary point)**.** *A point x is called stationary for a problem $\min_x f(x)$ if*

$$0 \in \partial f(x)$$

holds, where $\partial f(x)$ denotes the subdifferential at x.

2.1.5 Moreau envelope

Moreau envelope is an extremely important concept in convex analysis. It is closely related to proximal mapping, an idea that can be seen as generalization of projection. We state its definition here, as well as some related concepts such as monotone operator and resolvent. Those are not as important for our purpose but are needed for some proofs later in the paper. Also note that Moreau envelope and proximal mapping are not restricted to convex functions. In particular, for weakly convex functions, differentiability of the envelope still holds, just as it does for convex functions. Other nice properties, including monotonicity, do require convexity.

Definition 10 (Moreau envelope and Proximal mapping)**.** *For a function f and a real $\alpha > 0$, the Moreau envelope and the proximal mapping are defined by*

$$f_\alpha(x) = \inf_z f(z). + \frac{1}{2\alpha}\|z - x\|^2$$

$$prox_{\alpha f}(x) = argmin_z f(z) + \frac{1}{2\alpha}\|z - x\|^2$$

Lemma 2 (Differentiability of Moreau envelope [7])**.** *Consider a ρ-weakly convex function f. Then for any $\alpha \in (0, 1/\rho)$, the Moreau envelope is C^1 smooth with gradient given by*

$$\nabla f_\alpha(x) = \frac{1}{\alpha}(x - prox_{\alpha f}(x)).$$

Definition 11 (Monotone operator)**.** *A (multivalued) operator $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is monotone if $\langle u - v, x - y\rangle \geq 0$ for all $u \in Tx$ and $v \in Ty$.*

Proposition 1 (Monotonicity of subdifferential)**.** *If $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is closed and convex, then ∂f is monotone.*

Definition 12 (Resolvent)**.** *For any operator T and $\alpha > 0$, the resolvent is $(I + \alpha T)^{-1}$. In particular, $prox_{\alpha f} = (I + \alpha \partial f)^{-1}$.*

Theorem 1. *If T is monotone, then $(I + \alpha T)^{-1}$ is monotone, single valued and 1-Lipschitz continuous on its domain. In particular, $prox_{\alpha f}$ for a convex and closed f is monotone, single-valued and 1-Lipschitz continuous.*

2.2 Optimization algorithms

In this section we present (non)convex optimization algorithms that help to inform the algorithmic approaches in later chapters. We also include at the end summary tables (Table 2.1 to Table 2.3) of convergence rates of commonly used gradient based methods.

2.2.1 Proximal gradient and PALM

Proximal gradient descent is probably the most commonly seen algorithm for additive composite nonconvex optimization. Consider the problems of the form

$$\min_x f(x) + g(x),$$

where f is usually differentiable and g is potentially nonsmooth but has relatively easy-to-find proximal mapping. Proximal gradient descent algorithm updates $x^{(k)}$ at each iteration by optimizing the local model

$$x^{(k+1)} \leftarrow \operatorname*{argmin}_x f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)}\rangle + g(x^{(k)}) + \frac{1}{2t}\|x - x^{(k)}\|^2,$$

or more compactly,

$$x^{(k+1)} \leftarrow \operatorname{prox}_{tg}(x^{(k)} - \frac{1}{t}\nabla_x f(x^{(k)})).$$

Proximal alternating linearized minimization (PALM) [10] is an extension to prox-gradient descent that allows alternating updates on blocks of variables. Roughly speaking, given blocks of variables, PALM alternates updates on each block in a Gauss-Seidel fashion using prox-gradient step. Formally, it solves the problem class

$$\min_{x,z} F(x,z) + g(x) + h(z)$$

where

- F is differentiable with respect to x, z
- g, h are be lower-semicontinuous
- $\nabla_x F, \nabla_z F$ are Lipschitz continuous with Lipschitz constants $L_1(z), L_2(x)$ respectively,

PALM (algorithm 1) converges to stationary points of problems in this class under weak assumptions. Precise conditions are detailed in the original paper [10].

Algorithm 1 PALM

Require: $x_0, z_0, \gamma_1, \gamma_2$

for $k = 0, 1, ...$ **do**

$c_k \leftarrow \gamma_1 L_1(z^{(k)})$

$x^{(k+1)} \leftarrow \text{prox}_{c_k g}(x^{(k)} - \frac{1}{c_k}\nabla_x F(x^{(k)}, z^{(k)}))$

$d_k \leftarrow \gamma_2 L_2(x^{(k)})$

$z^{(k+1)} \leftarrow \text{prox}_{d_k h}(z^{(k)} - \frac{1}{d_k}\nabla_y F(x^{(k+1)}, z^{(k)}))$

2.2.2 Composite optimization

Consider the composite problem class

$$\min_x h(c(x))$$

where c is smooth and h may be nonsmooth but usually convex [11]. c is usually required to be β-smooth, whereas for h different assumptions are imposed. For instance, in [12] h is required to be finite-valued and Lipschitz on a certain domain; in [13] h is required to be 1-Lipschitz and possibly with a sharp minimum if stronger convergence results are desired; in [6], h is assumed to be Lipschitz globally; and in [11] h is required to be "prox-regular", though not necessarily convex.

Even with these regularity conditions, the class is very broad, and covers many practical problems, including nonlinear least squares, robust phase retrieval and matrix factorization [6]. A common way to solve this type of problem is to use what is called the prox-linear method [11] [6]. In each iteration, the method updates x by optimizing a linearized local model with a proximal term

$$x^{(k+1)} \leftarrow \text{argmin}_x h(c(x^{(k)}) + \nabla c(x^{(k)})(x - x^{(k)})) + \frac{1}{2t}\|x - x^{(k)}\|^2.$$

While the term 'prox-linear' may be recent, the idea traces back to earlier works such as [12]. In particular, [12] presents an alternative construction of subproblem that uses a so-called "casting function" instead of the standard proximal term $\frac{1}{2t}\|x - x^{(k)}\|^2$. We will use this idea in designing algorithm for hybrid system application.

In the case of nonlinear least squares or problem of solving nonlinear system of equations with a penalty function, prox-linear method is related to Gauss-Newton method [6]. [13] in particular presented a modified Gauss-Newton scheme that is based on essentially the same type of local linearized model as in prox-linear method. [13] further proves global and local convergence as well as worst-case complexity bounds for convex and sharp outer function h.

We refer readers to [11] [6] and references therein for details.

2.2.3 Variable Projection

Variable projection is a technique used to reduce the dimension of optimization space. Consider the problem class

$$\min_{x \in \mathbb{R}^m, z \in \mathbb{R}^n} f(x, z)$$

where f is C^2-smooth. We can reduce the optimization space from $\mathbb{R}^{m+n}$ to $\mathbb{R}^m$ by solving z in terms of x, i.e. solve

$$\bar{z}(x) = \underset{z}{\operatorname{argmin}} f(x, z), \text{ and define } \bar{f}(x) \equiv f(x, \bar{z}(x)).$$

In other words, the variable z is "projected" out. Optimizing $f(x, z)$ is equivalent to optimizing $\bar{f}(x)$. More importantly, the magic is that

$$\nabla \bar{f}(x) = \nabla_x f(x, z)|_{z=\bar{z}(x)},$$

which means that we only need to compute the gradient of f with respect x and optimal $\bar{z}(x)$, not the Jacobian of $\bar{z}(x)$ with respect to x. Hence using variable projection does not complicate the computation of derivatives. This result is based on the implicit function theorem [14] that proves crucially the existence, uniqueness and differentiability of $\bar{z}(x)$. The theorem is rephrased and stated below.

Theorem 2 (Theorem 2 in [14])**.** *Consider minimizing a twice differentiable function $f(x, z)$ with $x \in U, z \in V$ and define*

$$\bar{f}(x) = \min_z f(x, z).$$

Suppose $\exists x^* \in U, z^* \in V$ such that $\nabla_z f(x^*, z^*) = 0$ and $\nabla_z^2 f(x^*, z^*) \succ 0$. Then there exists a twice differentiable function $\bar{z}(x) : U \to V$ such that

$$\bar{z}(x) = \underset{z}{\operatorname{argmin}} f(x, z)$$

and is a unique minimizer of $f(x, \cdot)$. Further,

$$\nabla \bar{f}(x) = \nabla_x f(x, \bar{z}(x)), \nabla_x^2 \bar{f}(x) = \nabla_x^2 f(x, \bar{z}(x)) + \nabla_x \nabla_z f(x, \bar{x}(z)) \nabla_x \bar{f}(x).$$

Recent work [15] proves a similar result for functions with a nonsmooth part. The authors consider the problem class

$$\min_{x \in \mathbb{R}^m, z \in \mathbb{R}^n} f(x, z) + r(z)$$

where $r(z)$ can be nonsmooth. The assumptions and theorem regarding this problem class is provided below. We use variable projection on the nonsmooth objective in our hybrid systems application.

Theorem 3 (Theorem 2.2 in [15]). *Suppose $\nabla_x f$ and $\nabla_z f$ exist and are Lipschitz-continuous for all (x, z), and $f(x, z)$ is strongly convex in z for all x. Define*

$$\bar{z}(x) = \operatorname*{argmin}_z f(x, z) + r(z), \bar{f}(x) \equiv f(x, \bar{z}(x)) + r(\bar{z}(x)).$$

Then

$$\nabla \bar{f}(x) = \nabla_x f(x, z)|_{z=\bar{z}(x)}.$$

2.3 Summary of convergence rates

Table 2.1 to Table 2.3 summarizes convergence rates for commonly used first-order gradient based methods. Table 2.1 lists rates on convex functions; Table 2.2 lists rates on smooth functions and includes nonconvex case; Table 2.3 lists rates on classes of nonconvex nonsmooth functions with special structures.

Across all cases, stronger convergence (better convergence rates) becomes available as we make more assumptions on the functions, such as being differentiable, convex, strongly convex or having Lipschitz gradient. If we make little to none assumptions on function properties, we can only obtain asymptotic convergence. In particular, the two problems we consider in this paper do not fall into any of the categories listed in the tables, and we show only asymptotic convergence for each case. Stronger results may be available with further analysis or on special function classes.

2.4 Kalman filtering and smoothing

Kalman filtering and smoothing is most well-known as a class of methods for noisy dynamic systems inference. The classic Kalman filters and smoothers are usually formulated as recursive equations, however as pointed out in [18] they can also be formulated from the optimization viewpoint as a maximum a posteriori (MAP) problem, providing an elegant yet flexible framework for incorporating various extensions. We give as an example the formulation of a simple linear Gaussian system below, upon which our model in the hybrid systems application is built.

	Has a nonsmooth part?	Strongly convex ?	# iterations
Gradient descent			$O(\epsilon^{-1})$
Gradient descent		✓	$O(-\log(\epsilon))$
Proximal gradient	✓		$O(\epsilon^{-1})$
FISTA [16]	✓		$O(\epsilon^{-0.5})$
Proximal gradient	✓	✓	$O(-\log(\epsilon))$

Table 2.1: Table of convergence rates to ϵ-suboptimal point of convex functions. Lipschitz gradient on the smooth part is assumed in all cases.

	Convex ?	Gradient Lipschitz?	# Iterations
Lower bound [17]		✓	$\mathrm{O}(\epsilon^{-2})$
Gradient descent		✓	$\mathrm{O}(\epsilon^{-2})$
Gradient descent	✓	✓	$\mathrm{O}(\epsilon^{-1})$
Accelerated gradient descent	✓	✓	$\mathrm{O}(\epsilon^{-0.5})$

Table 2.2: Table of convergence rates to a ϵ-stationary point of smooth functions. This table is based on [9].

	Type	Gradient Lipschitz on smooth part?	# iterations
PALM [10]	additive composite	✓	$O(\epsilon^{-1})$
Prox-linear [6]	general composite	✓	$O(\epsilon^{-2})$
Proximal subgradient [7]	weakly convex		$O(\epsilon^{-4})$

Table 2.3: Table of convergence rates to ϵ-stationary point of structured nonconvex nonsmooth functions.

Example 1 (Linear Gaussian System)**.** *Consider the system*

$$x_t = G_t x_{t-1} + w_t$$
$$y_t = H_t x_t + v_t$$

where w_t, v_t are mutually independent Gaussian noise terms with covariance matrix Q_t and R_t respectively. The first equation is usually called process model whereas the second observation model.

Using Bayes' theorem, the posterior likelihood of $\{x_t\}$ given y_t can be written as

$$\begin{aligned}\mathbb{P}(\{x_t\}|\{y_t\}) &\propto \mathbb{P}(\{y_t\}|\{x_t\})\mathbb{P}(\{x_t\}) = \Pi_{t=1}^{T}\mathbb{P}(\{v_t\})\mathbb{P}(\{w_t\}) \\ &\propto \Pi_{t=1}^{T}\exp\left(-\frac{1}{2}(y_t - H_t x_t)^T R_t^{-1}(y_t - Hx_t) - \frac{1}{2}(x_t - G_t x_{t-1})^T Q_t^{-1}(x_t - G_t x_{t-1})\right).\end{aligned}$$

Maximizing the above likelihood is equivalent to minimizing its negative log likelihood, which leads to a quadratic minimization problem

$$\min_{\{x_t\}} \sum_{t=1}^{T} \frac{1}{2}(y_t - H_t x_t)^T R_t^{-1}(y_t - H_t x_t) + \frac{1}{2}(x_t - G_t x_{t-1})^T Q_t^{-1}(x_t - G_t x_{t-1}) \tag{2.1}$$

To relate the solution of (2.1) *to that people normally obtain using classic recursive Kalman filter and smoother, let us define*

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_T \end{bmatrix}, y = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix}, u = \begin{bmatrix} x_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and

$$\mathcal{R} = diag(R_1, ..., R_t), \mathcal{Q} = diag(Q_1, ..., Q_T), \mathcal{H} = diag(H_1, ..., H_T),$$
$$\mathcal{G} = \begin{bmatrix} I & 0 & & \\ -G_2 & I & \ddots & \\ & \ddots & \ddots & 0 \\ & & -G_T & I \end{bmatrix}$$

so that (2.1) *can be rewritten as*

$$\min_x \frac{1}{2}(y - \mathcal{H}x)^T \mathcal{R}^{-1}(y - \mathcal{H}x) + \frac{1}{2}(\mathcal{G}x - u)^T \mathcal{Q}^{-1}(\mathcal{G}x - u) \tag{2.2}$$

Solving (2.2) *is equivalent to solving the normal equations*

$$(\mathcal{H}^T \mathcal{R}^{-1} \mathcal{H} + \mathcal{G}^T \mathcal{Q}^{-1} \mathcal{G})x = \mathcal{H}^T \mathcal{R}^{-1} y + \mathcal{G}^T \mathcal{Q}^{-1} u \tag{2.3}$$

by taking the gradient of the objective and setting it to 0.

Solving the linear system (2.3) *with a forward Gauss elimination followed by a backward substitution is equivalent to the standard Kalman filter and smoother respectively. A detailed steps and explanation of equivalence can be found in [18].*

We can easily extend Example 1 to account for more complicated systems. For instance, if the system is nonlinear, we can replace matrices G and H with general mapping $\mathcal{G}$ and $\mathcal{H}$ in (2.1); if the system is time-dependent, we can have G_t and H_t instead in (2.1); if the noise is non-Gaussian, we can replace the quadratic penalty with neg log likelihood of other distributions in (2.1). Regardless of the modifications, (2.1) remains to be a well structured problem that can be solved readily with existing techniques.

The optimization perspective has been particularly fruitful in dynamic systems inference. The connection between filtering and smoothing and least squares problems was made early on in the papers [19–21]. The optimization perspective has allowed several innovations, including Gauss-Newton methods for dynamic inference of nonlinear models [22], interior point methods for incorporating linear and nonlinear constraints [23], modeling errors and innovations using piecewise linear quadratic penalties [24–26], extensions of this idea to heavy-tailed models [27,28], and time-series models with singular covariances $\mathcal{Q}, \mathcal{R}$ [29]. The survey [30] provides an overview of the optimization perspective with many examples.

We consider a significant extension, where different process models $G_{t,i}$ may be describing the dynamics at any time point i. The technology developed for this extension can be applied together with other innovations, by appropriately modifying the algorithms. The thesis includes these developments for heavy tailed error models and nonlinear $\mathcal{G}$ and $\mathcal{H}$

models. Significant additional work and new approaches would be required to extend to the piecewise linear-quadratic models [26].

2.5 Portfolio selection

Portfolio optimization is an extensively studied topic in literature. The general portfolio selection problem can be formulated as follows. We are given a total of n candidate assets and a certain selection criterion, usually formulated through an objective $f(w)$. Portfolio selection models also impose bounds on w, often using the simplex:

$$\Delta_1(w) = \left\{w : 0 \leq w_i \leq 1,\ 1^T w = 1\right\}.$$

The basic portfolio optimization problem is given by

$$\min_{w \in \Delta_1} f(w) \tag{2.4}$$

with the two aforementioned objectives $f(w)$ given below.

Mean-variance(Markowitz) [31] portfolios is a very well-studied problem in this field. In its basic form, Markowitz portfolio selection is just a simple quadratic problem, however efforts have been put to include various constraints to make it more practical: [32] constrains the norm of portfolios; [33] adds constraints related to transaction cost; [34] considers tracking error constraint. We refer readers to the survey [35] for a more comprehensive discussion.

Example 2 (Mean-Variance)**.** *In mean-variance portfolio selection, also known as Markowitz selection,*

$$f(w) = w^T \Sigma w - \gamma \mu^T w \tag{2.5}$$

where $\Sigma \in \mathbb{R}^{n \times n}$ is the covariance matrix and $\mu \in \mathbb{R}^n$ the expected return vector. Quantities Σ and μ are computed from historical returns:

$$\mu = \frac{1}{N}\sum_{j=1}^{N} r_j, \quad \Sigma = \frac{1}{N}\sum_{j=1}^{N}(r_j - \mu)(r_j - \mu)^T,$$

where r_j are historical return vectors, and N is the total number of samples. The parameter γ controls the weight between variance (as a measure of risk) and return.

The objective f can be interpreted as a trade-off between risk (first term) and average return (second term). By varying the value of γ, one can obtain pairs of optimal risk and return values corresponding to each γ, which together forms a curve known as 'pareto frontier'. The word 'pareto' refers to 'pareto optimality', indicating that there exist competing metrics, in this case risk and return.

A more recent approach to portfolio selection replaces the variance by Conditional Value-at-Risk (CVaR) [36]. Intuitively, CVaR chooses portfolios with low risk of high loss, by focusing on the expected value of loss in the tail of the historical distribution. It is related to another measure for risk, known as Value-at-Risk (VaR). β-VaR is the minimum loss value such that with probability β, the loss will not exceed α. The emergence of CVaR as an alternative risk measure over VaR is due to its better mathematical characteristics, such as positive homogeneity and convexity, and theoretical properties, such as coherence [36]. We present the mathematical formulation of CVaR below, and point readers to [36] for a summary of the history of CVaR and detailed derivations.

Example 3 (Conditional Value-at-Risk)**.** *Consider a loss function $l(x, y)$ with an underlying distribution $p(y)$. Define*

$$\psi(x, \alpha) = \int_{l(x,y)\leq\alpha} p(y)dy,$$

or in words, $\psi(x, \alpha)$ to be the probability that the loss is not exceeding α at a given x. We further define the β-VaR and β-CVaR value given some x respectively as

$$\alpha_\beta(x) = \min\{\alpha : \psi(x, \alpha) \geq \beta\}$$

$$\phi_\beta(x) = (1-\beta)^{-1} \int_{l(x,y)\geq\alpha_\beta(x)} l(x, y)p(y)dy.$$

The relationship between α and β is

$$P(loss > \alpha) \leq 1 - \beta.$$

Roughly speaking, β-CVaR value $\phi_\beta(x)$ is the expected loss over the $(1-\beta)$ right tail of the distribution of y. [36] proves that instead of having the minimization problem over α in the

expression for area of integration, one can equivalently write ϕ_β *as*

$$\phi_\beta(x) = \min_\alpha \alpha + (1-\beta)^{-1} \int_y [l(x,y) - \alpha]_+ p(y) dy.$$

In the context of portfolio optimization where x *is the portfolio weight* w *and loss is* $-w^T r_j, j = 1, ..., N$, r_j *being a return vector for assets* j, *the CVaR objective can be written as*

$$f_\beta(w) = \min_\alpha \alpha + \frac{1}{N(1-\beta)} \sum_{j=1}^{N} [-w^T r_j - \alpha]_+. \tag{2.6}$$

Chapter 3

PORTFOLIO SELECTION MODELS [1]

3.1 Cardinality constrained models

In portfolio selection, it is sometimes desirable to construct a sparse portfolio. In other words, the number of assets that can be selected from a pool is limited. Problems involving such constraints are known as cardinality constrained portfolio selection. Prior art for cardinality constrained [2] portfolio selection comprises two classes of methods: heuristic algorithms such as genetic algorithms and particle swarm algorithms [37–39], and mixed integer programming [40–42]. Both lines of research have primarily focused on the Markowitz (mean-variance) criterion. One exception is [43], which projects asset returns onto a reduced space, then uses clustering to identify similar subgroups; [43] also assumes a quadratic loss function.

We propose a different approach by formulating the problem as a continuous optimization problem over a highly nonconvex set induced by the intersection of cardinality and simplex constraints. We then develop a relaxation method using auxiliary variables, and create an efficient projection map onto the nonconvex set. These innovations allow recently developed techniques for structured nonsmooth nonconvex optimization [10] to bear on the problem. The proposed approach is not limited to Markowitz portfolio selection, but it can also be applied to a variety of portfolio criteria, both smooth and nonsmooth. As a key example, we look at the nonsmooth CVaR portfolio selection problem with cardinality constraints.

---

[1] JOINT WORK WITH T. LEUNG, A. ARAVKIN.

[2] "cardinality constraints" is the term used in literature. In this paper we will use it interchangeably with "combinatorial constraints".

3.1.1 Combinatorial constraints

Combinatorial constraints restrict the number of stocks to purchase, within specified subgroups and/or across the entire portfolio. We consider the constraint set Ω given by

$$\Omega = \left\{ p_i \leq 1^T w_i \leq q_i,\ \|w_i\|_0 \leq k_i \in \mathbb{N}_+,\ i = 1, .., m \right\}, \tag{3.1}$$

where the portfolio weights w are divided into subgroups $w = [w_1, ..., w_i, ..., w_m]$, and

$$\|w_i\|_0 = \text{card}(j : w_i^j \neq 0). \tag{3.2}$$

We use w_i to denote a vector, and w_i^j to denote the jth entry of w_i. The ℓ_0 norm in (3.2) counts the number of nonzero entries of its argument. The set Ω captures a wide variety of realistic trading constraints. We can group assets by sectors, industries or other criteria. The constraints restrict both the fraction of the portfolio, as well as certain number of assets from each group. For example, we can require that the trader buy no more than 5 stocks from healthcare, between 3 and 8 stocks from tech, and that stocks in consumer goods should comprise between 10% to 25% of the total portfolio.

We now consider the constrained optimization problem

$$\min_{w \in \Omega \cap \Delta_1} f(w) \tag{3.3}$$

with Ω as in (3.1), simplex constraint Δ_1, and portfolio criterion $f(w)$, such as Markowitz (2.5) or CVaR (2.6). The inclusion of the l_0-norm in Ω means that w lies in the intersection of a highly nonconvex set and a compact convex set given by box-constraints and the 1-simplex.

To develop an efficient method for (3.3), we first relax the problem by introducing an auxiliary variable $v \in \mathbb{R}^m$ to lie in Ω, and forced to be close to $w \in \Delta_1$ using a quadratic penalty term. This type of relaxation was recently shown to be effective for nonsmooth nonconvex optimization [**?**]. The relaxed problem is given by

$$\begin{aligned} \min_{w,v} \quad & f(w) + \frac{\nu}{2}\|w - v\|^2 \\ \text{s.t.} \quad & v \in \Omega, \quad w \in \Delta_1 \end{aligned} \tag{3.4}$$

As ν increases, we have $\|w - v\| \leq \frac{C}{\nu}$ for some constant C, and problem (3.4) approximates (3.3).

3.1.2 l_1 norm vs l_0 norm

Since l_0 norm constraints are highly nonconvex in general, it is common to replace them with convex l_1 norm constraints. However, for portfolio optimization 1-norm constraints are not a useful relaxation, since $w \in \Delta_1$ forces $\|w\|_1 = 1$. Thus requiring $\|w\|_1 \leq \tau$ makes the problem infeasible for $\tau < 1$ and is meaningless for $\tau > 1$. For cardinality-constrained optimization, we need the l_0 norm for its simple and direct interpretation: $\|w\|_0 \leq k$ means exactly that we allow no more than k assets.

Even though the constraint is very nonconvex, there is a simple form for the projection onto the set $\mathbb{B}_0^k := \|w\|_0 \leq k$. Let $\omega_k(z)$ be the set of indices corresponding to the largest k entries of z (by absolute value). Then

$$\text{Proj}_{\mathbb{B}_0^k}(z)_i = \begin{cases} z_i & i \in \omega_k(z) \\ 0 & i \notin \omega_k(z). \end{cases}$$

The projection onto the entire set Ω is more complicated, and is developed in the next section.

3.1.3 Projection map

In this section we develop the projection onto Ω as defined in (3.1). First we introduce a useful lemma for this projection.

Lemma 3. *Suppose $y \in \mathbb{R}^l$. Let K be any size-k subset of $I = \{1, ..., l\}$ and $\mathcal{K}$ the union of all such Ks, so that $I - K$ denotes the complement of K in I. Let $\mathcal{C} \in \mathbb{R}_+$ be a closed convex subset of $\mathbb{R}_+ = \{x : x \geq 0\}$. Without loss of generality, we reorder y such that $y_1 \geq y_2 \geq ... \geq y_l$. We claim that the optimal K for the problem*

$$\min_{z_K \in \mathcal{C}, z_{I-K}=0, K \in \mathcal{K}} \frac{1}{2}\|y - z\|^2$$

is $K_{opt} = \{1, 2, .., k\}$*, i.e. the indices corresponding to the* k *largest components in* y*.*

Proof: The problem can be stated as

$$\begin{aligned}
&\min_{z_K \in \mathcal{C}, K \in \mathcal{K}} \frac{1}{2} \sum_{j \in K} (y_j - z_j)^2 + \frac{1}{2} \sum_{j \in I-K} y_j^2 \\
\Leftrightarrow &\min_{z_K \in \mathcal{C}, K \in \mathcal{K}} \frac{1}{2} \|y_K - z_K\|^2 + \frac{1}{2} \|y_{I-K}\|^2 \\
\Leftrightarrow &\min_{z_K \in \mathcal{C}, K \in \mathcal{K}} \frac{1}{2} \|y_K - z_K\|^2 - \frac{1}{2} \|y_K\|^2 + \frac{1}{2} \|y\|^2.
\end{aligned}$$

Note that the last term $\frac{1}{2}\|y\|^2$ does not depend on z_K, so we can focus on the first two terms, i.e.

$$\min_{z_K \in \mathcal{C}, K \in \mathcal{K}} \frac{1}{2} \|y_K - z_K\|^2 - \frac{1}{2} \|y_K\|^2.$$

Suppose there is some K' that is different from K_{opt} and denote the corresponding y as $y_{K'}$. Define $f(y)$ and $g(t)$ by

$$\begin{aligned}
f(y) &= -\frac{1}{2}\|y\|^2 + \min_{z \in \mathcal{C}} \frac{1}{2} \|y - z\|^2, \\
g(t) &= f((1-t) y_{K_{opt}} + t y_{K'}).
\end{aligned}$$

Then we have

$$\begin{aligned}
f(y_{K'}) - f(y_{K_{opt}}) &= g(1) - g(0) = \int_0^1 g'(t) dt, \\
g'(t) &= \nabla f((1-t) y_{K_{opt}} + t y_{K'})^T (-y_{K_{opt}} + y_{K'}),
\end{aligned}$$

where $\nabla f(y) = -y + y - z^* = -z^* \in -\mathcal{C}$ given that $\mathcal{C}$ is convex and $z^* = \text{argmin}_{z \in \mathcal{C}} \frac{1}{2}\|y - z\|^2$. Since $\mathcal{C} \subset \mathbb{R}_+$, $\nabla f(y)$ is nonpositive in all components. Therefore, $\nabla f((1-t) y_{K_{opt}} + t y_{K'}) \leq 0$. Further $-y_{K_{opt}} + y_{K'} \leq 0$ because $v_{K_{opt}}$ contains the k-largest components of v. As a result,

$$g'(t) \geq 0 \Rightarrow \int_0^1 g'(t) dt \geq 0 \Rightarrow f(y_{K'}) \geq f(y_{K_{opt}}).$$

This shows that K_{opt} must be the optimal choice. □

Similar projection maps are used by [44] and [45], but the key difference in our case is the intersection with a convex set $\mathcal{C} \in \mathbb{R}_+$. We also do not require y to be nonnegative.

Applying Lemma 3 to the projection of each subgroup i in (3.1), for each w_i we pick its k_i-largest components and project them onto the set

$$\left\{z \geq 0, \quad p_i \leq 1^T z \leq q_i\right\}.$$

The problem thus reduces to solving m independent projection subproblems

$$\min_{z\geq 0, \quad p_i \leq 1^T z \leq q_i} \|u_i - z\|^2, \quad z \in \mathbb{R}^{k_i} \tag{3.5}$$

where $u_i \in \mathbb{R}^k$ contains any selection of k_i-largest components in w_i. This is a quadratic problem with linear constraints and can be solved exactly. In particular we relate it to projection onto simplex via the following lemma.

Lemma 4. *Consider the problem*

$$\min_{z\geq 0, p\leq 1^T z\leq q} \|u - z\|^2, z \in \mathbb{R}^k, \quad 0 \leq p \leq q \leq 1.$$

Let m be the number of positive entries in u and z^* the minimizer. Then we have

$$1^T u_{1:m} < p \quad \Rightarrow \quad z^* = \mathit{argmin}_{z\in\Delta_p}\|u - z\|^2 \tag{3.6}$$

$$p \leq 1^T u_{1:m} \leq q \quad \Rightarrow \quad z^*_{1:m} = u_{1:m}, z^*_{m+1:k} = 0 \tag{3.7}$$

$$1^T u_{1:m} > q \quad \Rightarrow \quad z^* = \mathit{argmin}_{z\in\Delta_q}\|u - z\|^2. \tag{3.8}$$

Proof: We prove each case.

If $1^T u_{1:m} < p$, then

$$1^T z^* \geq p > 1^T u_{1:m} \geq 1^T u \Rightarrow 1^T(z^* - u) > 0,$$

which implies that the vector $z^* - u$ must have some positive entries. Let J be the index set of those entries. If $1^T z^* = p + \epsilon > p$, we can decrease $z^*_j - u_j$ for some $j \in J$ to $\max(0, z^*_j - \epsilon - u_j)$ by reducing the value of z^*_j. Thus the objective value is decreased, indicating that such a z^* is not optimal. In other words $1^T z^* = p$ must hold. Hence the problem reduces to

$$\min_{z\geq 0, 1^T z = p} \|u - z\|^2 \Leftrightarrow \min_{z\in\Delta_p} \|u - z\|^2.$$

If $p \leq 1^T u_{1:m} \leq q$,

$$\min_{z \geq 0, p \leq 1^T z \leq q} \|u - z\|^2 \geq \min_{z \geq 0} \|u - z\|^2 = \|u_{m+1:k}\|^2$$

where the equality can be achieved when $z^*_{1:m} = u_{1:m}$ and $z^*_{m+1:k} = 0$.

If $1^T u_{1:m} > q$, then

$$1^T u_{1:m} > 1^T z^* \geq 1^T z^*_{1:m} \Rightarrow u_{1:m} - z^*_{1:m} > 0,$$

which means that $u_{1:m} - z^*_{1:m}$ must have some positive entries. Let J be the index set of those entries. If $1^T z^* = q - \epsilon < q$, we can always decrease $u_j - z^*_j$ for some $j \in J$ to $\max(0, u_j - (z^*_j + \epsilon))$ by increasing the value of z^*_j. This indicates z^* is not optimal, hence $1^T z^* = q$ must hold. Then the problem reduces to

$$\min_{z \geq 0, 1^T z = q} \|u - z\|^2 \Leftrightarrow \min_{z \in \Delta_q} \|u - z\|^2.$$

□

3.1.4 Algorithms

To solve Problem (3.4) we use proximal alternating linearized minimization (PALM) [10], with alternating updates on w and v. We refer readers to the original paper as well as preliminary section for description on PALM. To use PALM on problem (3.4), $f(w)$ needs to be differentiable and ∇f Lipschitz continuous. Such conditions are satisfied for the examples considered herein.

The update on v is always a projection, whereas the update on w is problem dependent.

Mean-variance portfolio selection The relaxed problem for mean-variance (Markowitz) portfolio optimization is

$$\min_{v,w \in \Delta_1} g(v, w) := w^T(\Sigma + \lambda I)w - \gamma \mu^T w + \sum_{i=1}^{m} \frac{\nu}{2} \|w_i - v_i\|^2$$

$$\text{s.t. } v_i \geq 0, p_i \leq 1^T v_i \leq q_i, \|v_i\|_0 \leq k_i.$$

where λI is a ridge regularization term in case Σ is singular. Since the problem involves only one nonsmooth term in w (the simplex constraint), we can use proximal gradient step directly to update w, as shown in Algorithm 2.

Algorithm 2 PALM for Cardinality-Constrained Markowitz

Require: $w, v \in \mathbb{R}^n, \gamma, \lambda, \nu$

$\mathcal{V}_i = \{z \geq 0 : p_i \leq 1^T z \leq q_i, \|z\|_0 \leq k_i\}$

while not converged **do**

  $w \leftarrow \text{Proj}_{\Delta_1}(w - \delta(\nabla_w f(w) + \nu(w - v))$

  for $i = 1, ..., m$ **do**

    $v_i \leftarrow \text{Proj}_{\mathcal{V}_i}(w_i)$

return w, v

Conditional value-at-risk (CVar) portfolio selection The relaxed objective for CVaR is

$$\begin{aligned} \min_{\alpha, w \in \Delta_1, v} \quad & g(w, \alpha, v) \\ \text{s.t.} \quad & v_i \geq 0, p_i \leq 1^T v_i \leq q_i, \|v_i\|_0 \leq k_i, \end{aligned}$$

where

$$g(w, \alpha, v) = \alpha + \frac{1}{N(1-\beta)} \sum_{j=1}^{N} [-w^T r_j - \alpha]_+ + \sum_{i=1}^{m} \frac{\nu}{2} \|w_i - v_i\|^2.$$

Unlike the mean variance problem, here the problem has two nonsmooth terms in w, the simplex constraint and the hinge loss $[-w^T r_j - \alpha]_+$, which we also relax. We introduce a second auxiliary variable $u \in \mathbb{R}^N$ and let $-Rw - \alpha = u$, $R \in \mathbb{R}^{N \times n}$ is the matrix of all

samples r_j stacked together. The objective now becomes

$$\begin{aligned}\min_{\alpha,w\in\Delta_1,v,u} \quad & \alpha + \frac{1}{N(1-\beta)}\sum_{j=1}^{N}[u_j]_+ \\ & + \frac{\gamma}{2}\|Rw+\alpha\mathbf{1}+u\|^2 + \sum_{i=1}^{m}\frac{\nu}{2}\|w_i - v_i\|^2 \\ \text{s.t.} \quad & v_i \geq 0, p_i \leq 1^T v_i \leq q_i, \|v_i\|_0 \leq k_i.\end{aligned}$$

We can partially minimize in α, since the objective with respect to α is an unconstrained quadratic:

$$\alpha^*(u,w) = -\frac{1+\gamma 1^T(Rw+u)}{\gamma N}.$$

Plugging in α^*, the problem simplifies to

$$\begin{aligned}\min_{w\in\Delta_1,u,v\geq 0} \quad & \widetilde{g}(w,u,v) \\ \text{s.t.} \quad & v_i \geq 0, p_i \leq 1^T v_i \leq q_i, \|v_i\|_0 \leq k_i\end{aligned} \tag{3.9}$$

where

$$\begin{aligned}\widetilde{g}(w,u,v) =& \alpha^*(u,w) + \frac{1}{N(1-\beta)}\sum_{j=1}^{N}[u_j]_+ + \frac{\nu}{2}\|w-v\|^2 \\ & + \frac{\gamma}{2}\left\|\left(I - \frac{\mathbf{1}\mathbf{1}^T}{N}\right)(Rw+u) - \frac{1}{\gamma N}\mathbf{1}\right\|^2.\end{aligned} \tag{3.10}$$

Again we use proximal gradient step to update w and u.

Acceleration Strategies Empirically replacing proximal gradient step for w and u with momentum-based updates can help speeding up the convergence (see Algorithm 3). Theoretical guarantees for accelerated algorithms in the nonconvex setting is an open problem, but FISTA updates often improve empirical performance. The algorithm for (3.9) accelerated with FISTA updates is detailed in Algorithm (4).

In each iteration we perform one F-update for w and then one for u, followed by a projection step on v. The y and t values corresponding to both w and v are saved and used in the next iteration.

Algorithm 3 FISTA-Update

Require: t_k, y_k, x_k

$x_{k+1} \leftarrow \text{ProxGrad}(y_k)$

$t_{k+1} \leftarrow \frac{1+\sqrt{1+4t_k^2}}{2}$

$y_{k+1} \leftarrow x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k+1})$

return $(x_{k+1}, y_{k+1}, t_{k+1})$

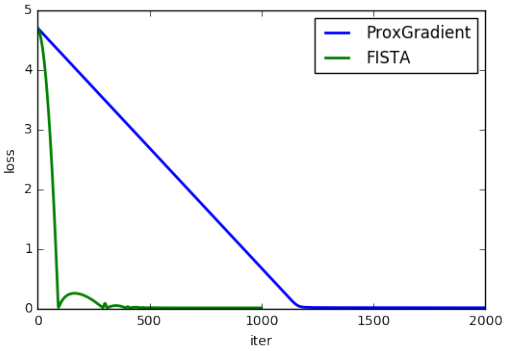


Figure 3.1: FISTA vs prox-gradient for CVaR loss $\widetilde{g}$ (3.10), with $\beta = 0.9$.

Figure 3.1 shows a comparison of FISTA updates and prox-gradient updates on w and u. The value of β is set to be 0.9 and the maximum number of stocks selected from each sector is restricted to be 5. Varying the β value and constraints yields similar plots.

When there is no sector grouping, the problem reduces to

$$\min_{w \in \Delta_1} f(w), \|w\|_0 \leq k,$$

we do not need relaxation and can solve the problem using proximal gradient descent.

Remark 1. *In some cases, using a continuation on ν can help avoid local optima. The continuation strategy adds an outer loop, increasing ν as the algorithms proceed.*

Algorithm 4 Accelerated PALM for CVaR (3.9)

Require: $w, v \in \mathbb{R}^n, \gamma$

$\mathcal{V}_i = \{z \geq 0 : p_i \leq 1^T z \leq q_i, \|z\|_0 \leq k_i\}$

while not converged **do**

$w \leftarrow$ FISTA-Update$(w; \gamma, \nu)$

$u \leftarrow$ FISTA-Update$(u; \gamma, \nu)$

for $i = 1, ..., m$ **do**

$v_i \leftarrow \text{Proj}_{\mathcal{V}_i}(w_i)$

return (w, v, u)

3.1.5 Stationarity

As mentioned earlier, the PALM algorithm converges to stationary points in nonconvex setting with weak assumptions [10]. A natural question to ask is what stationary points mean in the context of l_0-norm constraints. In this section, we consider a simple formulation

$$\min_{w \in \Delta_1^n} f(w), \quad \|w\|_0 \leq k \tag{3.11}$$

where f is a smooth loss function and Δ_1^n denotes the 1-simplex in $\mathbb{R}^n$. This problem can be solved with proximal gradient descent as discussed in previous section.

We address two questions related to stationarity:

Q_1 If we pick K of size k, $K \subset \{1, ..., n\}$ and solve

$$w^* \in \arg\min_w f(w), \quad w_K \in \Delta_1^k, w_j = 0, j \notin K \tag{3.12}$$

is w^* automatically a stationary point for (3.11)?

Q_2 If w^* is Is a stationary point of (3.11), is it always a fixed point of the proximal gradient method, as in the convex case? Or is it possible that the algorithm can move to a new stationary point with a lower objective value?

Q_1 asks whether w being a stationary point of (3.12) implies that w is also a stationary point for (3.11). To answer that, we look at stationarity conditions for both problems.

Let

$$B \equiv \{w : w_K \in \Delta_1^k, w_j = 0, j \notin K\}$$
$$C \equiv \{w \in \Delta_1^n : \|w\|_0 \leq k\}.$$

Note that $B \subset C$, and B is convex while C is nonconvex. From elementary convex analysis, if w is a stationary point of (3.12), then

$$0 \in \nabla f(w) + \partial \delta_B(w)$$

where $\partial \delta_B(w)$ denotes the subdifferential of the indicator function δ_B at w. Since B is convex, $\partial \delta_B(w)$ is the normal cone to B at w, defined by

$$N_B(\bar{w}) = \{v : \langle v, w - \bar{w} \rangle \leq 0 \ \forall w \in B\}.$$

Similarly if w is a stationary point of (3.11), then

$$0 \in \nabla f(w) + \partial \delta_C(w)$$

where $\partial \delta_C(w)$ is the limiting subdifferential of indicator function δ_C at w. For nonconvex functions, limiting subdifferential is defined in terms of Frechet subdifferentials; see the preliminaries.

We state in Lemma 5 a condition under which a stationary point of (3.12) is also a stationary point of (3.11).

Lemma 5. *For any $K \subset \{1, ..., n\}$ and its corresponding B.*

- *If $w \in intB$, then $\partial \delta_C(w) = \partial \delta_B(w)$. This implies that $0 \in \nabla f(w) + \partial \delta_B(w) \Leftrightarrow 0 \in \nabla f(w) + \partial \delta_C(w)$.*

- *If w is on the boundary of B, then $\partial \delta_C(w) \subset \partial \delta_B(w)$.*

Proof: We first need to determine $\partial\delta_C(w)$, by looking at the Frechet subdifferential. The interesting case is when $u \in B$, in which case the expression reduces to

$$\hat{\partial}\delta_C(w) = \{v : \lim_{u \neq w} \inf_{u \to w} \frac{-\langle v, u - w\rangle}{\|u - w\|} \geq 0, u \in B\}.$$

Suppose $v \in \hat{\partial}\delta_C(w)$, then

$$\begin{aligned} v \in \hat{\partial}\delta_C(w) &\Rightarrow \lim_{u \in B \to w} \inf -\langle v, u - w\rangle \geq 0 \\ &\Rightarrow -\langle v, u - \bar{w}\rangle \geq 0 \ \forall u \in B \Rightarrow v \in N_B(\bar{w}). \end{aligned}$$

Hence $\hat{\partial}\delta_C(w) \subseteq \partial\delta_B(w)$ always holds.

- $w \in \text{int}B$

  If $w \in \text{int}B$, then as $u \to w$, $u \in B$. Hence

  $$\begin{aligned} v \in N_B(w) &\Rightarrow \langle v, u - w\rangle \leq 0 \ \forall u \in B \\ &\Rightarrow \lim_{u \neq w} \inf_{u \to w} \frac{-\langle v, u - w\rangle}{\|u - w\|} \geq 0 \\ &\Rightarrow v \in \hat{\partial}\delta_C(w). \end{aligned}$$

  In other words, $\partial\delta_B(w) = \hat{\partial}\delta_C(\bar{w}) = \partial\delta_C(\bar{w})$.

- w is on the boundary of B

  Let $K' = (K \cup \{j\})\backslash\{k\}$ for some $j \notin K$ and $k \in K$, and call its corresponding set B'. Take $w \in B \cap B'$, which means $w_i = 0$ for $i \in (\{1, .., n\} - K) \cup \{k\}$ and $w_i \geq 0$ for $i \in K\backslash\{k\}$.

  We can find a vector v such that $v_i = 0$ for $i \in K$ and $v_i > 0$ for $i \notin K$. For any $u \in B$, since $u_i - w_i = 0 \ \forall i \notin K$,

  $$\langle v, u - w\rangle = \sum_{i \in K} v_i(u_i - w_i) + \sum_{i \notin K} v_i(u_i - w_i) = 0 + 0 = 0 \Rightarrow v \in \partial\delta_B(w).$$

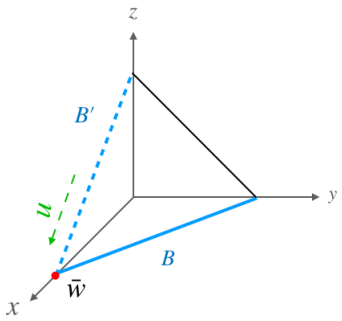


Figure 3.2: Stationary points for (3.11) and (3.12) only agree on the interior of some B; otherwise the former set is smaller. To illustrate, take $w \in \mathbb{R}^3$ and require that $\|w\|_0 \leq 2$. Let the line segment in xy plane be B and its left endpoint $\bar{w}$ (red dot). Then as $u \to \bar{w}$, it can approach along B' in addition to B. In that case there exists $u \in B'$ such that $\langle v, u - \bar{w} \rangle \geq 0$ for some $v \in N_B(\bar{w})$. Hence $v \in N_B(\bar{w})$ but $v \notin \hat{\partial}\delta_C(\bar{w})$.

On the other hand, for some $u \in B'$ such that $u_j > 0$, $u_i - w_i = 0, i \notin K'$,

$$\begin{aligned}
\langle v, u - w \rangle =& v_j(u_j - w_j) + \sum_{i \notin K'} v_i(u_i - w_i) + \sum_{i \in K' \setminus \{j\}} v_i(u_i - w_i) \\
=& v_j u_j + 0 + \sum_{i \in K \setminus \{k\}} v_i(u_i - w_i) \\
=& v_j u_j + 0 + 0 > 0
\end{aligned}$$

Hence $v \notin \partial\delta_C(w)$.

□

Lemma 5 says that if a stationary point w of (3.12) lies in the interior of any B, then w is also a stationary point of (3.11). When a w stationary for (3.12) is on the boundary, it may easily fail to be stationary for (3.11); Figure 3.2 provides a detailed illustration in $\mathbb{R}^3$.

Now suppose we solved (3.12) and found a stationary point w in the interior of some B. If we then use the obtained w as an initial guess and solve (3.11) with proximal gradient descent, would we be trapped at w given it is stationary? In other words, is a stationary

point necessarily a fixed point for proximal gradient descent? This is what Q2 is asking. The answer is true for convex functions [46, Proposition 3.1], but not for nonconvex functions. In the convex case, we have the following Lemma.

Lemma 6. *Consider the problem of*

$$\min_x f(x) + g(x)$$

where $f(x)$ is smooth and $g(x)$ is convex but nonsmooth. Then for any $\alpha > 0$,

$$0 \in \nabla f(x) + \partial g(x) \Leftrightarrow x = prox_{\alpha g}(x - \alpha \nabla f(x)).$$

Proof: The preliminaries (monotonicity of the subdifferential, definition of resolvent, and Theorem 1) imply that the proximal operator for convex function is single valued. We can argue that stationary points and fixed points are equivalent in the convex case. The arguments below are standard, see e.g. [46, Proposition 3.1].

$$\begin{aligned} 0 \in \nabla g(x) + \partial f(x) \Leftrightarrow & -\nabla f(x) \in \partial g(x) \\ \Leftrightarrow & (x - \alpha \nabla f(x)) \in x + \alpha \partial g(x) \\ \Leftrightarrow & (x - \alpha \nabla f(x)) \in (I + \alpha \partial g)(x) \\ \Leftrightarrow & x = \text{prox}_{\alpha g}(x - \alpha \nabla f(x)) \end{aligned}$$

since $(I + \alpha \partial g)^{-1}$ is single valued. This proof would not work more generally (in the nonconvex case), since the proximal operator may not be single valued. □

When g is not convex, e.g. when $g = \delta_C$, then $(I + \alpha \partial g)^{-1}$ can be multi-valued, which means that $x \in \text{prox}_{\alpha g}(x - \alpha \nabla f(x))$. In that case x is not necessarily a fixed point and we may move to some other points with smaller objective values. Empirically we see that 'bad' stationary points are seldom fixed points, and the algorithm will move past these points to stationary points with lower objective values.

To demonstrate the above claim, we conducted an experiment with the following procedure:

Markowitz (n/k)	1	2	3	4
15	0	0.75	0.9	1
30	0	0.72	0.92	0.98
45	0	0.71	0.85	0.95
CVaR (n/k)	1	2	3	4
15	0	0.31	0.67	0.8
30	0	0.3	0.56	0.75
45	0	0.2	0.42	0.64

Table 3.1: Fraction of times algorithm decreases objective when starting at a stationary point.

1. Randomly pick a subset K of size k from $\{1, ..., n\}$
2. Solve the problem $\min_w f(w), w_K \in \Delta_1^k, w_j = 0, j \notin K$ and denote the minimizer as w_{init}
3. Use w_{init} as an initial guess for (3.11) and run our algorithm
4. Check if we find another w with lower objective value

We repeat this procedure with varying k and n. For each k, n pair we run 100 trials and record the percentage of times we find a lower objective value; the results are presented in Table 3.1. As n increase we are more likely to stay at w_{init}, while as k increases we are more likely to move to a point with a lower objective value. When $k = 1$ we always stay at w_{init}, but in this case we can simply do a linear scan over all assets to find the best one. For the mean-variance model, the percentage is high in all cases except $k = 1$. The study suggests that for moderate k we can expect the algorithm to easily move past a spurious stationary point.

3.1.6 Numerical results

In this section we present results from two sets of experiments. The first set of experiments compares our approach against the brute force searches on small examples where a brute force search is feasible. Since gradient based methods only guarantees convergence to local optima for nonconvex problems, this experiment gives a sense of how good those local optima are, at least for small cases. Brute-force search is impossible for even moderate-sized problems.

The second set of experiments compares Pareto efficient frontiers of models with and without cardinality constraints. For Markowitz models, efficient frontiers show the trade-off between portfolio variance against average return; for CVaR we plot β-VaR against β-CVaR values.

For all experiments, the underlying dataset comprises closing stock prices for 65 stocks from June 21st 2017 to June 21st 2018, taken from Yahoo finance. The stocks belong to 7 sectors, basic metals, consumer goods, finance, health care, industrial goods, service and technology.

Comparison with brute force search Since the problem is highly nonconvex, our method is not guaranteed to find a global minimum. For small simple examples, where exhaustive search is feasible, we can compare the quality of our solution (using function value) to that of the global minimum. Specifically, we do not impose sector groupings on w (since there is no straightforward way to implement a brute force search in this case) and limit the total number of candidate assets. Figure 3.3 summarizes the results. The top row shows the histograms of losses for mean-variance portfolio (with $\gamma = 0.1$) when we look for at most 5 from a total of 10, 15 or 20 assets (left, middle and right panels). The red bar shows the objective value of our solution. The bottom row shows the corresponding histograms for CVaR with $\beta = 0.9$. In the experiment, our solution was a true global minimum in all three simulations for mean-variance portfolios, and the CVaR solutions were nearly optimal in function value.

As the total number of assets grows, exhaustive search quickly becomes prohibitively

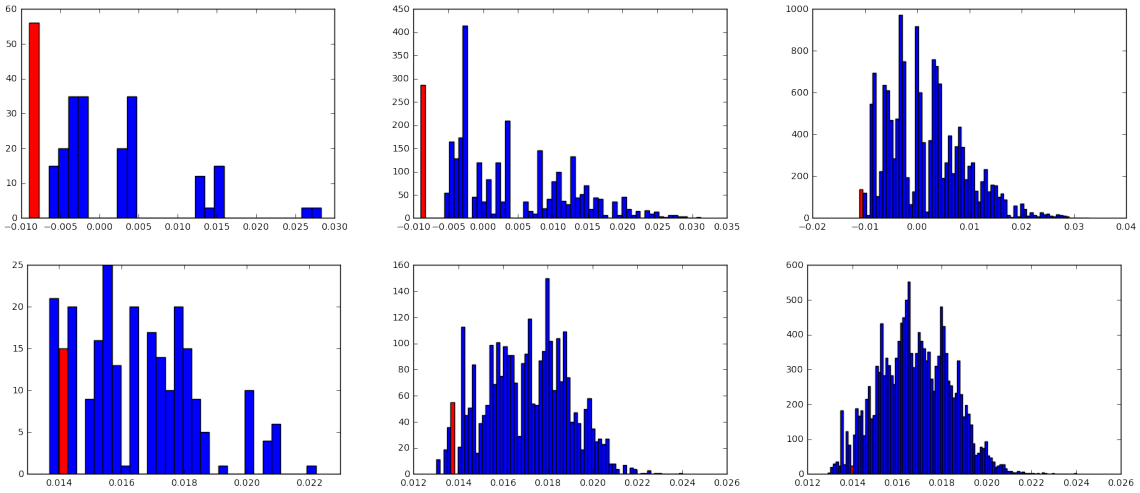


Figure 3.3: Histograms of losses. Each plot shows the histogram of loss obtained via exhaustive search and red bar indicates where result from our method lies. Top row: mean-variance; Bottom row: CVaR. Left: 10 choose 5. Middle: 15 choose 5; Right: 20 choose 5. For mean-variance we found global minima in all three cases; for CVaR we found local minima that are close to global minima.

Markowitz	avg time	min time	max time
Proposed method	0.022	0.020	0.030
Randomized brute force	1.77	0.018	17.39
CVaR	avg time	min time	max time
Proposed method	2.03	1.80	2.64
Randomized brute force	71.08	1.83	222.69

Table 3.2: Proposed method vs. randomized brute force search: time (seconds) for choosing 10 assets from 30 (across 20 trials).

expensive. For instance, choosing 10 assets out of 30 requires solving more than 30 million optimization problems over the subsets. Alternative approaches include using mixed integer quadratic programming. To test our method for these larger scenarios, we first run our algorithm to obtain a solution; then we randomly choose asset combinations, solve a minimization problem over that combination, and repeat until we find a solution with same or smaller function value. We compare the average elapsed time using our method versus brute force search. Table 3.2 shows the results. The task is to choose 10 out of 30 assets and the average is taken over 20 runs. Overall, the randomized brute force search takes a very long time to match the quality of the solution (measured by function value) that is quickly discovered by continuous optimization.

Efficient Frontiers We plot the Pareto efficient frontier for classic portfolio optimization strategies against those with different stock cardinalities specified by the cardinality constraints. For Markowitz models, an efficient frontier is traced out by varying γ, the weight on average return, while for CVaR it is traced out by varying the probability level β.

Figure 3.4 shows the plot of portfolio variance vs. portfolio return in a Markowitz model as γ varies from 0 to 1.5. The blue line shows the Pareto curve without cardinality constraints;

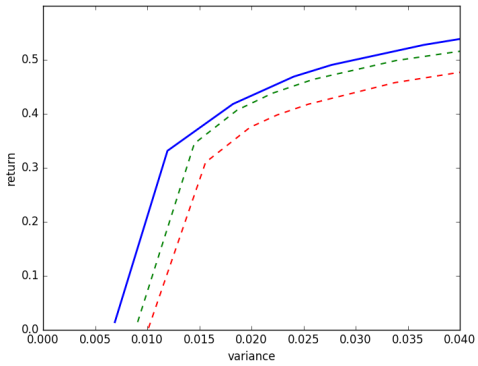


Figure 3.4: Variance v. return for unconstrained (solid line) and cardinality-constrained (dashed lines) Markowitz portfolios as γ varies. The green dashed uses a model that excludes industrial goods and allows at most 2 stocks from other sectors. The red dashed line uses a model that excludes industrial goods and services, and allows at most 2 stocks from other sectors.

and the dashed lines show the effects of the constrains on the frontier. In particular, the green dashed line is the Pareto frontier using a model that excludes industrial goods and allows at most 2 stocks from other sectors. The red dashed line is the Pareto frontier using a model that excludes industrial goods and services, and allows at most 2 stocks from other sectors. Progressively more stringent constraints shift us to less efficient regimes; however, the loss in efficiency relative to an unconstrained model is minor (particularly for the first model) and can be quantified using the Pareto approach.

Figure 3.5 shows the plot of β-VaR and β-CVaR values as β varies from 0.5 to 0.95. β-VaR corresponds to α value in the objective and β-CVaR corresponds to $\phi_\beta = F_\beta(\cdot, \alpha_\beta)$ [36] where

$$F_\beta(w, \alpha) = \alpha + \frac{1}{N(1-\beta)} \sum_{j=1}^{N} [-w^T r_j - \alpha]_+.$$

Recall that α is the value such that $P(\text{loss} > \alpha) \leq 1 - \beta$ and ϕ_β is the expected loss given loss $\geq \alpha$. In Figure 3.5 blue line plots the relation without any constraints; green and red dashed lines correspond to the same constraints as for Figure 3.4. When constraints are

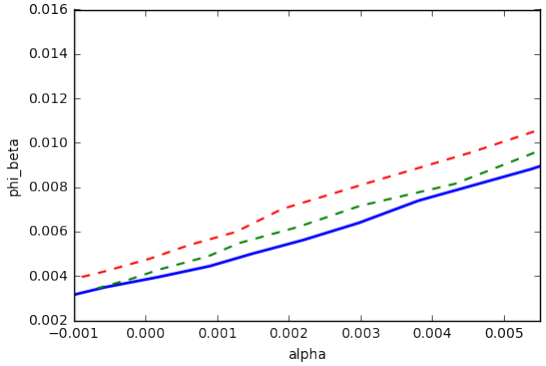


Figure 3.5: β-VaR (α) v. β-CVaR (ϕ_β) for unconstrained (solid line) and cardinality constrained (dashed lines) CVaR portfolios as β varies. The green dashed uses a model that excludes industrial goods and allows at most 2 stocks from other sectors. The red dashed line uses a model that excludes industrial goods and services, and allows at most 2 stocks from other sectors.

imposed, at a given α the value of ϕ_β increases, meaning that the expected tail loss increases. Analogously to the Markowitz case, the increase in tail-loss is small relative what is possible in the unconstrained case, and can be quantified using this approach. These results allow fund manages and investors to incorporate realistic constraints transaction costs, minimum lot sizes, execution efficiency, and investor preferences directly into portfolio optimization, and to evaluate the resulting portfolios against idealized settings.

3.2 Mean reverting portfolio selection

Mean reversion trading is a major class of trading strategies used by professional traders and fund managers. The strategy typically involves a portfolio of positions in two or more highly cointegrated assets (with a strong financial or economic relationship that prevents them from diverging), such as stocks and exchange-traded funds (ETFs), or derivatives, such as futures, across many asset classes. The challenge is to systematically construct a portfolio whose value over time exhibit mean-reverting behaviors. Once such a portfolio is

identified, then the pattern can be exploited by traders and the estimated parameters can inform the optimal trading strategies, such as those developed in [47]. There are also a number of studies on trading mean-reverting prices [48, 49] and the empirical performance of pairs trading [50].

Previous works on mean-reverting portfolio design have used different empirical proxies for mean reversion and can usually be converted into semi-definite programming problems (see e.g. [51], [52], [53]). In this paper, we instead consider using an Ornstein-Uhlenbeck (OU) process [54] as a measure for mean reversion. Given an arbitrary set of assets with their price histories, our main goal is to design a mean-reverting portfolio whose evolution over time can be characterized by an OU process [54] through penalized maximum likelihood estimation (MLE). A major feature of our joint optimization approach is that we simultaneously solve for the optimal portfolio and the corresponding parameters for maximum likelihood using gradient-based method. This unified approach is different from prior work since a) it does not rely on SDP, and b) we do not break the problem up into stage-wise computations. For example, [53] first determines optimal weights using mean-reversion proxies other than OU, and then fit the resulting portfolio to an OU process. Since our formulation is based on MLE of OU and does not involve other proxies, it is more natural in our case to perform simultaneous optimization than a two-stage procedure. Conversely, [47] fits an OU process to each of a range of candidate (pair) portfolios, and takes the candidate with the highest OU likelihood. Our unified approach looks for the best OU-representable portfolio from a set of candidates, making the quality of the OU fit part of the optimization problem.

We first present the maximum likelihood formulation for simultaneously selecting a portfolio from a set of assets, and representing that selection using an Ornstein-Uhlenbeck (OU) process. We also make several theoretical observations about the well-posedness of the estimation problem. We then extend the maximum likelihood formulation to allow selection of higher mean reversion and parsimony in the portfolio.

3.2.1 OU MLE via optimization

We are given historical data for m assets, with $S^{(T+1)\times m}$ the matrix for assets values over time. Our first goal is to find w, the linear combination of assets that comprise our portfolio, such that the corresponding portfolio price process $x_t := S_t w$ best follows an OU process. We first show that solving for the portfolio with the optimal OU likelihood leads to the optimization problem

$$\min_{a,c,\theta,\|w\|_1=1} \frac{1}{2}\ln(a) + \frac{1}{2Ta}\|\mathcal{A}(c)w - \theta(1-c)\mathbf{1}\|^2, \tag{3.13}$$

where $\mathcal{A}(c) = S_{1:T} - cS_{0:T-1}$, w is the portfolio to be selected, and a, c, θ are likelihood parameters. The objective function is nonconvex, since $\mathcal{A}(c)$ multiplies w, and also includes a nonconvex constraint $\|w\|_1 = 1$. The 1-norm constraint limits both long and short positions. We are primarily interested in the relative not the absolute magnitude of w_i's. The portfolio weights w_i's and thus value of the constraint (i.e. 1 on the right-hand side) can be scaled, and our method can still be applied (see remark 6). The derivations of problem (3.13) are presented below.

An OU process is defined by the SDE

$$dx_t = \mu(\theta - x_t)dt + \sigma dB_t, \tag{3.14}$$

where B_t is a standard Brownian motion under the physical probability measure. The likelihood of an OU process observed over a sequence $\{x_t\}_{t=1}^T$ is given by

$$\prod_{t=1}^{T} f(x_t|x_{t-1}) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \times \exp\left(-\frac{(x_t - x_{t-1}\exp(-\Delta t\mu) - \theta(1-\exp(-\Delta t\mu)))^2}{2\tilde{\sigma}^2}\right)$$

where $\tilde{\sigma}^2 = \sigma^2 \frac{1-\exp(-\Delta t\mu)^2}{2\mu}$. Minimizing the negative log-likelihood results in the optimization problem

$$\min_{\mu,\sigma^2,\theta,w} \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln(\tilde{\sigma}^2(\mu,\sigma^2)) + \frac{\|A(\mu)w - y(\theta,\mu)\|^2}{2T\tilde{\sigma}^2(\mu,\sigma^2)}, \tag{3.15}$$

with $y = \theta(1 - \exp(-\Delta t\mu))\mathbf{1}$, and $A(\mu) \in \mathbb{R}^{T\times m}$ defined as

$$A(\mu) := S_{1:T} - \exp(-\Delta t\mu)S_{0:T-1},$$

where the subscripts denote ranges for t.

Remark 2. *The objective function in (3.15) is unbounded. Set $w = 0, \theta = 0$; the objective function is then given by*

$$\frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln(\sigma^2) + \frac{1}{2}\ln\left(\frac{1 - \exp(-2\mu\Delta t)}{2\mu}\right),$$

which goes to $-\infty$ as $\sigma^2 \to 0$.

To solve the issue exposed in Remark 2, we add a 1-norm equality constraint on w, setting $\|w\|_1 = 1$. This constraint is also convenient from a modeling perspective, as it eliminates the need to select which assets in the portfolio are to be long or short *a priori*.

To obtain formulation (3.13), we denote

$$a = \tilde{\sigma}^2 = \frac{\sigma^2(1 - \exp(-2\Delta t\mu))}{2\mu}, c = \exp(-\Delta t\mu). \tag{3.16}$$

Applying the linear approximation $e^x \approx 1 + x$ to (3.16), we obtain simplified expressions for a and c:

$$a = \Delta t\sigma^2, \quad c = 1 - \Delta t\mu. \tag{3.17}$$

We can recover μ and σ^2 once we know a and c.

Remark 3. *The term $\frac{1}{2}\ln(2\pi)$ is dropped from the objective as it is simply a constant. In the subsequent sections when we mention negative log likelihood it refers to value without this constant term.*

Promoting Sparsity and Mean Reversion Given a set of candidate assets, we want to select a small parsimonious subset to build a portfolio. To add this feature to the model, we want to impose a sparsity penalty on w. While the 1-norm is frequently used, in our case we have already imposed the 1-norm equality constraint $\|w\|_1 = 1$. To obtain sparse

solutions under this constraint, we add a multiple of the *nonconvex* constraint $\|w\|_0 \leq \eta$ to the maximum likelihood (3.13). This constraint limits the maximum number of assets to be η.

In addition to sparsifying the solution, we may also want to promote other features of the portfolio. The penalized likelihood framework is flexible enough to allow these enhancements. An important feature is encapsulated by the mean-reverting coefficient μ; a higher μ may be desirable for trading, where positions are opened when deviations are observed, and closed when the portfolio returns to the mean. We can obtain a higher μ by promoting a lower c, e.g. with a linear penalty on $c = 1 - \Delta t\mu$ with a constant penalization coefficient γ. The augmented likelihood function is

$$\min_{a,c,\theta,\|w\|_1=1,\|w\|_0\leq\eta} \frac{\ln(a)}{2} + \frac{\|A(c)w - \theta(1-c)\mathbf{1}\|^2}{2Ta} + \gamma c. \tag{3.18}$$

A higher γ drives c to be lower, and hence drives μ higher.

3.2.2 Algorithm

We develop an algorithm to solve the nonsmooth, nonconvex problem (3.18) by exploiting its rich structure. We define the following nested value functions:

$$\begin{aligned} f(w,a,c,\theta) &= \frac{\ln(a)}{2} + \gamma c + \frac{\|A(c)w - \theta(1-c)\mathbf{1}\|^2}{2Ta} \\ f_1(w,a,c) &= \min_{\theta} f(w,a,c,\theta) \\ f_2(w,a) &= \min_{c} f_1(w,a,c) = \min_{c,\theta} f(w,a,c,\theta) \\ f_3(w) &= \min_{a} f_2(w,a) = \min_{a,c,\theta} f(w,a,c,\theta). \end{aligned} \tag{3.19}$$

In other words we project out variables a, c, θ. This technique is known as variable projection, or partial minimization. Our main strategy is to use these value functions to recast (3.18) as the optimization problem

$$\min_{\|w\|_1=1,\|w\|_0\leq\eta} f_3(w), \tag{3.20}$$

and solve it using projected gradient descent as detailed in Algorithm 5.

To prove Algorithm 5 converges for (3.20) requires several steps. First, we establish the differentiability of f_3 and Lipschitz continuity of its gradient on region bounded away from the origin in Theorem 4. Second, we use projection map developed in Section 3.1.3. Finally we develop the convergence analysis in Theorem 5.

Differentiability of $f_3(w)$ and Lipschitz continuity of $\nabla f_3(w)$. We first make an assumption on the input data S: for any $\|w\|_2 \geq \epsilon$, we assume that

$$\|Bx(w)_{0:T-1}\|_2 \geq \delta > 0 \tag{3.21}$$

where $x = Sw$ and $B = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{T} \in \mathbb{R}^{T\times T}$. If $\|Bx(w)_{0:T-1}\|_2 = 0$ for some w, that implies

$$\exists w,\ x(w)_{0:T-1} = \frac{\mathbf{1}^T x(w)_{0:T-1}}{T}\mathbf{1}, \tag{3.22}$$

but this is a linear system with m (the number of assets) unknowns and T equations, where T usually is much larger than m. Intuitively, (3.22) says that the portfolio value $x(w)$ must be constant over time and exactly equal to its mean, which is very unlikely with stock market data. Hence assumption (3.21) is reasonable.

We now state the theorem.

Theorem 4. *Consider $w \in \{w : \|w\|_2 \geq \epsilon\}$. Problem (3.18) is equivalent to*

$$\min_{\|w\|_1=1, \|w\|_0\leq\eta} f_3(w)$$

where $f_3(w)$ is a differentiable function for small enough γ and ∇f_3 is Lipschitz continuous.

Proof: We start by deriving an explicit expression for the f_1 value function. Taking $\partial_\theta f = 0$, we get

$$\begin{aligned} 0 = \frac{\partial f}{\partial \theta} &= (1-c)\mathbf{1}^T(\theta(1-c)\mathbf{1} - A(c)w) \\ \Rightarrow \theta^*(c,w) &= \frac{\mathbf{1}^T(x(w)_{1:T} - cx(w)_{0:T-1})}{T(1-c)}. \end{aligned}$$

Plugging $\theta^*(c,w)$ into f, we get an explicit form of f_1:

$$f_1(w,a,c) = \frac{1}{2}\ln(a) + \gamma c + \frac{\|B(x(w)_{1:T} - cx(w)_{0:T-1})\|^2}{2Ta} \tag{3.23}$$

with $B = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{T}$ a projection matrix onto the space of vectors in $\mathbb{R}^T$ with mean 0. To simplify the following analysis, we define

$$b_1(w) := Bx(w)_{1:T}, \quad b_0(w) := Bx(w)_{0:T-1}.$$

We now apply a differential variant of the implicit function theorem to f_1. Let $F(w, y)$ be f_1 in (3.23) where $y = [a, c]$, so that $f_3(w) = \min_y F(w, y)$. From [14, Theorem 2], if there exists $\bar{w}, \bar{y}$ such that $F_y(\bar{w}, \bar{y}) = 0$ and $F_{yy}(\bar{w}, \bar{y})$ is positive definite, then in the neighborhood of $(\bar{w}, \bar{y})$ where $f_3(w)$ is defined, it is twice differentiable. In our case,

$$F_y(w, y) = \begin{bmatrix} \frac{1}{2a} - \frac{\|b_1 - cb_0(w)\|^2}{2Ta^2} \\ \gamma - \frac{1}{Ta} b_0(w)^T (b_1(w) - cb_0(w)) \end{bmatrix}$$

$$F_{yy}(w, y) = \begin{bmatrix} -\frac{1}{2a^2} + \frac{\|b_1(w) - cb_0(w)\|^2}{Ta^3} & \frac{b_0(w)^T (b_1(w) - cb0(w))}{Ta^2} \\ \frac{b_0(w)^T (b_1(w) - cb_0(w))}{Ta^2} & \frac{1}{Ta} b_0(w)^T b_0(w) \end{bmatrix}.$$

When $F_y(\bar{w}, \bar{y}) = 0$, we have

$$\bar{a}(\bar{w}) = \frac{1}{T} \|b_1(\bar{w}) - \bar{c} b_0(\bar{w})\|^2,$$
$$\gamma(\bar{w}) = \frac{1}{T\bar{a}} b_0(\bar{w})^T (b_1(\bar{w}) - \bar{c} b_0(\bar{w})),$$

and F_{yy} simplifies to

$$F_{yy}(\bar{w}, \bar{y}) = \begin{bmatrix} \frac{1}{2\bar{a}(\bar{w})^2} & \frac{\gamma}{\bar{a}(\bar{w})} \\ \frac{\gamma(\bar{w})}{\bar{a}(\bar{w})} & \frac{1}{T\bar{a}(\bar{w})} b_0(\bar{w})^T b_0(\bar{w}) \end{bmatrix}.$$

Given assumption in (3.21), when $\gamma = 0$, we immediately have that $F_{yy}(\bar{w}, \bar{y})$ is diagonal with positive entries. If $\gamma > 0$, we write $\bar{a}$ in terms of $\bar{w}$ by solving $F_y(\bar{w}, \bar{y}) = 0$ and

$$\bar{a} = \frac{\|b_0\|^2}{2T\gamma^2} - \frac{\sqrt{\|b_0\|^4 - 4\gamma^2(\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2)}}{2T\gamma^2}$$

where b_0, b_1 are evaluated at $\bar{w}$.

When $\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2 \neq 0$, in order for $\bar{a}$ to be a real number, γ has to be small enough so that

$$\|b_0\|^4 - 4\gamma^2(\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2) \geq 0 \ \forall \|w\|_2 \geq \epsilon$$
$$\Rightarrow 0 \leq \gamma \leq \inf_{\|w\|_2 \geq \epsilon} \frac{1}{2} \sqrt{\frac{\|b_0\|^4}{\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2}}.$$

The infimum can be attained because $\|b_0\|^2$ is bounded below by the assumption on input data and $\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2 \leq \|b_0\|^2\|b_1\|^2$ is bounded above.

Thus the determinant of $F_{yy}(\bar{w}, \bar{y})$ is

$$\det(F_{yy}(\bar{w}, \bar{y})) = \frac{\|b_0\|^2 - 2T\bar{a}\gamma^2}{2T\bar{a}^3} > 0$$

using the expression for $\bar{a}$. Since $F_{yy}(\bar{w}, \bar{y})$ is a 2×2 matrix with a positive first minorant and positive determinant, it must be positive definite. Hence the conditions in Theorem 2, [14] are satisfied, implying that f_3 is twice differentiable on $\{w : \|w\|_2 \geq \epsilon\}$. Moreover, the eigenvalues of F_{yy} depend continuously on w, which is restricted to a compact set $\mathcal{W}$. Hence the operator norm of F_{yy} has an upper bound for all $w \in \mathcal{W}$, and this value is also a Lipschitz constant for $\nabla f(w)$.

□

Remark 4. *The expression for $\bar{c}$ is*

$$\bar{c} = \frac{b_0^T b_1 - T\bar{a}\gamma}{\|b_0\|^2}.$$

There is no guarantee that $\bar{c}$ is positive. Indeed $\bar{c}$ can potentially be negative, in which case no corresponding positive μ exists. This means that the given data and γ do not permit the construction of a mean-reverting time series. The γ term in numerator drives $\bar{c}$ towards negative values, which means that the higher mean-reverting level we request, the less likely such a process can be constructed.

Remark 5. *When $\gamma > 0$, $f_3(w)$ is given by*

$$f_3(w) = \frac{1}{2}\ln(\bar{a}) + \frac{\|b_1\|^2}{2T\bar{a}} - \frac{(b_0^T b_1)^2}{2T\bar{a}\|b_0\|^2} - \frac{T\bar{a}\gamma^2}{2\|b_0\|^2} + \frac{\gamma b_0^T b_1}{\|b_0\|^2}.$$

When $\gamma = 0$, $f_3(w)$ simplifies to

$$f_3(w) = \frac{1}{2}\ln(\bar{a}) + 1/2.$$

In both expressions, $\bar{a}, b_0, b_1$ are evaluated at w as in the proof of Theorem 4.

Remark 6. *If we scale w to Kw, then*

$$\begin{aligned} f_3(Kw) =& \frac{1}{2}\ln(K^2\bar{a}) + \frac{K^2\|b_1\|^2}{2TK^2\bar{a}} - \frac{K^4(b_0^T b_1)^2}{2TK^4\bar{a}\|b_0\|^2} \\ & - \frac{TK^2\bar{a}\gamma^2}{2K^2\|b_0\|^2} + \frac{K^2\gamma b_0^T b_1}{K^2\|b_0\|^2} \\ =& \ln(K) + f_3(w). \end{aligned}$$

Let $v = Kw$. Then

$$\min_{\|v\|_1=K, \|v\|_0\le\eta} f_3(v) \Leftrightarrow \min_{\|w\|_1=1, \|w\|_0\le\eta} f_3(w).$$

Algorithm 5 Projected Gradient Descent for $f_3(w;\gamma,\eta)$ (3.19).

Require: $w \in \mathbb{R}^m, S, f_3, \gamma, \eta$

1: $\mathcal{W} = \{w : \|w\|_1 = 1, \|w\|_0 \le \eta\}$

2: **while** not converged **do**

3: $w^k \leftarrow \text{Proj}_{\mathcal{W}}(w^{k-1} - \delta_i \nabla_w f_3(w^{k-1};\gamma,\eta))$
Recover a, c, θ from w.

4: (δ_i denotes stepsize via line search.)

3.2.3 Convergence analysis

Algorithm 5 is projected gradient descent for the value function f_3 over the nonconvex set $\mathcal{W}$, which converges for a large class of nonconvex functions [55]. However our problem does not satisfy the assumptions of [55] because the gradient of loss function $f_3(w)$ is not globally Lipschitz. As shown in the previous section, when w is bounded away from the origin, the gradient is Lipschitz; when w approaches the origin, however, the function value goes to ∞ and the gradient is not Lipschitz. Figure 3.6 shows a schematic plot of the loss function f_3. Global Lipschitz of gradient is used to establish sufficient decrease in the loss, a key component of any convergence theory. We derive Lemma 7 to establish sufficient decrease of f_3, taking advantage of the fact that $\mathcal{W}$ is bounded away from the origin. We also include

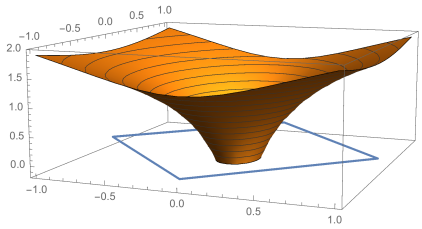


Figure 3.6: 3D plot of the objective function in (3.20) for $w \in \mathbb{R}^2$, with constraint set $\|w\|_1 = 1, \|w\|_0 \leq 2$. Our goal is to find the minimum value of f_3 (yellow 3D plot) restricted to $\mathcal{W}$ (edges of the blue diamond).

additional lemmas to provide a full picture of the analysis. The main result is presented in Theorem 5.

Theorem 5. *Consider the optimization problem*

$$\min_{w \in \mathcal{W}} f(w),$$

where f is the objective function f_3 in (3.20) and $\mathcal{W} = \{w : \|w\|_1 = 1, \|w\|_0 \leq \eta\}$ is the nonconvex constraint set in (3.20). In particular, f is nonconvex and is not smooth and has singularities near the origin.

Let $\{w^k\}$ be the sequence generated by the line search $w^+ \leftarrow \Pi_{\mathcal{W}}(w - t\nabla f(w))$ with $t \geq \underline{t}$, then

$$\nabla f(w^k) + \partial \delta_{\mathcal{W}}(w^k) \to 0$$

as $k \to \infty$. Here $\Pi_{\mathcal{W}}$ denotes the projection onto $\mathcal{W}, \underline{t}$ a lower bound on t, $\delta_{\mathcal{W}}(z) = \begin{cases} 0 & z \in \mathcal{W} \\ \infty & o.w. \end{cases}$, and $\partial\delta$ denotes the limiting subdifferential, which is the appropriate generalization of derivative for this situation; see e.g. [56].

Proof: This theorem is proved using the following lemmas (detailed below):

- Lemma 7 relates decrease in function values $f(w) - f(w^+)$ to consecutive differences

$\|w^+ - w\|^2$, using Lipschitz continuity of ∇f on the set

$$C = \mathbb{R}^n - \mathbb{B}_\epsilon(0) \supset \mathcal{W},$$

where $\mathbb{B}_\epsilon(0) = \{w : \|w\|_2 < \epsilon\}$ and $\epsilon \leq \sqrt{2}/2$.

- Lemma 8 uses Lemma 7 to show that $\|w^{k+1} - w^k\| \downarrow 0$.

- Lemma 9 shows that elements in the subdifferential $\nabla f + \delta_{\mathcal{W}}$ converge to 0 using Lemma 8.

Lemma 7. *Let $C = \mathbb{R}^n - \mathbb{B}_\epsilon(0)$ where $\mathbb{B}_\epsilon(0) = \{w : \|w\|_2 < \epsilon\}$ and $\epsilon \leq \sqrt{2}/2$. In other words, $\mathbb{B}_\epsilon(0)$ is inside the 1-norm sphere $\{w : \|w\|_1 = 1\}$. Let $L(\epsilon)$ be the upper bound such that*

$$\|\nabla f(w) - \nabla f(w')\| \leq L(\epsilon)\|w - w'\| \ \forall w, w' \ \in C.$$

Suppose $w \in \mathcal{W}$, and let $w^+ \leftarrow \Pi_{\mathcal{W}}(w - t\nabla f(w))$. Then we have

$$f(w^+) \leq f(w) - \frac{1/t - 15L(\epsilon)}{2}\|w^+ - w\|^2.$$

Proof: If the line segment from w to w^+ does not go through $\mathbb{B}_\epsilon$, then by $L(\epsilon)$-Lipschitz,

$$f(w^+) \leq f(w) + \langle w^+ - w, \nabla f(w)\rangle + \frac{L(\epsilon)}{2}\|w^+ - w\|^2.$$

Otherwise, let w_1, w_4 denote the intersection of the line segment with the closed ball $\bar{\mathbb{B}}_\epsilon$ and $w_0 = w, w_5 = w^+$. We can find a 2D circle centered at the origin with diameter 2ϵ that passes through w_1, w_4. Then we can find a tight box with length 2ϵ that contains the circle. Let w_2, w_3 be two vertices on the box, through which we can define a path from w_1 to w_4 along the box. This path does not go through $\mathbb{B}_\epsilon$.

By $L(\epsilon)$-Lipschitz of f on C,

$$\begin{aligned}
&f(w_{i+1}) \leq f(w_i) + \langle w_{i+1} - w_i, \nabla f(w_i)\rangle \\
&+ \frac{L(\epsilon)}{2}\|w_{i+1} - w_i\|^2 \\
\Rightarrow &f(w^+) \leq \sum_{i=0}^{4}\langle w_{i+1} - w_i, \nabla f(w_i)\rangle + \frac{L(\epsilon)}{2}\|w_{i+1} - w_i\|^2 \\
&+ f(w) \\
\Rightarrow &f(w^+) \leq f(w) + \langle w^+ - w, \nabla f(w)\rangle \\
&+ \sum_{i=0}^{4}\langle w_{i+1} - w_i, \nabla f(w_i) - \nabla f(w)\rangle + \frac{L(\epsilon)}{2}\|w_{i+1} - w_i\|^2 \\
\Rightarrow &f(w^+) \leq f(w) + \langle w^+ - w, \nabla f(w)\rangle \\
&+ \sum_{i=0}^{4} L(\epsilon)\|w_{i+1} - w_i\|\|w_i - w_0\| + \frac{L(\epsilon)}{2}\|w_{i+1} - w_i\|^2 \\
\Rightarrow &f(w^+) \leq f(w) + \langle w^+ - w, \nabla f(w)\rangle + \frac{15L(\epsilon)}{2}\|w^+ - w\|^2.
\end{aligned}$$

By the definition of projection,

$$\begin{aligned}
w^+ =&\text{argmin}_{y\in\mathcal{W}}\frac{1}{2}\|w - t\nabla f(w) - y\|^2 \\
\Rightarrow&\frac{1}{2}\|w - w^+ - t\nabla f(w)\|^2 \leq \frac{1}{2}\|t\nabla f(w)\|^2 \\
\Rightarrow&\frac{1}{2t}\|w - w^+\|^2 + \langle w^+ - w, \nabla f(w)\rangle \leq 0.
\end{aligned}$$

Adding them together yields

$$\begin{aligned}
f(w^+) + \frac{1}{2t}\|w - w^+\|^2 &\leq f(w) + \frac{15L(\epsilon)}{2}\|w^+ - w\|^2, \\
f(w^+) &\leq f(w) - \frac{1/t - 15L(\epsilon)}{2}\|w^+ - w\|^2.
\end{aligned}$$

Lemma 8. *Let $\{w^k\}$ be a sequence generated by $w^+ \leftarrow \Pi_{\mathcal{W}}(w - t\nabla f(w))$ with initial guess $w^0 \in C$, and let $K = 15L(\epsilon)$. If we choose t_k at each step such that $\underline{t} \leq t_k < \frac{1}{K}$, then*

$$\sum_{k=1}^{\infty}\|w^{k+1} - w^k\|^2 < \infty \Rightarrow \lim_{k\to\infty}\|w^{k+1} - w^k\| = 0.$$

Proof: Since $\underline{t} \leq t_k < \frac{1}{K}$, the expression $\frac{2}{1/t_k - K}$ is bigger than 0, and is upper bounded by some $M > 0$ for all k. By Lemma 7

$$\begin{aligned}\|w^{k+1} - w^k\|^2 &\leq \frac{2}{1/t_k - K}[f(w^k) - f(w^{k+1})] \\ &\leq M[f(w^k) - f(w^{k+1})].\end{aligned}$$

Summing up k from 0 to $N-1$ gives

$$\begin{aligned}&\sum_k \|w^{k+1} - w^k\|^2 \leq M \sum_k f(w^k) - f(w^{k+1}) \\ &= M[f(w^0) - f(w^N)] \leq M[f(w^0) - f(w^*)].\end{aligned}$$

Taking $N \to \infty$ yields the desired result.

Lemma 9. *Let $\{w^k\}$ be a sequence generated by $w^+ \leftarrow \Pi_{\mathcal{W}}(w - t\nabla f(w))$. Define*

$$A^k = \frac{1}{t_{k-1}}(w^{k-1} - w^k) + \nabla f(w^k) - \nabla f(w^{k-1}).$$

Then $A^k \in \nabla f(w^k) + \partial \delta_{\mathcal{W}}(w^k)$ and $A^k \to 0$ as $k \to \infty$.

Proof: By the definition of projected gradient step,

$$\begin{aligned}&0 \in \nabla f(w^{k-1}) + \frac{1}{t_{k-1}}(w^k - w^{k-1}) + \partial \delta_{\mathcal{W}}(w^k) \\ \Rightarrow &\frac{1}{t_{k-1}}(w^{k-1} - w^k) \in \nabla f(w^{k-1}) + \partial \delta_{\mathcal{W}}(w^k).\end{aligned}$$

Hence,

$$\begin{aligned}A^k \in &\nabla f(w^{k-1}) + \partial \delta_{\mathcal{W}}(w^k) + \nabla f(w^k) - \nabla f(w^{k-1}) \\ = &\partial \delta_{\mathcal{W}}(w^k) + \nabla f(w^k).\end{aligned}$$

In turn, we have

$$\begin{aligned}\|A^k\| \leq &\frac{1}{t_{k-1}}\|w^{k-1} - w^k\| + L(\epsilon)\|w^k - w^{k-1}\| \\ &\leq \left(\frac{1}{\underline{t}} + L(\epsilon)\right)\|w^k - w^{k-1}\|.\end{aligned}$$

By Lemma 3, as $k \to \infty$, $A^k \to 0$.

3.2.4 Numerical results

Algorithm 5 is much faster than the standard projected gradient descent on all unknowns [**?**, Section IV B]. We give additional examples to show how the approach identifies mean-reverting time series using simulated data. We simulate five time series; four from an OU process with (μ, σ, θ) given by $(1, 1, 0), (4, 1, 1), (1, 0.5, 1), (4, 0.5, 0)$, and one is non-OU time series with $\sigma = .1$ (the fifth time series). All have $T = 500$ and $\Delta t = 0.01$. We divide the data into training set (70% of data) and test sets (30% of data).

Table 3.3 compares the estimated OU parameters and weight vectors as we tune γ and η. Top three rows correspond to $\gamma = 0$, and bottom three rows $\gamma = 0.5$. When $\gamma = 0, \eta = 5$, the model puts 64% of the weights into the pair of OU time series with $\sigma = 0.5$. It is evident from the results that the model favors OU time series with a lower σ value but is less sensitive to μ values. With larger η we reach lower negative log likelihood (nll) since that means more freedom in choosing assets.

η	μ	σ^2	θ	w	nll (train,test)
5	2.4	0.09	0.07	[.12, -.11,.33,.31,-.12]	-(3.03,3.04)
4	2.9	0.10	0.32	[.13, .12, .38, .36, 0]	-(2.96,2.94)
3	2.6	0.11	0.23	[0.16, 0,0.43,0.42,0]	-(2.90,2.90)
5	5.0	0.09	0.08	[.11, -.11,-.33,.32,-.12]	-(3.02,2.99)
4	5.8	0.10	0.18	[.14,0,.35,.34,.16]	-(2.93,2.84)
3	4.7	0.11	0.27	[0,0, .40, .40, -.19]	-(2.88,2.83)

Table 3.3: Estimated parameters, weights, and nll. We set $\gamma = 0$ for top three rows, $\gamma = 0.5$ for bottom three rows.

Remark 7. *As noted in Remark 8, γ will drive $\bar{c}$ to be negative. If γ is large, the model may not find a feasible time series combination. In addition, γ controls the balance between negative log likelihood and mean-reversion promoting term (i.e. γc). If γ is too large, the model may choose a portfolio that has high negative log likelihood, i.e. low likelihood. Also, since $\mu = -\frac{1}{\Delta t}\log(c)$ and $c \in (0, 1)$, a small increase in c will be amplified in μ. Hence a*

small γ usually suffices and is preferred.In practice it is a good idea to start with $\gamma = 0$. The tuning of η is straightforward. One can set it to be the desired number of assets for the portfolio.

Real data. We performed experiments with empirical price data from three groups of selected assets: precious metals, large equities and oil companies. Data were taken from Yahoo Finance, and give closing stock prices for each asset over the past five years. The first 70% of data (over time) is used for training, and the rest for testing.

For each group, we progressively augmented the set of candidate assets in pairs, and applied our approach. The model determined asset weights, along with negative log-likelihoods of portfolios and of individual assets are given in Table 3.4. The portfolios' negative log likelihoods are generally smaller than negative log likelihoods of individual assets in that portfolio and decrease as we include more assets, which means we can obtain more OU-representable portfolios as the candidate sets expand. The negative log likelihood on the test set can sometimes be significantly larger than that on the training set. In Group 2, individual assets such as GOOG, JNJ and MCD in have significantly larger negative log likelihood on test than on training. The discrepancy in likelihood indicates that those assets have very different patterns before and after the split of training/test. As a result, the constructed portfolios also tend to have larger negative log likelihood on test set. This also suggests that one should check individual asset patterns before generalizing fitted model to another time period.

We also conducted experiments varying γ to promote larger μ. As summarized in Table 3.5, when $\gamma > 0$, we see increasing μ across asset groups. As $c = \exp(-\Delta t\mu) \approx 1 - \Delta t\mu$, the change in c due to γ will be magnified in μ, hence we may see fairly drastic increase in μ.

Comparison with Pairs Trading We compared our approach with that in chapter 2 of [47] on pairs trading. In [47], two assets are selected first, from which a portfolio is constructed as

$$X = S_1 - \beta S_2 \tag{3.24}$$

where S_1 and S_2 are asset price time series. This "β-method" requires longing the first

Assets	2	4	6	indiv. nll (train,test)
GLD	-0.17	-0.08	-0.07	0.77,0.44
GDX		-0.21	-0.29	0.05,-0.30
GDXJ			0.03	0.70, 0.38
SLV	0.83	0.44	0.30	-0.69, -1.0
GG			0.10	-0.04, -0.44
ABX		0.27	0.21	-0.24 ,-0.54
Port.	-1.48,-1.72	-1.95,2.12	-2.18,-2.35	
GOOG				2.66, 3.06
JNJ		-0.12	-0.10	0.40, 0.86
NKE	-0.49	-0.36	-0.27	0.09, 0.43
MCD		-0.11	-0.07	0.49, 1.09
SBUX	0.51	0.41	0.36	0.02, 0.05
SPY			0.12	0.95 ,1.00
VIG				-0.07 ,0.01
VO			-0.08	0.53 ,0.45
Port.	-0.70,-0.08	-0.77,-0.14	-1.07,-0.52	
BP			-0.01	-0.09, -0.33
COP		-0.01	-0.01	0.46, 0.25
CVX		-0.02	-0.01	0.79, 0.73
OIL	-0.59	-0.57	-0.57	-0.84, -1.25
USO	0.41	0.41	0.41	-0.45, -0.86
VLO			0.002	0.58, 0.43
XOM				0.48, 0.26
Port.	-2.89,-3.29	-2.94,-3.35	-2.96,-3.37	

Table 3.4: Negative log-likelihood (nll) of assets groups for $\eta \in \{2, 4, 6\}$ (no. of assets in portfolio) and $\gamma = 0$. The bottom row shows the (training, testing) nll of our optimal portfolios.

γ	η	μ	σ^2	θ	nll (train, test)
0	2	2.69	4.77	-6.42	-1.48, -1.72
0.5	2	4.51	4.78	-6.14	-1.48, -1.72
0	4	2.28	1.87	-2.90	-1.95, -2.13
0.5	4	7.06	2.35	-2.65	-1.84, -2.07
0	6	1.20	1.17	-3.30	-2.18, -2.35
0.5	6	12.70	1.11	-0.98	-2.21, -2.39
0	2	5.74	22.85	-0.57	-0.70, -0.08
0.5	2	11.49	23.11	-0.56	-0.69 -0.04
0	4	1.90	19.76	-22.84	-0.77, -0.14
0.5	4	4.12	19.63	-23.61	-0.77, 0.01
0	6	3.54	10.87	1.00	-1.07, -0.52
0.5	6	6.35	10.64	-1.78	-1.08, -0.46
0	2	11.80	0.28	0.89	-2.89, -3.29
0.5	2	34.43	0.29	1.00	-2.87, -3.26
0	4	16.84	0.26	0.47	-2.94, -3.35
0.5	4	37.80	0.27	0.44	-2.92, -3.31
0	6	17.39	0.25	0.49	-2.95, -3.37
0.5	6	42.63	0.26	0.63	-2.93, -3.31

Table 3.5: Model estimations with different γ and η for precious metals, large equities, and oil companies.

asset and shorting the other. With the weight of the first asset fixed to be 1, this method first determines, for each fixed β, the model parameters that maximizes the OU likelihood of the corresponding portfolio X. Then, in a separate step, it searches over a range of β for the MLE. For this approach to identify the optimal pairs, one needs to further find two optimal β's by switching positions of two assets in (3.24). In contrast, our model solves for the optimal portfolio in a single step. For the examples in Table 3.6, we can simply take the results from our model with $\eta = 2$ from Table 3.4.

	β	portfolio weights
GLD - βSLV	3.68	[0.21, -0.79]
-βGLD + SLV	0.19	[-0.17,0.83]
our model	–	[-0.17,0.83]
NKE - βSBUX	0.61	[0.63, -0.37]
- βNKE + SBUX	0.52	[-0.33, 0.67]
our model	–	[-0.49, 0.51]
OIL - βUSO	0.67	[0.59, -0.41]
- βOIL + USO	1.42	[-0.59, 0.41]
our model	–	[-0.59, 0.41]

Table 3.6: Summary of portfolio weights from our model and method of (3.24) applied to different pairs.

Chapter 4

HYBRID SYSTEMS[1]

This paper considers the problem of using noisy measurements from a piecewise-continuous trajectory to estimate a hybrid system's state. A hybrid dynamical system switches between dynamic regimes at time- or state-triggered events. The state estimation problem has been extensively studied in classical dynamical systems whose states evolve according to one (possibly time–varying) smooth model. This problem is fundamentally more challenging for hybrid systems since the set of discrete state[2] sequences generally grows combinatorially in time.

When the discrete state sequence and switching times are known a priori or directly measured, only the continuous state needs to be estimated, yielding a classical state estimation problem; this approach has been applied to piecewise-linear systems [57, Chap. 4.5] and to nonlinear mechanical systems undergoing impacts [58]. When the discrete state is not known or measured, estimating both the discrete and continuous states simultaneously improves estimation performance. One approach uses a bank of filters, each tuned to one discrete state, and selects the discrete states as the filter with the lowest residual [59, §4.1]. This filter bank method has been applied to hybrid systems with linear dynamics [60, §4.1] [61], nonlinear dynamics [62], and jumps in the continuous state when the discrete state changes [63]. Likewise, particle filter methods for hybrid systems [64–66] use a collection of filters, identified as particles, and are applicable to more general nonlinear process dynamics. Particle filters

---

[1]JOINT WORK WITH A. PACE, S. BURDEN.

[2]The state of the hybrid system is specified by the discrete and continuous components. We refer to the discrete component of the hybrid system state as the discrete state, and refer to the continuous component as the continuous state.

and filter banks are effective when the number of discrete states and dimension of continuous state spaces are small.

Another approach formulates a moving-horizon estimator over both the continuous and discrete states, resulting in a mixed-integer optimization problem [67]. The inherently discrete nature of the problem formulation enables estimation of the exact sample when the discrete state switches, at the expensive of combinatorial growth of the set of discrete decision variables as the horizon increases. Multiple methods have been developed to mitigate the challenge posed by this combinatorial complexity. One approach entails summarizing past measurements and state estimates with a penalty term in the the objective function [68]. Another approach, applicable to systems with bounded noise, entails restricting the set of possible discrete state sequences using a priori knowledge of the system [69, 70].

An alternative approach to circumventing the combinatorial challenge entailed by exactly estimating the discrete state sequence involves relaxing the discrete state estimate to take on continuous values as in [71, 72]. The latter reference uses a sparsity-promoting convex program whose objective incorporates a nonsmooth penalty across all possible discrete state sequences, and guarantees the estimate converges to the true continuous and discrete states. Both approaches are formulated for piecewise-linear systems whose continuous states do not jump when switching between subsystems; in the language of hybrid systems, the continuous states are reset using the identity function.

Our approach and contributions

We propose an offline algorithm for estimating the state of hybrid systems with nonlinear dynamics, non–identity resets, and noisy process and observation models. Although prior work accommodates aspects of our problem formulation, to the best of our knowledge no work simultaneously allows nonlinear dynamics and non–identity resets: [63] does not allow nonlinear dynamics, [64] and [73] do not allow non–identity reset, and [71] does not allow either nonlinear dynamics nor non–identity resets. Our starting point is the optimization perspective on generalized and robust state estimation [74,75]. To formulate state estimation

as a continuous optimization problem, we relax the discrete state to take on continuous values as in prior work. Unlike prior work on state estimation for hybrid systems, we model process noise using the Student's t distribution, which allows large innovations and makes the method applicable to systems with non–identity resets.

In combination, these elements yield a nonsmooth nonconvex continuous optimization formulation for offline state estimation (Sec. 4.1). We develop a Gauss-Newton type algorithm to solve this problem and prove the algorithm globally converges to stationary points (Sec. 4.2). The algorithm is compared to a class of state-of-the-art algorithms (Sec. 4.5) and evaluated on piecewise-linear and -nonlinear hybrid system models (Sec. 4.6).

4.1 Problem formulation

We consider observational data periodically sampled from a continuous-time hybrid dynamical system [76] that undergoes occasional jumps in continuous state, such as a mechanical system undergoing intermittent impacts [77]. We utilize a discrete-time switched system as the process model for this sampled data. The process model is chosen to capture the salient features of a hybrid dynamical system model, e.g. the continuous-time dynamics differing between discrete states, while shifting the challenge of non–identity resets to the process noise. As we explain below, combining this process model with a Student's t distribution for the process noise captures the salient features of the underlying system dynamics while enabling our derivation of a computationally efficient state estimation algorithm.

4.1.1 Process and observation models

We use a discrete-time switched system

$$\begin{aligned} x_{t+1} &= \sum_{m=1}^{M} \mathcal{F}_m(x_t) w_t[m] + \sigma_t \\ y_t &= \mathcal{H}_t(x_t) + \delta_t \end{aligned} \tag{4.1}$$

where $m \in \{1, ..., M\}$ indexes the continuously-differentiable process model $\mathcal{F}_m \colon \mathbb{R}^n \to \mathbb{R}^n$, $M \in \mathbb{N}$ is the number of process models, $\mathcal{H}_t \colon \mathbb{R}^n \to \mathbb{R}^d$ is the continuously-differentiable ob-

servation model that generates observations $y_t \in \mathbb{R}^d$ of the hidden continuous state $x_t \in \mathbb{R}^n$, σ_t, δ_t are process and measurement noises, and $w_t \in \mathcal{D}^M$ is a one-hot vector[3]that indicates which process model is active at time t. Note that the observation model does not depend explicitly on the active model $\mathcal{F}_m$, which must be inferred from measurements of the continuous state x_t.

The model $\mathcal{F}_m$ that is active during each time step may be determined by an exogenous signal, prescribed as a function of time or state, or some combination thereof. Thus, the equation in (4.1) can represent the process and observation models of a wide variety of hybrid systems. Appendix B provides an overview of the construction of a switched system by sampling a general hybrid dynamical system. We are motivated theoretically and experimentally to focus on cases where the active model $\mathcal{F}_m$ is constant for many time steps, only occasionally switching to a new model. When the sampling rate of a continuous-time hybrid dynamical system is much faster than the dwell-time [78], consecutive measurements will often be from the hybrid system in the same discrete state.

The problem of when measurements from a switched-system as in (4.1) with no process noise $\sigma_t \sim 0$, and no measurement noise $\delta_t \sim 0$, can reconstruct the true discrete and continuous state (i.e. when is the system is observable) is well studied continuous time switched linear systems [79] [72, Chpt. 2]. For the more general linear hybrid system, when the continuous state undergoes occasional jumps, observability tests with particular assumptions have been proposed [60]. To the best of our knowledge there is not a general observability test that applies to nonlinear hybrid systems with non–identity resets; a class of hybrid systems considered in this paper.

When the discrete state changes in a hybrid system, the continuous state may change abruptly according to a reset map. As an example, the velocity of a rigid mass changes abruptly when it impacts a rigid surface [80]. Empirically, these discrete reset dynamics are much more poorly characterized than their continuous counterparts. For instance, whereas

[3] $w \in \mathbb{R}^M$ is one-hot if $w[i] \in \{0, 1\}$ for all $i \in \{1, \ldots, M\}$ and $1^T w = 1$; $\mathcal{D}^M \subset \mathbb{R}^M$ denotes the set of one-hot vectors.

the ballistic trajectory of a rigid mass is well-approximated by Newton's laws, the abrupt change in velocity that occurs at impact is not consistent with any established impact law [81]. Including such a reset in the system model (4.1) will introduce bias into the state estimate because the model will generate erroneous predictions at resets, diminishing the accuracy of estimated states at nearby times. This observation motivates us in the next section to account for the effect of unknown resets as part of the process noise.

4.1.2 Process noise and observation noise models

Instead of incorporating continuous state resets explicitly into the model (4.1), we introduce a distributional assumption on the process noise σ_t that accepts large instantaneous changes in the continuous state estimate. Specifically, we assume that process noise σ_t follows a Student's t distribution. However, we emphasize that this is a modeling assumption. It does not imply that process noise from real hybrid system has to follow this distribution. Compared with the commonly-used Gaussian distribution, the heavy-tailed Student's t is tolerant to large deviations in the estimate of the hidden continuous state x_t [28]. Hence, the Student's t error model allows an instantaneous change in the state that is consistent with (4.1) before and after the change. The negative log-likelihood of the Student's t (as a function of σ_t) is given by

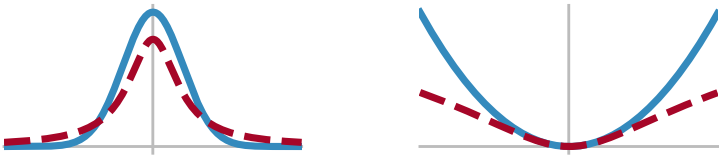
$$r \log \left(r + \left\| Q^{-1/2} \sigma_t \right\|^2 \right) - C(r), \tag{4.2}$$

where r is the degrees-of-freedom parameter of the Student's t, and Q is the covariance matrix, and $C(r)$ is a term independent of σ_t.

If the continuous state x_t was known, then any residual between the predicted observations $\mathcal{H}_t(x_t)$ and actual measurements y_t at time t is due to measurement noise; in particular, the residual does not exhibit large deviations due to continuous state resets at switching times. Thus, we assume the measurement noise δ_t follows the usual Gaussian distribution, with negative log-likelihood

$$\frac{1}{2} \left\| R^{-1/2} \delta_t \right\|^2, \tag{4.3}$$

where R is the covariance matrix. The plots below provide a comparison between the probability density (left) and the negative log-likelihood (right) for the scalar Gaussian (solid blue) and Student's t distributions (dashed red; degree-of-freedom $r = 1$).



probability density negative log-likelihood

4.1.3 State estimation problem formulation

We derive the objective function for estimating states of (4.1) using maximum a posteriori (MAP) likelihood. Including the constraint on w, we obtain the optimization problem

$$\min_{x_t \in \mathbb{R}^n, w_t \in \mathcal{D}^m} \sum_{t=0}^{T-1} l_{\text{meas}}(x_t, y_t) + l_{\text{proc}}(x_t, y_t, w_t) \tag{4.4}$$

where

$$l_{\text{meas}}(x_t, y_t) = \frac{1}{2} \left\| R^{-1/2} \left(y_t - \mathcal{H}_t(x_t) \right) \right\|^2$$

and

$$l_{\text{proc}}(x_t, y_t, w_t) = r \log \left(r + \left\| Q^{-1/2} \left(x_{t+1} - \sum_{m=1}^{M} \mathcal{F}_m(x_t) w_t[m] \right) \right\|^2 \right).$$

Problem (4.4) is a nonlinear mixed-integer program with respect to both the continuous (x_t) and discrete (w_t) decision variables, with the discrete variable constrained to be a one-hot vector ($w_t \in \mathcal{D}^M$). We can significantly simplify the structure by establishing the following lemma.

Lemma 10 (Formulation Equivalence)**.** *Given $w \in \mathcal{D}^M$, any vectors x_1, x_2, models $\mathcal{F}_i$, and any penalty functional g, we have*

$$\begin{aligned} &\min_{w \in \mathcal{D}^M} \quad g\left(x_2 - \sum_{m=1}^{M} w[m]\mathcal{F}_m(x_1)\right) \\ =&\min_{w \in \mathcal{D}^M} \quad \sum_{m=1}^{M} w[m] g\left(x_2 - \mathcal{F}_m(x_1)\right) \end{aligned}$$

and

$$\begin{aligned} &\operatorname*{argmin}_{w \in \mathcal{D}^M} g\left(x_2 - \sum_{m=1}^{M} w[m]\mathcal{F}_m(x_1)\right) \\ =&\operatorname*{argmin}_{w \in \mathcal{D}^M} \quad \sum_{m=1}^{M} w[m] g\left(x_2 - \mathcal{F}_m(x_1)\right). \end{aligned}$$

Proof:

Since $w \in \mathcal{D}^M$ for both problems, there are only M possible values for both objective functions, i.e.

$$g(x_2 - \mathcal{F}_1(x_1)), \quad g(x_2 - \mathcal{F}_2(x_1)), \quad \ldots, \quad g(x_2 - \mathcal{F}_M(x_1)).$$

Hence, the minimum objective value for both problems will be $\min_m g(x_2 - \mathcal{F}_m(x_1))$ and every minimizer is a one-hot vector that selects a minimum value. □

Based on Lemma 10, an equivalent formulation to (4.4) is given by

$$\begin{aligned} \min_{x_t \in \mathbb{R}^n, w_t \in \mathcal{D}^M} \sum_{t=0}^{T-1} \Bigg(&\frac{1}{2} \left\| R^{-1/2}\left(y_t - \mathcal{H}_t(x_t)\right)\right\|^2 + \\ &\sum_{m=1}^{M} w_t[m] r \log\left(r + \left\| Q^{-1/2}\left(x_{t+1} - \mathcal{F}_m(x_t)\right)\right\|^2\right)\Bigg). \end{aligned} \tag{4.5}$$

Although still a mixed-integer program, this reformulation exhibits linear coupling between the discrete variables w_t and continuous variables x_t. We will leverage this linear coupling when we develop our estimation algorithm based on the relaxed problem formulation introduced in the next section.

4.1.4 Relaxed state estimation problem formulation

Ultimately, the discrete state estimate will be specified as a one-hot vector, $w_t \in \mathcal{D}^M \subset \mathbb{R}^M$. To formulate a continuous optimization problem that approximates the mixed-integer problem formulated in the previous section, we relax the decision variable w_t to take values in the convex hull Δ^M of $\mathcal{D}^M$.We use $\Delta^M := \{w \in [0,1]^M : 1^T w = 1\}$ to denote the simplex in $\mathbb{R}^M$. The optimal relaxed w_t will generally lie on the interior of the simplex, so we project the result from our relaxed optimization problem to return the one-hot discrete state estimate. Since this relaxation-optimization-projection process tends to induce frequent changes in the discrete state estimate, we introduce a smoothing term on w_t,

$$\nu \|w_{t+1} - w_t\|_2^2,$$

yielding the continuous relaxation of (4.5) given by

$$\begin{aligned} \min_{x_t \in \mathbb{R}^n, w_t \in \Delta^M} f(x,w) := \sum_{t=0}^{T-1} \Bigg(& \frac{1}{2} \left\| R^{-1/2} \left(y_t - \mathcal{H}_t(x_t) \right) \right\|^2 \\ & + \sum_{m=1}^{M} w_t[m] r \log \left(r + \left\| Q^{-1/2} \left(x_{t+1} - \mathcal{F}_m(x_t) \right) \right\|^2 \right) \\ & + \nu \|w_{t+1} - w_t\|_2^2 \Bigg), \end{aligned} \tag{4.6}$$

where x is the concatenated variable containing all x_t, w is the concatenated variable containing all w_t, and ν is a parameter controlling the strength of smoothing. The optimal relaxed discrete state estimate $w_t \in \Delta^M$ is projected onto $\mathcal{D}^M$ by choosing the (unique) one-hot vector whose $\operatorname{argmax}_m w_t[m]$ component is equal to 1.

4.2 State estimation algorithm

In this section, we derive an algorithm to solve the relaxed state estimation problem formulated in (4.6) using two key ideas:

1. nonsmooth variable projection;

2. Gauss-Newton descent with Student's t penalties.

These two ideas are explained in the next two subsections, followed by a convergence analysis in the third subsection.

4.2.1 Nonsmooth variable projection

The first idea is to pass to the value function, projecting out (partially minimizing over) the w variables, so as to reduce the number of variables to optimize over. Define

$$v(x) := \min_{w \in \Delta} f(x, w) = \min_{w} f(x, w) + \delta_{\Delta}(w) \tag{4.7}$$

with $f(x, w)$ as in (4.6). Assuming w_0 is known, the objective $f(x, w)$ is strongly convex in w. By Theorem 2.2 in [15], since $f(x, w)$ has Lipschitz-continuous gradient with respect to both x and w, and is strongly convex in w, $v(x)$ is differentiable and its gradient is given by

$$\nabla v(x) = \partial_x f(x, w)|_{w=w(x)}. \tag{4.8}$$

Plugging $w(x)$ back into (4.6) we obtain the problem

$$\begin{aligned} \min_x v(x) =& \frac{1}{2} \sum_{t=0}^{T-1} \|y_t - \mathcal{H}(x_t)\|^2_{R^{-1}} + \nu \|w_{t+1}(x) - w_t(x)\|_2^2 \\ &+ \sum_{m=1}^{M} w_{t,m}(x) r \log \left(1 + \frac{\|x_{t+1} - \mathcal{F}_m(x_t)\|^2_{Q^{-1}}}{r} \right), \end{aligned} \tag{4.9}$$

where $w_{t,m}(x) \equiv w_t[m](x)$.

4.2.2 Gauss-Newton updates on Student's t penalties

We apply a modified Gauss-Newton algorithm to solve (4.9) based on [13] and [28]. [13] derives a general Gauss-Newton scheme for solving composite minimization, whereas [28] suggests a particular construction of local subproblems suitable for objectives involving Student's t penalty.

We decompose the objective into two parts $v(x) = f_1(x) + f_2(x)$ where

$$\begin{aligned} f_1(x) =& \frac{1}{2}\sum_{t=0}^{T-1}\sum_{m=1}^{M} w_{t,m}(x) r \log\left(1 + \frac{\|x_{t+1} - \mathcal{F}_m(x_t)\|^2_{Q^{-1}}}{r}\right) \\ &+ \nu\|w_{t+1}(x) - w_t(x)\|_2^2 \\ f_2(x) =& \frac{1}{2}\|F_2(x)\|^2_{R^{-1}}, \quad F_2(x) = \mathcal{H}(x) - y. \end{aligned}$$

Equivalently, we can think of v as a composite function $v = \phi \circ F$ where

$$\phi(u) = u + \delta_{\mathbb{R}_{[0,\infty)}}(u), \quad F = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}.$$

We then define a local model $\psi(x; z)$ by linearizing F as in [13],

$$\psi(x; z) = f_1(x) + f_2(x) + \nabla f_1(x)(z - x) + \nabla f_2(x)(z - x) \tag{4.10}$$

and x-update is obtained by solving a local subproblem

$$x^+ \leftarrow \operatorname*{argmin}_z \psi(x; z) + \frac{C}{2}(z - x)^T(\nabla F_2(x)^T R^{-1} \nabla F_2(x) + U(x))(z - x) \tag{4.11}$$

where $U(x)$ is a positive definite Hessian approximation of the Student's t term in $f_1(x)$. Note that (4.11) is different from that presented in [13] since we do not use the usual proximal term $\frac{L}{2}\|z - x\|^2$. The choice of $U(x)$, proposed in [28, (5.5), (5.6)], is shown to improve computational efficiency; it is of the form

$$U = \begin{bmatrix} U_1 & A_2^T & 0 & \\ A_2 & U_2 & A_3^T & 0 \\ 0 & \ddots & \ddots & \ddots \\ & 0 & A_T & U_T \end{bmatrix} \tag{4.12}$$

with

$$A_t = -r\sum_{m=1}^{M} w_{t-1,m}(x) \frac{Q^{-1}\nabla \mathcal{F}_m(x_{t-1})}{r + \|x_t - \mathcal{F}_m(x_{t-1})\|^2_{Q^{-1}}},$$

$$U_t = r\sum_{m=1}^{M} \frac{w_{t,m}(x)\nabla\mathcal{F}_m(x_t)^T Q^{-1}\nabla\mathcal{F}_m(x_t)}{r + \|x_{t+1} - \mathcal{F}_m(x_t)\|^2_{Q^{-1}}} + \frac{w_{t-1,m}(x)Q^{-1}}{r + \|x_t - \mathcal{F}_m(x_{t-1})\|^2_{Q^{-1}}}$$

for $1 \leq t \leq T-1$, and

$$U_T = \frac{r w_{T-1,m}(x)Q^{-1}}{r + \|x_T - \mathcal{F}_m(x_{T-1})\|^2_{Q^{-1}}}.$$

We can rewrite $U(x)$ as

$$U(x) = \sum_m G_m(x)^T \tilde{Q}_m(w(x))^{-1} G_m(x),$$

where

$$G_m(x) = \begin{bmatrix} I & 0 & 0 & \\ -\nabla\mathcal{F}_m(x_2) & I & 0 & 0 \\ 0 & \ddots & \ddots & \ddots \\ \dots & 0 - \nabla\mathcal{F}_m(x_T) & & I \end{bmatrix}$$

and

$$\tilde{Q}_m(w(x))^{-1} = \mathrm{diag}(\tilde{Q}_{m,t}(w(x))^{-1})$$
$$\tilde{Q}_{m,t}(w(x))^{-1} = \frac{r w_{t-1,m}(x)Q^{-1}}{r + \|x_t - \mathcal{F}_i(x_{t-1})\|^2_{Q^{-1}}}.$$

$U(x)$ is positive semidefinite. Let

$$\Sigma(x) = \nabla F_2(x)^T R^{-1} \nabla F_2(x) + U(x).$$

Since R^{-1} is positive definite and $U(x)$ is positive semidefinite, $\Sigma(x)$ is positive semidefinite; we show later that $U(x)$ is actually positive definite, which implies that $\Sigma(x)$ is also positive definite. We further define

$$\psi_C(x,z) = \psi(x;z) + \frac{C}{2}(z-x)^T\Sigma(x)(z-x), \quad V_C(x) = \mathrm{argmin}_z \psi_C(x,z), \quad f_C(x) = \psi_C(x, V_C(x))$$

for some positive number C.

Our proposed Gauss-Newton updates follows from [13] and proceeds as in Algorithm 6. That the line search in the algorithm is well-defined will be discussed in the next section. Incorporating Gauss-Newton, the full algorithm is given in Algorithm 7.

Algorithm 6 Gauss-Newton for $v_\beta(x)$ (4.9).

Require: x_0, c_l, c_u

1: **for** $k = 0, 1, 2, 3, ...$ **do**

2: $\quad x^{(k+1)} \leftarrow V_{C_k}(x^{(k)})$ such that $v_\beta(V_{C_k}(x^{(k)})) < f_{C_k}(x^{(k)}), C_k \in [c_l, c_u]$

Algorithm 7 Variable Projection with Gauss-Newton for (4.9).

Require: $x_0, w_0, Q, R, r, \nu, \beta$

1: **for** $k = 0, 1, 2, 3, ...$ **do**

2: $\quad w^{(k)} \leftarrow \text{InnerSolver}_{\Pi_t \Delta}(x^{(k)})$

3: $\quad x^{(k+1)} \leftarrow$ one step Gauss-Newton as in algorithm 6 given $x^{(k)}, w^{(k)}$

4: $\quad \text{loss}_k \leftarrow v(x^{(k+1)})$

4.3 Convergence of state estimation algorithm

In this section we show that the algorithm converges to stationary points in the limit. The proof follows the framework in [13] with adjustment for our specific model. Denote

$$\Lambda(v) = \{x : v(x) \leq v\},$$

and in particular

$$\Lambda_0 = \{x : v(x) \leq v_0 \equiv v(x^{(0)})\}$$

where $x^{(0)}$ is the initial value for x. For subsequent analysis we always consider $x \in \Lambda_0$.

The statement we would like to prove is as follows:

$$x^{(k)} \to \bar{x}, \quad x^{(k)} \in \Lambda_0, \quad \text{and } 0 \in \partial v(\bar{x}) \tag{4.13}$$

where $x^{(k)}$ is the sequence generated by Gauss-Newton updates. We show later (Lemma 11) that Λ_0 is compact. Hence, to prove the above statement, it suffices to prove the following two limits:

$$\lim_{k\to\infty} \|x^{(k)} - x^{(k+1)}\| = 0, \quad \lim_{k\to\infty} \Delta_r(x^{(k)}) = 0 \tag{4.14}$$

where

$$\Delta_r(x) = v(x) - \min_z\{\psi(x;z) : \|z - x\| \le r\}.$$

In other words, Δ_r measures how close the local model $\psi(x;z)$ (4.10) is to the true function. To see its relationship to (4.13), observe that since $\psi(x;z) = v(x)$ when $z = x$,

$$\min_z\{\psi(x;z) : \|z - x\| \le r\} \le v(x) \Rightarrow \Delta_r(x) \ge 0.$$

Hence when $\Delta_r(\bar{x}) = 0$, $z = \bar{x}$ is a minimizer for $\psi(\bar{x};z)$, which implies that

$$0 \in \partial_z\psi(\bar{x};z)|_{\bar{x}} \Leftrightarrow 0 \in \nabla f_1(\bar{x}) + \nabla f_2(\bar{x}) \Leftrightarrow 0 \in \partial v(\bar{x}),$$

or $\bar{x}$ is a stationary point for the original objective v. Therefore if $\bar{x}$ is a limit of $x^{(k)}$, we can conclude that the limiting point of sequence $\{x^{(k)}\}$ is a stationary point for $v(x)$.

We give a summary on how the following lemmas are connected to prove (4.14). This is also shown pictorially in Figure 4.1.

- Lemma 11, 12 and 13 are auxiliary lemmas to prove Lemma 15 and Lemma 16.
- The monotonicity of $x^{(k)}$ follows from Lemma 14.
- Lemma 15 bounds distance between consecutive $x^{(k)}$ whereas Lemma 16 bounds $\Delta_r(x^{(k)})$.
- Theorem 6 is proved using Lemma 15 and Lemma 16, and the fact that the sequence $\{x^{(k)}\}$ is non-increasing.
- (4.14), also stated in Lemma 17, is a corollary following from Theorem 6.

Lemma 11. Λ_0 *is a compact set.*

Proof: We show that the set is closed and bounded.

- Closedness is straightforward. Since v is a continuous function, it is closed. Hence Λ_0 is closed.

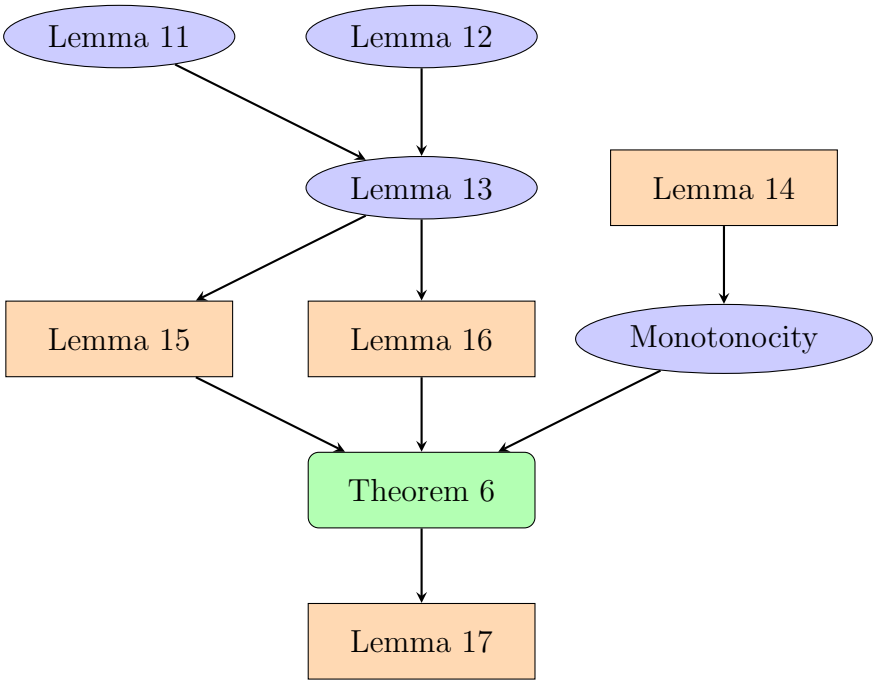


Figure 4.1: Roadmap of the analysis

- To show boundedness, we prove by contradiction. Suppose Λ_0 is not bounded, then as $\|x\| \to \infty, v(x) < \infty$. Since $v(x) = f_1(x) + f_2(x)$ where $f_1(x) \geq 0, f_2(x) \geq 0$, it must be the case that both $f_1(x) < \infty$ and $f_2(x) < \infty$ as $\|x\| \to \infty$. In particular, recall that $f_1(x)$ has the term

$$\frac{1}{2}\sum_{t=0}^{T-1}\sum_{m=1}^{M} w_{t,m}(x) r \log\left(1 + \frac{\|x_{t+1} - \mathcal{F}_m(x_t)\|^2_{Q^{-1}}}{r}\right).$$

  If $\|x\| \to \infty$, then $\exists t+1, \|x_{t+1}\| \to \infty$, which further implies that $\|x_t\| \to \infty$ since $\mathcal{F}_m$'s are assumed to be proper functions. Iterating this gives the limiting case that $f_1(x) < \infty$ as $\|x_1\| \to \infty$. However we start at some finite x_0, hence $\|x_1 - \mathcal{F}_m(x_0)\| \to \infty$ as $\|x_1\| \to \infty$. Therefore a contradiction.

This completes the proof. □

Lemma 12. *$U(x)$ is a positive definite matrix.*

Proof: To show that $U(x)$ is positive definite, recall that we can rewrite $U(x)$ as

$$U(x) = \sum_m G_m(x)^T \tilde{Q}_m(w(x))^{-1} G_m(x) \succeq 0.$$

If there exists some d such that $d^T U(x) d = 0$, then

$$\begin{aligned}
&d^T \left(\sum_m G_m(x)^T \tilde{Q}_m(w(x))^{-1} G_m(x) \right) d \\
=&\sum_m \underbrace{d^T G_m(x)^T}_{z_m(x)^T} \tilde{Q}_m(w(x))^{-1} \underbrace{G_m(x) d}_{z_m(x)} \\
=&\sum_m z_m(x)^T \tilde{Q}_m(w(x))^{-1} z_m(x) = 0, \\
\Rightarrow& z_m(x)^T \tilde{Q}_m(w(x))^{-1} z_m(x) = 0 \ \forall i \\
\Rightarrow& z_{m,t}(x)^T \tilde{Q}_{m,t}(w(x))^{-1} z_{m,t}(x) = 0 \forall t \ \forall i
\end{aligned}$$

since $\tilde{Q}_m(w(x))^{-1} = \text{diag}(\tilde{Q}_{m,t}(w(x))^{-1})$, and

$$\tilde{Q}_{m,t}(w(x))^{-1} = \frac{r w(x)_{t,m} Q^{-1}}{r + \|x_{t+1} - \mathcal{F}_m(x_t)\|^2_{Q^{-1}}}$$

are positive semidefinite. However because each $w_t \in \Delta$, there has to be some $\tilde{Q}^{-1}_{m,t} \succ 0$ for each t. Therefore $U(x)$ must be positive definite. □

Lemma 13. *There exists some $\lambda_{\max}, \lambda_{\min} > 0$ independent of x such that*

$$\lambda_{\min} = \inf_x \lambda_{\min}(\Sigma(x)), \quad \forall x \in \Lambda_0$$

$$\lambda_{\max} = \sup_x \lambda_{\max}(\Sigma(x)), \quad \forall x \in \Lambda_0.$$

Proof: $\Sigma(x)$ is positive definite due to positive definiteness of $U(x)$. Since the mapping $x \in \Lambda_0 \mapsto \lambda_{\min}(\Sigma(x))$ is continuous, and Λ_0 is compact, its image is compact. Hence there exists some $\lambda_{\min} > 0$ independent of x such that $\lambda_{\min} = \inf_x \lambda_{\min}(\Sigma(x))$ for all $x \in \Lambda_0$. Likewise for $\lambda_{\max}$. □

Lemma 14. *Let L denote the Lipschitz constant of $\nabla F(x)$. For any $x, z \in \Lambda_0$,*

$$v(z) - v(x) \leq \frac{L}{2} \|z - x\|^2.$$

Proof: By Lipschitz continuity of F, we have

$$\|F(z) - F(x) - \nabla F(x)(z-x)\| \leq \frac{L}{2}\|z-x\|^2,$$

hence

$$v(z) - \psi(x;z) = \phi(F(z) - F(x) - \nabla F(x)(z-x)) \leq \frac{L}{2}\|z-x\|^2$$

□

In particular, if we take $C \geq L/\lambda_{\min}$, then

$$v(z) - \psi(x;z) \leq \frac{\lambda_{\min} C}{2}\|z-x\|^2 \leq \frac{C}{2}(z-x)^T\Sigma(x)(z-x) \tag{4.15}$$

$$\Rightarrow v(V_C(x)) \leq \psi(x; V_C(x)) + \frac{C}{2}(V_C(x)-x)^T\Sigma(x)(V_C(x)-x) = f_C(x) \tag{4.16}$$

This inequality 4.16 justifies the line search step in algorithm 6, and that the sequence generated by it is non-increasing.

Lemma 15. *For any $x \in \Lambda$, we have*

$$\delta_C(x) \geq \frac{\lambda_{\min} C}{2} r_C^2(x)$$

where

$$r_C(x) = \|V_C(x) - x\|$$
$$\delta_C(x) = v(x) - f_C(x)$$

Proof: Fix an arbitrary x and consider the function

$$\xi(t) = \min_z \{\psi(x;z) + \frac{1}{2t}(z-x)^T\Sigma(x)(z-x)\}$$

We claim that the second term $\frac{1}{2t}(z-x)^T\Sigma(x)(z-x)$ is jointly convex in z, t. If it is convex, since $\psi(x;z)$ is convex in z by construction, $\psi(x;z) + \frac{1}{2t}(z-x)^T\Sigma(x)(z-x)$ is also jointly convex in z, t. Hence $\xi(t)$ is convex in t from elementary convex analysis. By the property of subgradient for convex functions,

$$v(x) = \xi(0) \geq \xi(t) + \xi'(t)(-t) = f_{1/t}(x) + \frac{1}{2t}(V_{1/t}(x) - x)^T U(x)(V_{1/t}(x) - x) \geq f_{1/t}(x) + \frac{\lambda_{\min}}{2t} r_{1/t}^2(x).$$

By plugging in $t = \frac{1}{C}$, we get the stated result.

Now to show the claim that $\frac{1}{2t}(z-x)^T\Sigma(x)(z-x)$ is convex, we show that its equivalent statement – that its epigraph is a convex set. Its epigraph is the set $\{(z,t,\alpha) \in \mathbb{R}^{Tn} \times \mathbb{R}^2_+ : (z-x)^T U(x)(z-x) \leq \alpha t\}$.

Let (z_1, t_1, α_1) and (z_2, t_2, α_2) be two arbitrary points in the set, and consider their convex combination,

$$
\begin{aligned}
&(\lambda z_1 + (1-\lambda)z_2 - x)^T U(x)(\lambda z_1 + (1-\lambda)z_2 - x) \\
=&\lambda^2 (z_1 - x)^T U(x)(z_1 - x) + (1-\lambda)^2 (z_2 - x)^T U(x)(z_2 - x) \\
&+ 2\lambda(1-\lambda)(z_1 - x)^T U(x)(z_2 - x) \\
\leq&\lambda^2 \alpha_1 t_1 + (1-\lambda)^2 \alpha_2 t_2 + 2\lambda(1-\lambda)\sqrt{(z_1 - x)^T U(x)(z_1 - x)(z_2 - x)^T U(x)(z_2 - x)} \\
\leq&\lambda^2 \alpha_1 t_1 + (1-\lambda)^2 \alpha_2 t_2 + 2\lambda(1-\lambda)\sqrt{\alpha_1 t_1 \alpha_2 t_2} \\
=&\lambda^2 \alpha_1 t_1 + (1-\lambda)^2 \alpha_2 t_2 + 2\lambda(1-\lambda)\sqrt{\alpha_1 t_2}\sqrt{\alpha_2 t_1} \\
\leq&\lambda^2 \alpha_1 t_1 + (1-\lambda)^2 \alpha_2 t_2 + 2\lambda(1-\lambda)(\alpha_1 t_2 + \alpha_2 t_1) \\
=&(\lambda\alpha_1 + (1-\lambda)\alpha_2)(\lambda t_1 + (1-\lambda)t_2),
\end{aligned}
$$

hence their convex combination is also in the set, so the set is convex. □

Lemma 16. *For any $x \in \Lambda$ and $r > 0$, we have*

$$\delta_C(x) \geq \lambda_{\max} C r^2 \kappa\left(\frac{1}{\lambda_{\max} C r^2}\Delta(x)\right)$$

where

$$\Delta_r(x) = v(x) - \min_z\{\psi(x;z) : \|z - x\| \leq r\}$$

$$\kappa(t) = \begin{cases} t - \frac{1}{2} & t \geq 1 \\ \frac{t^2}{2} & t \in [0,1] \end{cases}.$$

And the right hand side is a decreasing function of C.

Proof: Let $h \in \text{argmin}_h\{\psi(x; x+h) : \|h\| \leq r\}$. Then

$$\begin{aligned} f_C(x) &\leq \min_{\tau\in[0,1]} \phi(F(x) + \tau\nabla F(x)h) + \frac{\tau^2 C}{2} h^T\Sigma(x)h \\ &= \min_{\tau\in[0,1]} \phi\left((1-\tau)F(x) + \tau(F(x) + \nabla F(x)h)\right) + \frac{\tau^2 C}{2} h^T\Sigma(x)h \\ &= \min_{\tau\in[0,1]} (1-\tau)\phi(F(x)) + \tau\phi(F(x) + \nabla F(x)h) + \frac{\tau^2 C}{2} h^T\Sigma(x)h \\ &\leq \min_{\tau\in[0,1]} v(x) - \tau\Delta(x) + \frac{\lambda_{\max} C}{2}\tau^2 r^2 \end{aligned}$$

Thus,

$$\delta_C(x) \geq \max_{\tau\in[0,1]} \tau\Delta_r(x) + \frac{\lambda_{\max}}{2} C\tau^2 r^2 = \lambda_{\max} C r^2 \kappa(\frac{1}{\lambda_{\max} C r^2}\Delta_r(x)).$$

□

Theorem 6. *Let v^* be the optimal value for $v_\beta(x)$. For any $r > 0$, the sequence generated by algorithm 6 with $c_l \leq L/\lambda_{\min}$ and $c_u \geq L/\lambda_{\min}$ satisfies*

$$v(x^{(0)}) - v^* \geq \frac{1}{2}\lambda_{\min} c_l \sum_{i=0}^{\infty} r^2_{C_i}(x^{(i)}) \geq \frac{1}{2}\lambda_{\min} c_l \sum_{i=0}^{\infty} r^2_{c_u}(x^{(i)})$$

$$v(x^{(0)}) - v^* \geq r^2\lambda_{\max} \sum_{i=0}^{\infty} C_i \kappa\left(\frac{1}{\lambda_{\max} C_i r^2}\Delta_r(x^{(i)})\right) \geq \lambda_{\max} c_u r^2 \sum_{i=0}^{\infty} \kappa\left(\frac{1}{\lambda_{\max} c_u r^2}\Delta_r(x^{(i)})\right)$$

Proof: Using telescoping sequence, for any $k > 0$,

$$\begin{aligned} v(x^{(0)}) - v^* \geq v(x^{(0)}) - v(x_k) &= \sum_{i=0}^{k-1} v(x^{(i)}) - v(x^{(i+1)}) \\ &\geq \sum_{i=0}^{k-1} v(x^{(i)}) - f_{C_i}(x^{(i)}) \\ &= \sum_{i=0}^{k-1} \delta_{C_i}(x^{(i)}) \end{aligned}$$

where the inequality follows from line search criteria in Algorithm 6. Using lemma 15 and 16, we get the first and second inequalities respectively. □

The above theorem in particular implies the following convergence result

Lemma 17. *Let the sequence* $\{x^{(k)}\}$ *be generated from Gauss-Newton updates. Then*

$$\lim_{k\to\infty} \|x^{(k)} - x^{(k+1)}\| = 0, \quad \lim_{k\to\infty} \Delta_r(x^{(k)}) = 0,$$

and since Λ_0 *is compact, at a limiting point* $\bar{x}$, $\Delta_r(\bar{x}) = 0$.

4.4 Parameter tuning for proposed algorithm

Before we present numerical results, we include a general guidance on parameter tuning for the new algorithm. We discuss both standard parameters (e.g. Q, R) that must be tuned by any algorithm for this application, as well as the parameters ν and r which are specific to our approach. We first give a rough outline of steps we have taken to tune the parameters, followed by more detailed guidelines to tune each individual parameter.

1. Start with large r for Student's t, i.e. distribution close to Gaussian.

2. If Q and R are unknown, they are tuned such that the smooth part of trajectories can be well approximated.

3. Decrease degrees of freedom r of Student's t so that the nonsmooth part of trajectories can be captured.

4. Adjust smoothing coefficient ν to reduce number of switches.

For degrees of freedom r, one can start with a large value, meaning that the distribution is close to Gaussian, and decrease it later to capture jumps in the continuous state.

For covariance matrices Q and R, if empirical estimations are available, they can be supplied to the model directly. There is existing literature on estimation methods for noise covariance matrices [82]. When such estimations are not available, we usually assume the matrices to be diagonal for simplicity, in which case the inverse of diagonal entries can also be interpreted as weights. The diagonal values of R represent variance for measurements.

When choosing R, we consider the relative scale of measurements, e.g. measurements with smaller magnitude usually have smaller variance. For choices of diagonal values of Q, we usually assign smaller variance for observed states, e.g. positions in our examples,and larger variance for unobserved states.

The choice of smoothing coefficient ν depends on modeler's belief in frequency of switches. One can start with a small value of ν (i.e. little penalty on frequent switches), and gradually increase it, till the pattern of switches is close to modeler's belief.

We recommend having a short piece of manually labeled trajectories as a training set for the purpose of parameter tuning. After tuning, the user can apply the same parameters on larger dataset collected from similar scenarios.

In terms of sensitivity of estimation results on parameters, we had the following observations when running our experiments:

- The estimation result is not very sensitive to r. We were able to decrease r fairly aggressively during parameter tuning.

- For the diagonals of Q and R, we found that it was important to have values in the correct ranges, but the exact values taken were not crucial.

- For smoothing coefficient ν, we noticed that the switching times were sensitive to ν when ν was very small relative to the diagonal entries of Q^{-1} and R^{-1}. Since we assumed that the discrete states should not change too frequently, we used a slightly larger ν.

4.5 Comparison with the Interacting Multiple Model (IMM) method

We compare the nonsmooth variable projection algorithm 7[4]with the Interacting Multiple Model (IMM) [83] algorithm implemented in the open-source package `filterpy` [84]. We

[4]We provide an implementation of 7 at `https://github.com/jizezhang/hds-state-estimation`.

consider two examples, in both cases the continuous state x is a scalar, and there are two discrete states. In the first example, the continuous state x undergoes no jumps, i.e. the reset is the identity function. In the second example, the continuous state x undergoes an instantaneous jump when the discrete state changes; i.e. a non–identity reset. The dynamics of the two discrete state process models are:

$$\dot{x} = -1 \qquad \mathcal{F}_{w=1},$$
$$\dot{x} = 1 \qquad \mathcal{F}_{w=2}.$$

For the second example with non–identity resets, when a discrete state switch occurs, the continuous state decreases by 5. In both examples the discrete state switches at $t = 1$ and $t = 2$. Additionally, the measurement noise has a variance of $R = [.0001]$, which is used as the measurement noise covariance for all models. IMM_1 uses a process noise model with covariance $Q = [.001]$ for both the internal Kalman filters while IMM_2 uses a process noise model with covariance $Q = [.2]$.

In the first example, algorithm 7 (VP) and IMM perform nearly identically (Figure 4.2). Both methods accurately recover the continuous state and discrete state. When the system undergoes instantaneous jumps in the continuous state at discrete state changes, algorithm 7 outperforms IMM (Figure 4.3). For IMM, there is a clear trade–off exists between recovering the continuous state and recovering the discrete state. When using a process noise model with large covariance, as in the case of IMM_2, the continuous state can be recovered at the expense of the discrete state. In the top subplot of Figure 4.3, $\widetilde{w}_{IMM_2}$ is nearly the same value for the duration of the simulation, with slight separation between the two modes. With a smaller covariance, as in IMM_1, the discrete state can be recovered. From $t = 1$ to near $t = 1.25$, IMM_1 incorrectly identifies the discrete state due to the continuous state jump direction being opposite of the continuous state dynamics for discrete state $w = 2$.

Both algorithm 7 and IMM require a similar number of parameters from the user. For both methods, covariance matrices for the process error model Q and measurement error model R need to be provided. IMM adjusts the estimated frequency of switching between

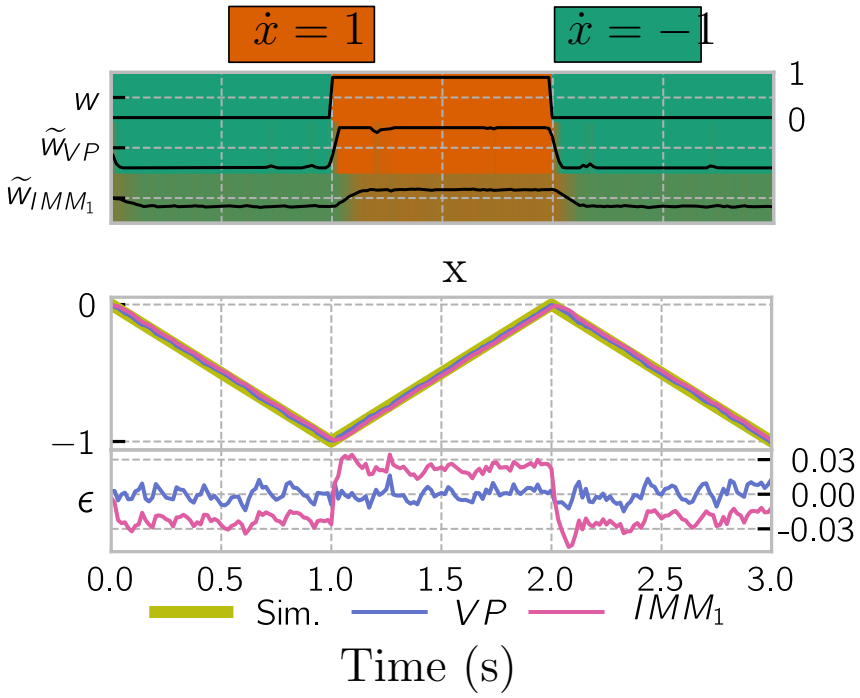


Figure 4.2: **algorithm 7 (VP) performs comparably to IMM when the continuous state does not undergo any resets.** The top plot shows the true state w and the simplex estimate of the true state from both methods $\widetilde{w}_{VP}$, $\widetilde{w}_{IMM_1}$. The simplex estimate is shown in color and the probability estimate of the discrete state being $w = 1$ is superimposed as a black line. The middle plot shows the actual value of the continuous state of the simulation and the estimates. The bottom plot shows the residual between true continuous state and the estimated continuous state.

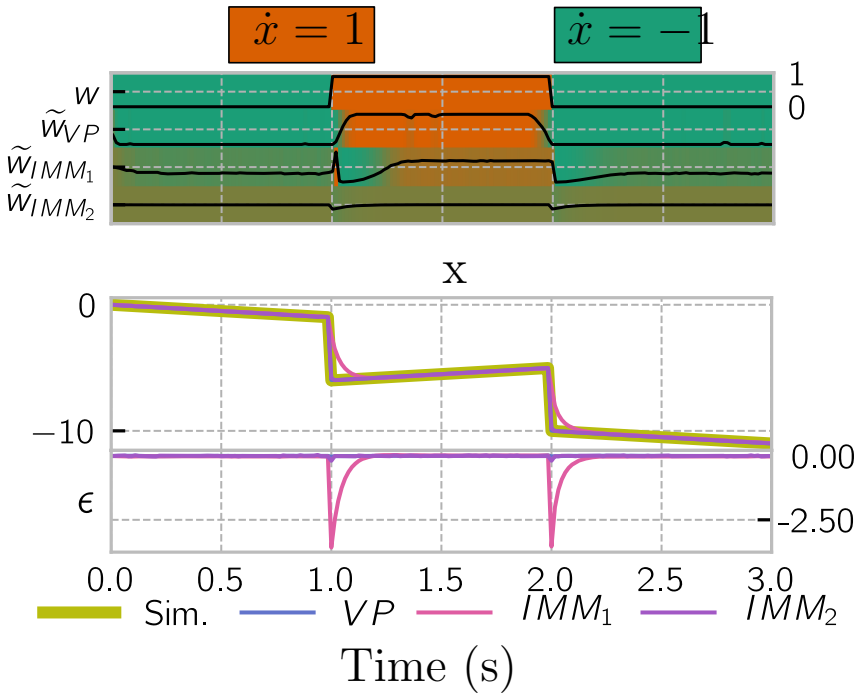


Figure 4.3: **Algorithm 7 (VP) outperforms the IMM when there are jumps in the continuous state.** The plots follow the convention laid out in Figure 4.2.

the discrete states via a probability transition matrix while algorithm 7 uses the smoothing parameter ν, Sec. 4.1.4. algorithm 7 has one additional parameter r due to the process noise model being Student's t distribution, which is crucial for obtaining accurate estimates with non–identity resets, Sec. 4.1.2.

4.6 Experiments with hybrid system models

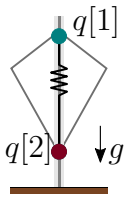
To evaluate the proposed approach to state estimation for hybrid systems, we apply our algorithm to linear and nonlinear impact oscillators. In addition to being well-studied ([85, §1.2], [86]), these mechanical systems were chosen since they are among the simplest physically-relevant models that have non–identity reset maps. The parameter and trajectory regime considered in what follows is representative of a jumping robot constructed from one limb of a commercially-available quadrupedal robot [87] and controlled with an event-triggered stiffness adjustment; Figure 4.4a contains a photograph of the limb. The jumping robot's hip and foot are constrained to move vertically in a gravitational field, so the rigid pantograph mechanism depicted in Figure 4.4b has two mechanical degrees-of-freedom (DOF) coupled through nonlinear pin-joint constraints. These two DOF are preserved, but their nonlinear coupling is neglected, in the piecewise-linear model illustrated in Figure 4.4c. The hybrid dynamics of these linear and nonlinear impact oscillators are specified in Section 4.6.1

We perform two sets of experiments. The first set of experiments in Sec. 4.6.2 concern the piecewise-linear model depicted in Figure 4.4c and explore the consequences of our modeling assumptions and the efficacy of our proposed algorithm:

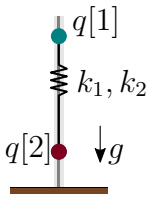
- Sec. 4.6.2 demonstrates the advantage of employing a Student's t distribution for process noise as compared to a Gaussian distribution;

- Sec. 4.6.2 demonstrates the superior convergence rate yielded by Gauss-Newton descent directions as compared to gradient (steepest) descent;

(a)



(b)



(c)

Figure 4.4: **Jumping robot and impact oscillator hybrid system models** (Sec. 4.6.1). (a) Photograph of the physical robot (one leg from a Minitaur [87]) that inspired the simulation models. (b) Nonlinear model consisting of two masses coupled with a linear spring and a nonlinear pantograph mechanism. (c) Linear model consisting of two masses coupled with a linear spring.

- Sec. 4.6.2 demonstrates the advantage of smoothing the relaxed discrete state estimate; and

- Sec. 4.6.2 demonstrates the algorithm's performance when onboard measurements are used instead of offboard measurements.

The second set of experiments in Sec. 4.6.3 evaluate our proposed approach using the nonlinear model depicted in Figure 4.4b.

Since this section is devoted to comparing estimated states to ground truth simulation results, and since our approach entails the determination of a relaxed discrete state estimate en route to obtaining the discrete state estimate, we now introduce notation that distinguishes these quantities:

- $w_t \in \mathcal{D}^M$ denotes the ground truth discrete state;

- $\widetilde{w}_t \in \Delta^M$ denotes the relaxed discrete state estimate;

- $\widehat{w}_t \in \mathcal{D}^M$ denotes the discrete state estimate.

This notational distinction was not introduced previously in the interest of readability since there was no ambiguity entailed by overloading notation in the problem formulation and algorithm specification.

4.6.1 Impact oscillator hybrid system models

The continuous state $x = (q, \dot{q}) \in \mathbb{R}^4$ for the jumping robot hybrid system model consists of the two-dimensional configuration vector $q \in \mathbb{R}^2$ and corresponding velocity $\dot{q} \in \mathbb{R}^2$, where $q[1]$ and $q[2]$ denote the vertical height of the hip and foot, respectively. The foot is not permitted to penetrate the ground, $q[2] \geq 0$, so the first part of the discrete state indicates whether this constraint is active: A (air) if $q[2] > 0$, G (ground) if $q[2] = 0$. To compensate for energy losses at impact, an event-triggered controller stiffens or softens a spring based on which direction the hip is traveling, so the second part of the discrete state indicates the direction of travel for $q[1]$: $\uparrow$ if up, $\downarrow$ if down. With $\ddot{q}_m(q, \dot{q}) \in \mathbb{R}^2$ denoting the acceleration of the hip and foot in discrete state $m \in \{\mathrm{A}\downarrow, \mathrm{G}\downarrow, \mathrm{G}\uparrow, \mathrm{A}\uparrow\}$,[5] formula for this acceleration are given in Table 4.1. At the moment of impact (when the discrete state changes from $w_t \in \{\mathrm{A}\downarrow, \mathrm{A}\uparrow\}$ to $w_{t+1} \in \{\mathrm{G}\downarrow, \mathrm{G}\uparrow\}$) the foot velocity $\dot{q}[2]$ is instantaneously reset to 0, corresponding to perfectly plastic impact. An example of the jump in continuous state when transitioning from $\mathrm{A}\downarrow$ to $\mathrm{G}\downarrow$ on the foot velocity $\dot{q}[2]$ is shown in Figure 4.5 near time 17.5s.

4.6.2 Piecewise-linear impact oscillator experiment

In this subsection, we employ the linear spring laws

$$k_1(q, \dot{q}) = 10(q[1] - q[2]) - 3,$$

and

$$k_2(q, \dot{q}) = 15(q[1] - q[2]) - 3,$$

[5]To simplify exposition we identify $m = \mathrm{A}\downarrow$ with $m = 1$, $m = \mathrm{G}\downarrow$ with $m = 2$, $m = \mathrm{G}\uparrow$ with $m = 3$, and $m = \mathrm{A}\uparrow$ with $m = 4$.

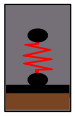
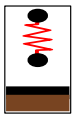
Discrete state w	**Icon**	$\ddot{q}_w(x)$
$w = \mathrm{A}\downarrow$		$\begin{bmatrix} \frac{1}{m_h}\left(-k_1(q,\dot{q})\right) - g \\ \frac{1}{m_t}\left(k_1(q,\dot{q})\right) - g \end{bmatrix}$
$w = \mathrm{G}\downarrow$		$\begin{bmatrix} \frac{1}{m_h}\left(-k_1(q,\dot{q})\right) - g \\ 0 \end{bmatrix}$
$w = \mathrm{G}\uparrow$		$\begin{bmatrix} \frac{1}{m_h}\left(-k_2(q,\dot{q})\right) - g \\ 0 \end{bmatrix}$
$w = \mathrm{A}\uparrow$		$\begin{bmatrix} \frac{1}{m_h}\left(-k_2(q,\dot{q})\right) - g \\ \frac{1}{m_t}\left(k_2(q,\dot{q})\right) - g \end{bmatrix}$

Table 4.1: **Discrete states and continuous dynamics for impact oscillator hybrid system models** (Sec. 4.6.1). Note that the continuous dynamics $\ddot{q}$ have the same general form for both the piecewise-linear and -nonlinear models, with the spring law k being a linear or nonlinear function of the continuous state $x = (q, \dot{q})$ depending on which model is considered.

with parameter values $m_h = 3, m_t = 1, g = 2$.

In our first demonstration the observed states are $q[1]$ and $q[2]$, position of the hip and foot, leaving the velocities unobserved:

$$\mathcal{H}_{\text{pos}}(x) = q. \tag{4.17}$$

State estimation results for this system are shown in Figure 4.8.

In the remainder of this subsection, we demonstrate the effects of the choices we made in our problem formulation (Sec. 4.1) and algorithm derivation (Sec. 4.2) using the piecewise-linear model as a running example. We also consider a variation where the measurements correspond to the leg length and velocity, which are more representative of the onboard measurements available to an autonomous robot operating outside of the laboratory.

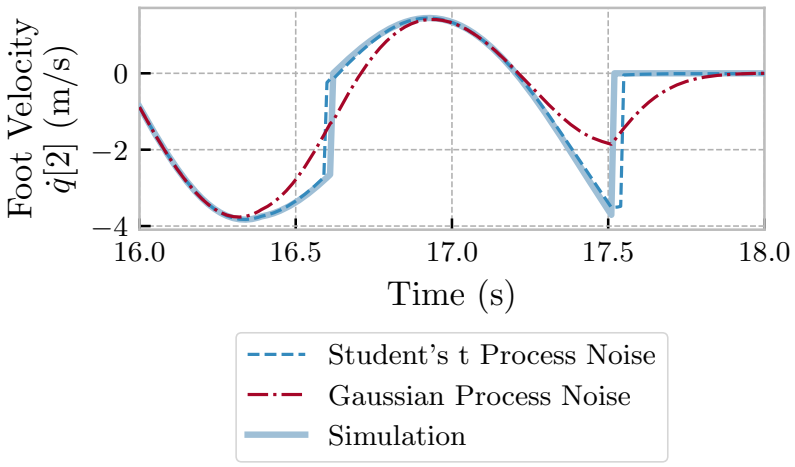


Figure 4.5: **The Student's t distribution process noise yields better estimates of instantaneous changes in continuous state** (Sec. 4.6.2). In this plot, estimates of the foot velocity are shown near two impacts ($\approx$ 16.6s, 17.5s).

Student's t versus Gaussian process noise

Figure 4.5 compares the estimation of foot velocity using Student's t with $r = 0.01$ versus using Gaussian for the process noise distribution; in both cases the true discrete state is given. The estimated trajectory for both distributions match the true simulated trajectory away from jumps, while near jumps, such as around times 16.6s and 17.5s, using the Student's t distribution enables closer tracking of the instantaneous change in the true foot velocity $\dot{q}[2]$ than when using a Gaussian distribution.

Gauss-Newton versus gradient (steepest) descent

We empirically compared convergence rates for continuous state x_t updates obtained using Gauss-Newton (algorithm 6) and gradient (steepest) descent directions. Figure 4.6 shows the log loss versus algorithm iteration for the two methods; the actual discrete state w_t was taken

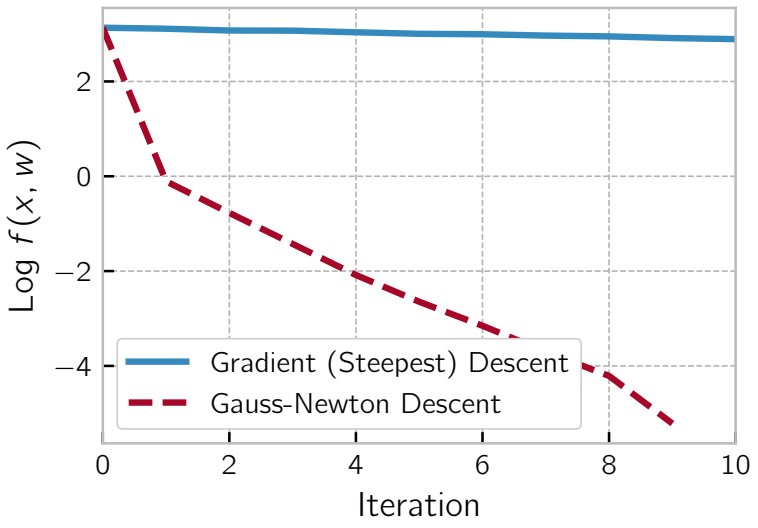


Figure 4.6: **Gauss-Newton descent directions yield faster convergence than gradient (steepest) descent** (Sec. 4.6.2). In this plot, the discrete state variables w are given and line 3 of algorithm 7 is modified to use either Gauss-Newton descent directions or gradient (steepest) descent to estimate the continuous state variables x by minimizing the relaxed objective function $f(x, w)$ (4.6).

as given to perform this comparison. As expected, the objective value decreases significantly faster when the search direction is determined by the Gauss-Newton scheme as compared to the direction of steepest descent, reaching the stopping criterion in ten times fewer iterations in our tests.

Smoothing the relaxed discrete state versus not

If the continuous states are given, the discrete state estimate returned by our algorithm (skipping line 3 of algorithm 7) is very close to the true discrete state regardless of whether a smoothing term is included in the relaxed problem formulation. When simultaneously estimating both the continuous and discrete states, the smoothing term becomes crucial, as

illustrated by comparing the discrete state estimates ($\widehat{w}_t$) in Figure 4.7 (without smoothing) and Figure 4.8 (with smoothing). In particular, the estimated discrete state switches rapidly without smoothing, whereas with smoothing the discrete state tends to remain constant for many samples and change mostly near ground-truth switching times.

Onboard versus offboard measurements

In the laboratory, the positions of the robot hip and foot can be directly measured offboard, e.g. with an external camera system. Outside of the laboratory, only the relative position of the hip and foot can be directly measured onboard our robot. Thus, we are motivated by this practical consideration to evaluate our algorithm's performance in the case where only the relative position and velocity of the hip and foot are measured,

$$\mathcal{H}_{\text{relative}}(x) = \begin{bmatrix} q[1] - q[2] \\ \dot{q}[1] - \dot{q}[2] \end{bmatrix}. \tag{4.18}$$

Although the full hybrid system state is formally unobservable with these relative measurements, our algorithm nevertheless yields good estimates of the discrete state as shown in Figure 4.9; due to large errors in the estimate of (unobservable) continuous states, we omit those results from the figure.

4.6.3 Piecewise-nonlinear impact oscillator experiment

To test Algorithm 7 on a nonlinear model, we included the kinematic constraints depicted in Figure 4.4b, resulting in a nonlinear spring force. In this model we set the two spring laws to be the same $k_1 = k_2$, decreasing the number of discrete states from four to two: $w = \text{A}$ when $q[2] > 0$ and $w = \text{G}$ when $q[2] = 0$. State estimation results compare favorably with the analogous results from the piecewise-linear system when using either absolute position measurements $\mathcal{H}_{\text{pos}}$ (4.17) (compare Figure 4.10 with Figure 4.8) or relative measurements $\mathcal{H}_{\text{relative}}$ (4.18) (compare Figure 4.11 with Figure 4.9).

In Figure 4.10 we see that the model can estimate continuous and discrete states in the

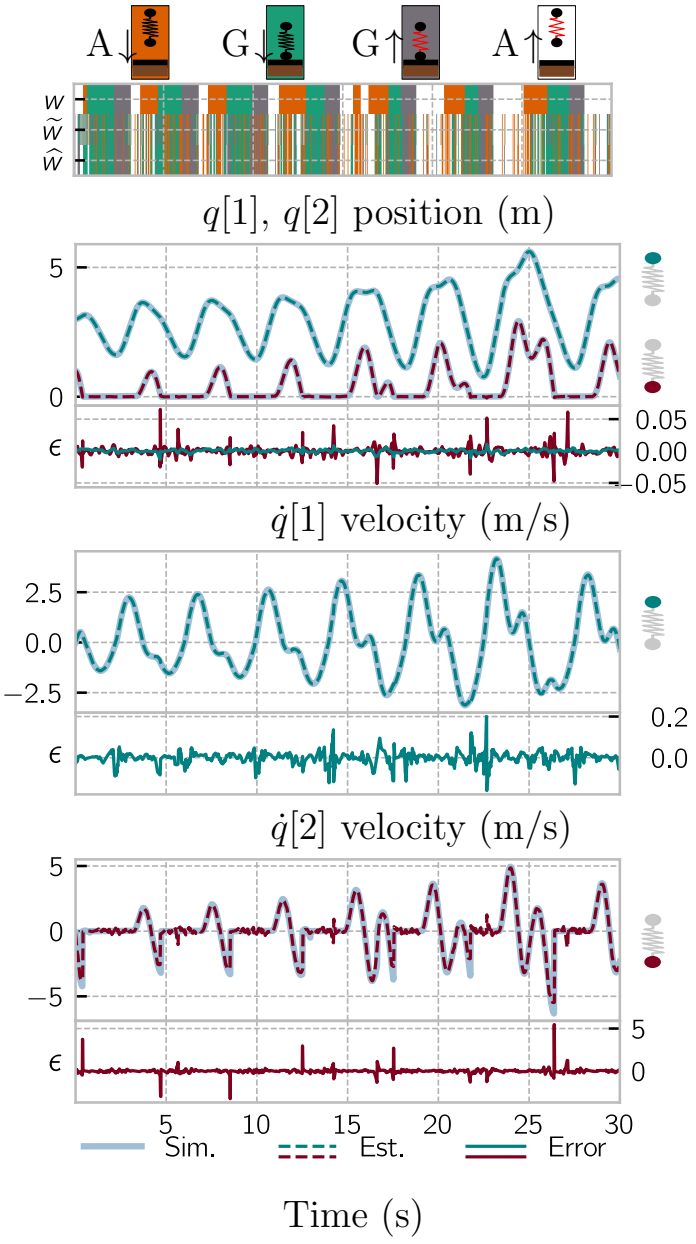


Figure 4.7: **Without smoothing ($\nu = 0$), the discrete state estimate switches frequently** (Sec. 4.6.2). The top plot shows the true discrete state of the system $w \in D^M$, the relaxed discrete state estimate $\widetilde{w} \in \Delta^M$, and the discrete state estimate $\widehat{w} \in D^M$ for a simulation of the piecewise-linear system. The subsequent plots show the estimate, simulation, and error ϵ values for position and velocity of the hip $q[1]$ and foot $q[2]$.

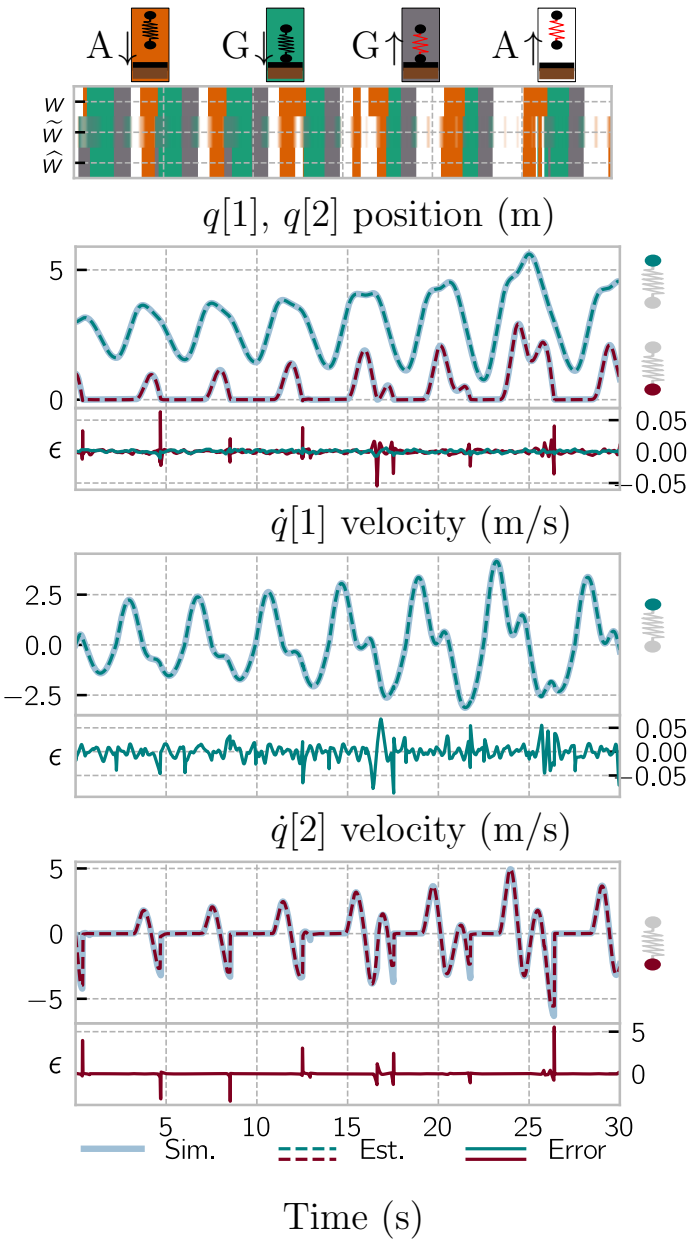


Figure 4.8: **With smoothing ($\nu > 0$), the discrete state estimate mostly switches near the true switching times.** (Sec. 4.6.2). This plot shows results from the piecewise-linear system; the notational and plotting conventions are adopted from Figure 4.7.

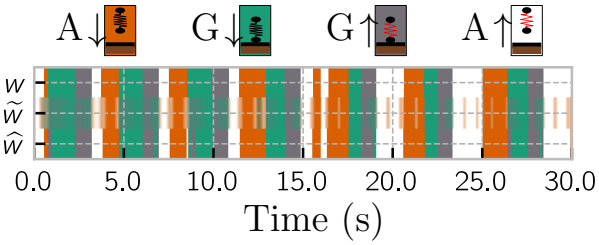


Figure 4.9: **Estimated discrete state using onboard (relative position and velocity) measurements $\mathcal{H}_{\text{relative}}$ (4.18) for the piecewise-linear system closely matches true discrete state.** (Sec. 4.6.2). Continuous state estimates are not shown since they are formally unobservable using only onboard measurements (in practice, they drift away from ground truth over time).

nonlinear setting. However, we do notice that the estimated trajectories are not as close to ground truth as in the linear case. In particular, when $q[2]$ has a value only slightly greater than 0 (e.g. between times 3s and 4s), the algorithm fails to detect the transition between $w = \text{A}$ and $w = \text{G}$.

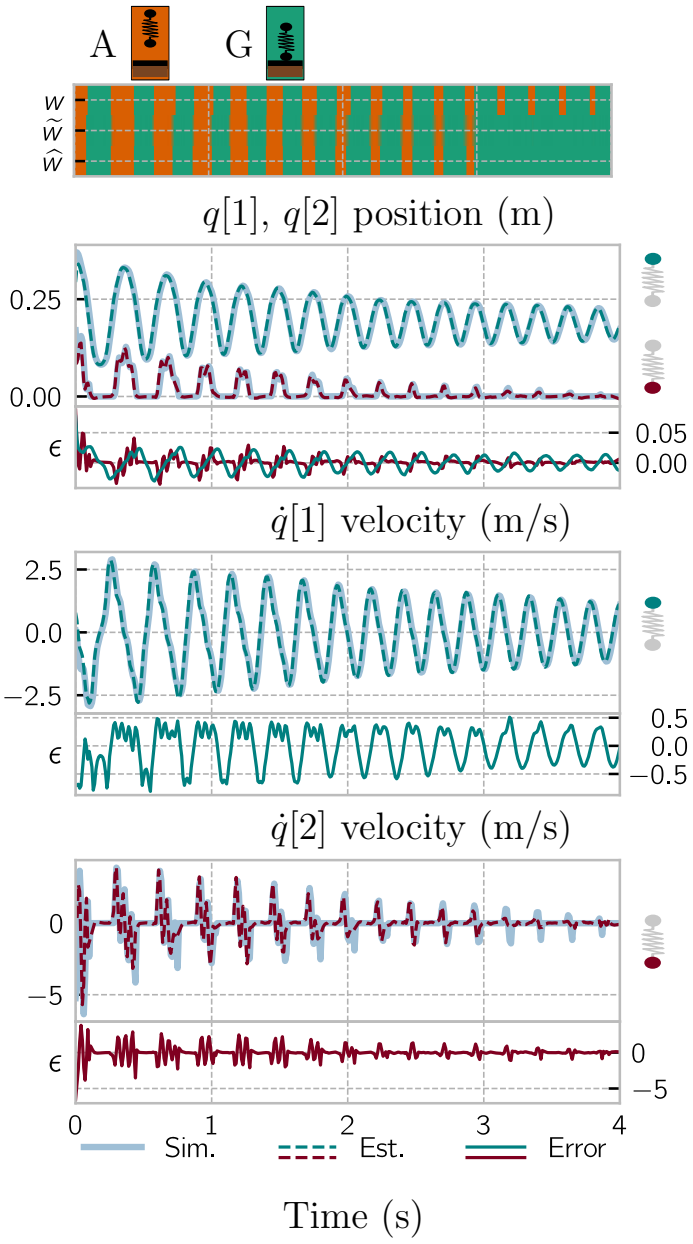


Figure 4.10: **Continuous and discrete states estimated for the piecewise-nonlinear model** (Sec. 4.6.3). Notational and plotting conventions are adopted from Figure 4.7; note that this model only has two discrete states (Sec. 4.6.1).

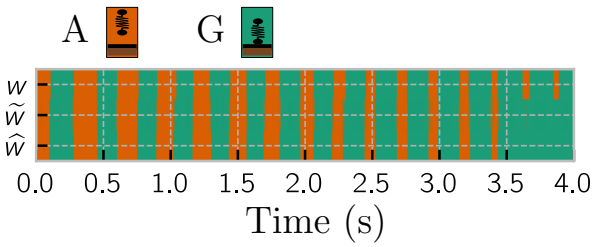


Figure 4.11: **Estimated discrete state using onboard (relative position and velocity) measurements $\mathcal{H}_{\text{relative}}$ (4.18) for the piecewise-nonlinear system closely matches true discrete state.** (Sec. 4.6.2). As with Figure 4.9, continuous state estimates are not shown since they drift from the true values over time; note that this nonlinear model only has two discrete states (Sec. 4.6.1).

Chapter 5

DISCUSSION

We discussed two applications in this thesis: portfolio selection and hybrid systems inference.

For portfolio selection, we proposed a new approach for general cardinality-constrained optimization, combining proximal gradient descent with recently developed relaxation technique. The approach is generalizable, computationally efficient, and is guaranteed to converge to a stationary point of the original problem. We do need to introduce additional hyperparameters however, such as the weight for quadratic penalty term during relaxation, in the proposed algorithm, which must be selected in practice. We showed that the quality of solutions found using our approach is high by comparing ours against global optima on small problems where exhaustive search is possible. While not performed in our work, it would be interesting and valuable to compile benchmark datasets for portfolio selection and to see how our method compares with other non-gradient based methods, including mixed integer optimization and genetic algorithms.

For hybrid systems inference, we proposed a new state estimation algorithm, analyzed its convergence properties, compared with IMM, and evaluated its performance on piecewise-linear and -nonlinear hybrid systems with non–identity resets. The model formulation leverages ideas from a well-developed research stream of generalized robust Kalman smoothing, able to capture the dynamic problem structure while allowing nonlinear dynamics and non – identity resets without hard-coded functions. The algorithm incorporates recently-developed nonsmooth variable projection technique with customized Gauss-Newton update, and its effectiveness was demonstrated on hybrid mechanical systems undergoing impact. While the approach performs decently well on the examples we looked at, we have not tested it out on

more complex systems in order to investigate both its overall performance and robustness to hyperparemeter selection. Another more practical concern is that while we assumed all dynamics are known in our model, that may rarely be the case in real scenarios. Sensitivity to noisy dynamics is certainly a question worth pondering, and our current approach does not yet take that into consideration. A further step would be to consider the problem of system identification in the complex setting of hybrid systems, where multiple systems may need to be inferred. This opens up a research area to a new type of physics-based unsupervised learning, which has both a lot of challenges and many potential applications.

BIBLIOGRAPHY

[1] R. Rockafellar and R. J.-B. Wets, Variational Analysis. Heidelberg, Berlin, New York: Springer Verlag, 1998.

[2] H. H. Bauschke, P. L. Combettes et al., Convex analysis and monotone operator theory in Hilbert spaces. Springer, 2011, vol. 408.

[3] Y. Nesterov, "Gradient methods for minimizing composite functions," Mathematical Programming, vol. 140, no. 1, pp. 125–161, 2013.

[4] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, "Stochastic subgradient method converges on tame functions," Foundations of computational mathematics, vol. 20, no. 1, pp. 119–154, 2020.

[5] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, "First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems," SIAM Journal on Optimization, vol. 28, no. 3, pp. 2131–2151, 2018.

[6] D. Drusvyatskiy and C. Paquette, "Efficiency of minimizing compositions of convex functions and smooth maps," Mathematical Programming, vol. 178, no. 1-2, pp. 503–558, 2019.

[7] D. Davis and D. Drusvyatskiy, "Stochastic model-based minimization of weakly convex functions," SIAM Journal on Optimization, vol. 29, no. 1, pp. 207–239, 2019.

[8] C. Blair, "Problem complexity and method efficiency in optimization (as nemirovsky and db yudin)," SIAM Review, vol. 27, no. 2, p. 264, 1985.

[9] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Accelerated methods for nonconvex optimization," SIAM Journal on Optimization, vol. 28, no. 2, pp. 1751–1772, 2018.

[10] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization or nonconvex and nonsmooth problems," Mathematical Programming, vol. 146, no. 1-2, pp. 459–494, 2014.

[11] A. S. Lewis and S. J. Wright, "A proximal method for composite minimization," Mathematical Programming, vol. 158, no. 1-2, pp. 501–546, 2016.

[12] J. V. Burke, "Descent methods for composite nondifferentiable optimization problems," Mathematical Programming, vol. 33, no. 3, pp. 260–279, 1985.

[13] Y. Nesterov, "Modified gauss–newton scheme with worst case guarantees for global performance," Optimisation methods and software, vol. 22, no. 3, pp. 469–483, 2007.

[14] B. M. Bell and J. V. Burke, "Algorithmic differentiation of implicit functions and optimal values," in Advances in Automatic Differentiation. Springer, 2008, pp. 67–77.

[15] A. Aravkin, D. Drusvyatskiy, and T. van Leeuwen, "Variable projection without smoothness," arXiv preprint arXiv:1601.05011, 2016.

[16] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," SIAM Journal on Imaging Sciences, vol. 2, no. 1, pp. 183–202, 2009.

[17] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Lower bounds for finding stationary points i," Mathematical Programming, pp. 1–50, 2019.

[18] A. Y. Aravkin, J. V. Burke, and G. Pillonetto, "Optimization viewpoint on kalman smoothing with applications to robust and sparse estimation," in Compressed sensing & sparse filtering. Springer, 2014, pp. 237–280.

[19] C. C. Paige and M. Saunders, "Least squares estimation of discrete linear dynamic systems using orthogonal transformations," SIAM Journal on Numerical Analysis, vol. 14, no. 2, pp. 180–193, 1977.

[20] L. Fahrmeir and H. Kaufmann, "On kalman filtering, posterior mode estimation and fisher scoring in dynamic exponential family regression," Metrika, vol. 38, no. 1, pp. 37–60, 1991.

[21] B. M. Bell and F. W. Cathey, "The iterated kalman filter update as a gauss-newton method," IEEE Transactions on Automatic Control, vol. 38, no. 2, pp. 294–297, 1993.

[22] B. M. Bell, "The iterated kalman smoother as a gauss–newton method," SIAM Journal on Optimization, vol. 4, no. 3, pp. 626–636, 1994.

[23] B. M. Bell, J. V. Burke, and G. Pillonetto, "An inequality constrained nonlinear kalman–bucy smoother by interior point likelihood maximization," Automatica, vol. 45, no. 1, pp. 25–33, 2009.

[24] A. Y. Aravkin, B. M. Bell, J. V. Burke, and G. Pillonetto, "An l1-laplace robust kalman smoother," IEEE Transactions on Automatic Control, vol. 56, no. 12, pp. 2898–2911, 2011.

[25] S. Farahmand, G. B. Giannakis, and D. Angelosante, "Doubly robust smoothing of dynamical processes via outlier sparsity constraints," IEEE Transactions on Signal Processing, vol. 59, no. 10, pp. 4529–4543, 2011.

[26] A. Y. Aravkin, J. V. Burke, and G. Pillonetto, "Sparse/robust estimation and kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory," The Journal of Machine Learning Research, vol. 14, no. 1, pp. 2689–2728, 2013.

[27] L. Fahrmeir and R. Künstler, "Penalized likelihood smoothing in robust state space models," Metrika, vol. 49, no. 3, pp. 173–191, 1999.

[28] A. Y. Aravkin, J. V. Burke, and G. Pillonetto, "Robust and trend-following student's t kalman smoothers," SIAM Journal on Control and Optimization, vol. 52, no. 5, pp. 2891–2916, 2014.

[29] J. Jonker, A. Aravkin, J. V. Burke, G. Pillonetto, and S. Webster, "Fast robust methods for singular state-space models," Automatica, vol. 105, pp. 399–405, 2019.

[30] A. Aravkin, J. V. Burke, L. Ljung, A. Lozano, and G. Pillonetto, "Generalized kalman smoothing: Modeling and algorithms," Automatica, vol. 86, pp. 63–86, 2017.

[31] H. Markowitz, "Portfolio selection," Journal of Finance, vol. 7, no. 1, pp. 77–91, 1952.

[32] V. DeMiguel, L. Garlappi, F. J. Nogales, and R. Uppal, "A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms," Management science, vol. 55, no. 5, pp. 798–812, 2009.

[33] M. S. Lobo, M. Fazel, and S. Boyd, "Portfolio optimization with linear and fixed transaction costs," Annals of Operations Research, vol. 152, no. 1, pp. 341–365, 2007.

[34] P. Jorion, "Portfolio optimization with tracking-error constraints," Financial Analysts Journal, vol. 59, no. 5, pp. 70–82, 2003.

[35] P. N. Kolm, R. Tütüncü, and F. J. Fabozzi, "60 years of portfolio optimization: Practical challenges and current trends," European Journal of Operational Research, vol. 234, no. 2, pp. 356–371, 2014.

[36] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," Journal of banking & finance, vol. 26, no. 7, pp. 1443–1471, 2002.

[37] T.-J. Chang, N. Meade, J. E. Beasley, and Y. M. Sharaiha, "Heuristics for cardinality constrained portfolio optimisation," Computers & Operations Research, vol. 27, no. 13, pp. 1271–1302, 2000.

[38] H. Soleimani, H. R. Golmakani, and M. H. Salimi, "Markowitz-based portfolio selection with minimum transaction lots, cardinality constraints and regarding sector capitalization using genetic algorithm," Expert Systems with Applications, vol. 36, no. 3, pp. 5058–5063, 2009.

[39] G.-F. Deng, W.-T. Lin, and C.-C. Lo, "Markowitz-based portfolio selection with cardinality constraints using improved particle swarm optimization," Expert Systems with Applications, vol. 39, no. 4, pp. 4558–4566, 2012.

[40] D. X. Shaw, S. Liu, and L. Kopman, "Lagrangian relaxation procedure for cardinality-constrained portfolio optimization," Optimisation Methods & Software, vol. 23, no. 3, pp. 411–420, 2008.

[41] D. Bertsimas and R. Shioda, "Algorithm for cardinality-constrained quadratic optimization," Computational Optimization and Applications, vol. 43, no. 1, pp. 1–22, 2009.

[42] X. T. Cui, X. J. Zheng, S. S. Zhu, and X. L. Sun, "Convex relaxations and miqcqp reformulations for a class of cardinality-constrained portfolio selection problems," J. of Global Optimization, vol. 56, no. 4, pp. 1409–1423, Aug. 2013. [Online]. Available: http://dx.doi.org/10.1007/s10898-012-9842-2

[43] W. Murray and H. Shek, "A local relaxation method for the cardinality constrained portfolio optimization problem," Computational Optimization and Applications, vol. 53, no. 3, pp. 681–709, 2012.

[44] A. Beck and Y. C. Eldar, "Sparsity constrained nonlinear optimization: Optimality conditions and algorithms," SIAM Journal on Optimization, vol. 23, no. 3, pp. 1480–1509, 2013.

[45] G. Banjac and P. J. Goulart, "A novel approach for solving convex problems with cardinality constraints," 2017.

[46] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," Multiscale Modeling & Simulation, vol. 4, no. 4, pp. 1168–1200, 2005.

[47] T. Leung and X. Li, Optimal Mean Reversion Trading: Mathematical Analysis and Practical Applications, ser. Modern Trends in Financial Engineering. World Scientific, Singapore, 2016.

[48] ——, "Optimal mean reversion trading with transaction costs and stop-loss exit," International Journal of Theoretical & Applied Finance, vol. 18, no. 3, p. 15500, 2015.

[49] Y. Kitapbayev and T. Leung, "Optimal mean-reverting spread trading: nonlinear integral equation approach," Annals of Finance, vol. 13, no. 2, pp. 181–203, 2017.

[50] E. Gatev, W. Goetzmann, and K. Rouwenhorst, "Pairs trading: Performance of a relative-value arbitrage rule," Review of Financial Studies, vol. 19, no. 3, pp. 797–827, 2006.

[51] Z. Zhao and D. P. Palomar, "Mean-reverting portfolio with budget constraint," IEEE Transactions on Signal Processing, vol. 66, no. 9, pp. 2342–2357, 2018.

[52] Z. Zhao, R. Zhou, and D. P. Palomar, "Optimal mean-reverting portfolio with leverage constraint for statistical arbitrage in finance," IEEE Transactions on Signal Processing, 2019.

[53] A. d'Aspremont, "Identifying small mean-reverting portfolios," Quantitative Finance, vol. 11, no. 3, pp. 351–364, 2011.

[54] L. S. Ornstein and G. E. Uhlenbeck, "On the theory of the Brownian motion," Physical Review, vol. 36, pp. 823–841, 1930.

[55] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods," Mathematical Programming, vol. 137, no. 1-2, pp. 91–129, 2013.

[56] R. T. Rockafellar and R. J.-B. Wets, Variational analysis. Springer Science & Business Media, 2009, vol. 317.

[57] R. R. Stengel, Optimal Control and Estimation. Dover, 1994.

[58] L. Menini and A. Tornambe, "Asymptotic tracking of periodic trajectories for a simple mechanical system subject to nonsmooth impacts," IEEE Transactions on Automatic Control, vol. 46, no. 7, pp. 1122–1126, Jul. 2001.

[59] A. Balluchi, L. Benvenuti, M. D. Di Benedetto, and A. L. Sangiovanni-Vincentelli, "Design of Observers for Hybrid Systems," in Proceedings of Hybrid Systems: Computation and Control (HSCC), vol. 2289. Springer Berlin Heidelberg, 2002, pp. 76–89. [Online]. Available: http://link.springer.com/10.1007/3-540-45873-5_9

[60] A. Balluchi, L. Benvenuti, M. Di Benedetto, and A. Sangiovanni-Vincentelli, "Observability for hybrid systems," in Proceedings of the IEEE Conference on Decision and Control, vol. 2. IEEE, 2003, pp. 1159–1164. [Online]. Available: http://ieeexplore.ieee.org/document/1272764/

[61] D. Gómez-Gutiérrez, S. Čelikovský, A. Ramírez-Treviño, J. Ruiz-Léon, and S. D. Gennaro, "Sliding mode observer for Switched Linear Systems," in IEEE International Conference on Automation Science and Engineering, 2011, pp. 725–730.

[62] N. Barhoumi, F. Msahli, M. Djemaï, and K. Busawon, "Observer design for some classes of uniformly observable nonlinear hybrid systems," Nonlinear Analysis: Hybrid Systems, vol. 6, no. 4, pp. 917–929, 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1751570X12000064

[63] A. Balluchi, L. Benvenuti, M. D. Di Benedetto, and A. Sangiovanni-Vincentelli, "The design of dynamical observers for hybrid systems: Theory and application to an automotive control problem," Automatica, vol. 49, no. 4, pp. 915–925, 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0005109813000381

[64] H. A. P. Blom and E. A. Bloem, "Exact Bayesian and particle filtering of stochastic hybrid systems," IEEE Transactions on Aerospace and Electronic Systems, vol. 43, no. 1, pp. 55–70, 2007.

[65] A. Doucet, N. J. Gordon, and V. Krishnamurthy, "Particle filters for state estimation of jump Markov linear systems," IEEE Transactions on Signal Processing, vol. 49, no. 3, pp. 613–624, 2001.

[66] C. E. Seah and I. Hwang, "State Estimation for Stochastic Linear Hybrid Systems with Continuous-State-Dependent transitions: An IMM approach," IEEE Transactions on Aerospace and Electronic Systems, vol. 45, no. 1, pp. 376–392, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1109/TAES.2009.4805286

[67] A. Bemporad, D. Mignone, and M. Morari, "Moving horizon estimation for hybrid systems and fault detection," in Proceedings of the American Control Conference, vol. 4, Jun. 1999, pp. 2471–2475 vol.4.

[68] G. Ferrari-Trecate, D. Mignone, and M. Morari, "Moving horizon estimation for hybrid systems," IEEE Transactions on Automatic Control, vol. 47, no. 10, pp. 1663–1676, Oct. 2002.

[69] A. Alessandri, M. Baglietto, and G. Battistelli, "Receding-horizon estimation for switching discrete-time linear systems," IEEE Transactions on Automatic Control, vol. 50, no. 11, pp. 1736–1748, Nov. 2005.

[70] ——, "Minimum-Distance Receding-Horizon State Estimation for Switching Discrete-Time Linear Systems," in Assessment and Future Directions of Nonlinear Model

Predictive Control, ser. Lecture Notes in Control and Information Sciences, R. Findeisen, F. Allgöwer, and L. T. Biegler, Eds. Springer-Verlag Berlin Heidelberg, 2007, no. 358, pp. 348–366.

[71] L. Bako and S. Lecoeuche, "A sparse optimization approach to state observer design for switched linear systems," Systems & Control Letters, vol. 62, no. 2, pp. 143–151, Feb. 2013.

[72] S. C. Johnson, "Observability and observer design for switched linear systems," Ph.D. dissertation, Purdue University, Dec. 2016.

[73] G. Ferrari-Trecate, D. Mignone, and M. Morari, "Moving horizon estimation for hybrid systems," IEEE transactions on automatic control, vol. 47, no. 10, pp. 1663–1676, Oct. 2002. [Online]. Available: http://dx.doi.org/10.1109/TAC.2002.802772

[74] A. Aravkin, J. V. Burke, L. Ljung, A. Lozano, and G. Pillonetto, "Generalized Kalman smoothing: Modeling and algorithms," Automatica, vol. 86, pp. 63–86, Dec. 2017.

[75] A. Aravkin, J. V. Burke, and G. Pillonetto, "Robust and Trend-following Kalman Smoothers using Student's t," IFAC Proceedings Volumes, vol. 45, no. 16, pp. 1215–1220, Jul. 2012.

[76] R. Goebel, R. Sanfelice, and A. Teel, "Hybrid dynamical systems," IEEE Control Systems, vol. 29, no. 2, pp. 28–93, Apr. 2009.

[77] A. M. Johnson, S. A. Burden, and D. E. Koditschek, "A hybrid systems model for simple manipulation and self-manipulation systems," The International Journal of Robotics Research, vol. 35, no. 11, pp. 1354–1392, Sep. 2016.

[78] J. P. Hespanha and A. S. Morse, "Stability of switched systems with average dwell-time," in Proceedings of the IEEE Conference on Decision and Control, vol. 3, Dec. 1999, pp. 2655–2660.

[79] R. Vidal, A. Chiuso, S. Soatto, and S. Sastry, "Observability of Linear Hybrid Systems," in Hybrid Systems: Computation and Control, O. Maler and A. Pnueli, Eds. Springer Berlin Heidelberg, 2003, vol. 2623, pp. 526–539. [Online]. Available: http://link.springer.com/10.1007/3-540-36580-X_38

[80] P. Lötstedt, "Mechanical Systems of Rigid Bodies Subject to Unilateral Constraints," SIAM Journal on Applied Mathematics, vol. 42, no. 2, pp. 281–296, 1982.

[81] N. Fazeli, S. Zapolsky, E. Drumwright, and A. Rodriguez, "Learning Data-Efficient Rigid-Body Contact Models: Case Study of Planar Impact," in Proceedings of the Conference on Robot Learning, 2017, p. 10.

[82] J. Duník, O. Straka, O. Kost, and J. Havlík, "Noise covariance matrices in state-space models: A survey and comparison of estimation methods—part i," International Journal of Adaptive Control and Signal Processing, vol. 31, no. 11, pp. 1505–1543, 2017.

[83] H. A. P. Blom and Y. Bar-Shalom, "The Interacting Multiple Model Algorithm for Systems with Markovian Switching Coefficients," IEEE Transactions on Automatic Control, vol. 33, no. 8, 1988.

[84] R. Labbe, "FilterPy," 2014–, [Online]. [Online]. Available: https://github.com/rlabbe/filterpy

[85] M. Di Bernardo, C. Budd, A. Champneys, and P. Kowalczyk, Piecewise-Smooth Dynamical Systems: Theory and Applications, ser. Applied Mathematical Sciences. Springer, 2008, no. 163.

[86] M. Schatzman, "Uniqueness and continuous dependence on data for one–dimensional impact problems," Mathematical and Computer Modelling, vol. 28, no. 4—8, pp. 1–18, 1998. [Online]. Available: http://dx.doi.org/10.1016/S0895-7177(98)00104-6

[87] G. Kenneally, A. De, and D. E. Koditschek, "Design Principles for a Family of Direct-Drive Legged Robots," IEEE Robotics and Automation Letters, vol. 1, no. 2, pp. 900–907, Jul. 2016.

Appendix A

PARTIAL MINIMIZATION IN SECTION ??

We start with the f_1 subproblem. Taking $\partial_\theta f = 0$, we get

$$\begin{aligned} 0 = \frac{\partial f}{\partial \theta} &= (1-c)\mathbf{1}^T(\theta(1-c) - A(c)w) \\ \Rightarrow \theta^* &= \frac{\mathbf{1}^T(x(w)_{1:T} - cx(w)_{0:T-1})}{T(1-c)}. \end{aligned}$$

Plugging $\theta^*(c, w)$ into f, we get an explicit form of f_1:

$$\begin{aligned} f_1(w, a, c) = &\frac{1}{2}\ln(a) + \gamma c \\ &+ \frac{1}{2Ta}\|B(x(w)_{1:T} - cx(w)_{0:T-1})\|^2 \end{aligned} \tag{A.1}$$

with $B = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{T}$ a projection matrix onto the space of vectors in $\mathbb{R}^T$ with mean 0.

We now minimize with respect to c.

$$\begin{aligned} 0 = \frac{\partial f_1}{\partial c} &= \gamma + \frac{1}{Ta}x_{0:T-1}^T B^T B(cx(w)_{0:T-1} - x(w)_{1:T}) \\ \Rightarrow c^*(a, w) &= \frac{(Bx(w)_{0:T-1})^T(Bx(w)_{1:T}) - Ta\gamma}{\|Bx(w)_{0:T-1}\|^2}. \end{aligned}$$

We can use this expression to explicitly write $f_2(w, a)$:

$$\begin{aligned} f_2(a, w) = &\frac{1}{2}\ln(a) + \gamma c^*(a, w) \\ &+ \frac{1}{2Ta}\|B(x(w)_{1:T} - c^*(a, w)x(w)_{0:T-1})\|^2. \end{aligned} \tag{A.2}$$

Remark 8. *The denominator in $c^*(a, w)$ equals zero when $Bx(w)_{0:T-1} = 0$, which indicates*

$$x(w)_{0:T-1} = \frac{\mathbf{1}^T x(w)_{0:T-1}}{T}\mathbf{1}.$$

This implies that $x(w)$ must be constant over time and exactly equal to its mean, which is very unlikely with stock market data. Therefore it is reasonable to assume that $\|Bx(w)_{0:T-1}\|^2 > 0$.

There is no guarantee that the numerator will be positive. Indeed, the optimal c^* can potentially be negative, in which case no corresponding positive μ exists. This means that the given data does not permit the construction of a mean-reverting time series. The γ term in numerator drives c^* towards negative values, which means that the higher mean-reverting level we request, the less likely such a process can be constructed.

Special case: when $\gamma = 0$, $c^*(a, w) = c^*(w)$ is independent of a. Denote

$$c^* = c^*(w), \quad b_1 = Bx(w)_{1:T}, \quad b_0 = Bx(w)_{0:T-1}.$$

We can now minimize f_2 (A.2) in a:

$$\begin{aligned} 0 = \frac{\partial f_2}{\partial a} &= \frac{1}{2a} - \frac{1}{2Ta^2}\|b_1 - c^* b_0\|^2 \\ \Rightarrow a^* &= \frac{\|b_1 - c^* b_0\|^2}{T} = \frac{1}{T}\left(\|b_1\|^2 - \frac{(b_0^T b_1)^2}{\|b_0\|^2}\right). \end{aligned} \tag{A.3}$$

Plugging a^* in, we get a closed-form expression for f_3:

$$\begin{aligned} f_3(w) &= \frac{1}{2}\ln(a^*) + \frac{1}{2Ta^*}\|b_1 - c^* b_0\|^2 \\ &= \frac{1}{2}\ln(a^*) + \frac{\|b_1\|^2}{2Ta^*} - \frac{(b_0^T b_1)^2}{2Ta^*\|b_0\|^2} \end{aligned} \tag{A.4}$$

More succinctly,

$$f_3(w) = \frac{1}{2}\ln(a^*) + \frac{1}{2Ta^*}(Ta^*) = \frac{1}{2}\ln(a^*).$$

The optimization problem in this case reduces to

$$\min_{\|w\|_1=1, \|w\|_0 \le \eta} \frac{1}{2}\ln(a^*) \equiv \frac{1}{2}\ln(\|BA(c^*(w))w\|^2). \tag{A.5}$$

Remark 9. *The objective function of* (A.5) *is an even function. Hence the problem does not have a unique minimizer.*

General case: $\gamma > 0$**.** Here, c^* depends on a. Denote $c^* = c^*(a, w)$. We have

$$f_3(w) = \min_a \frac{1}{2}\ln(a) + \frac{\|b_1\|^2}{2Ta} - \frac{(b_0^T b_1 - Ta\gamma)^2}{2Ta\|b_0\|^2},$$

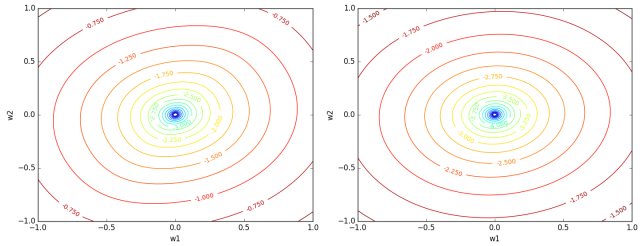


Figure A.1: Contour plot of (A.2) $\gamma \neq 0$ (left) and $\gamma = 0$ (right) for $w \in \mathbb{R}^2$.

$$= \frac{\gamma b_0^T b_1}{\|b_0\|^2} + \min_a \frac{1}{2}\ln(a) + \frac{\|b_1\|^2}{2Ta} - \frac{(b_0^T b_1)^2}{2Ta\|b_0\|^2} - \frac{Ta\gamma^2}{2\|b_0\|^2}. \tag{A.6}$$

Solving (A.6), we obtain the optimal a in closed form:

$$a^* = \frac{\|b_0\|^2}{2T\gamma^2} - \frac{\sqrt{\|b_0\|^4 - 4\gamma^2(\|b_0\|^2\|b_1\|^2 - (b_0^T b_1)^2)}}{2T\gamma^2}$$

and correspondingly

$$c^* = \frac{b_0^T b_1}{\|b_0\|^2} - \frac{1}{2\gamma} + \sqrt{\frac{(b_0^T b_1)^2}{\|b_0\|^4} + \frac{1}{4\gamma^2} - \frac{\|b_1\|^2}{\|b_0\|^2}}.$$

In these expressions, b_0 and b_1 are functions of w. The optimal solution a^* increases with respect to γ and c^* decreases with respect to γ when

$$0 < \gamma < \frac{1}{2}\sqrt{\frac{\|b_0\|^4}{-(b_1^T b_0)^2 + \|b_0\|^2\|b_1\|^2}}.$$

As $\gamma \to 0$, $a^* \to \frac{1}{T}\left(\|b_1\|^2 - \frac{(b_0^T b_1)^2}{\|b_0\|^2}\right)$ and $c^* \to \frac{b_0^T b_1}{\|b_0\|^2}$. These limits correspond to the optimal a and c derived in the section with $\gamma = 0$. We can also write down the final optimization problem in closed form:

$$f_3(w) = \frac{1}{2}\ln(a^*) + \frac{\|b_1\|^2}{2Ta^*} - \frac{(b_0^T b_1)^2}{2Ta^*\|b_0\|^2} - \frac{Ta\gamma^2}{2\|b_0\|^2} + \frac{\gamma b_0^T b_1}{\|b_0\|^2},$$

where b_i and a^* are all functions of w as detailed above. When $\gamma = 0$, we recover (A.4).

Figure A.1 illustrates the effect of γ on the shape of contours for a simple case where w has dimension 2. The contours are rotated by γ, which can affect which assets are selected (as it can change the intersection points with the 1-norm ball).

Appendix B

MATRICES USED IN HYBRID SYSTEM SIMULATION

The $\mathcal{G}_i$ maps are

$$\mathcal{G}_i(x) = x + \Delta_t A_i x + b_i$$

where

$$A_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{k}{m_b} & 0 & \frac{k}{m_b} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{k}{m_f} & 0 & -\frac{k}{m_f} & 0 \end{bmatrix}, b_1 = \begin{bmatrix} 0 \\ -g + \frac{kl_0}{m_b} \\ 0 \\ -g - \frac{kl_0}{m_f} \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{k}{m_b} & 0 & \frac{k}{m_b} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, b_2 = \begin{bmatrix} 0 \\ -g + \frac{kl_0}{m_b} \\ 0 \\ 0 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{\alpha k}{m_b} & 0 & \frac{\alpha k}{m_b} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{\alpha k}{m_f} & 0 & -\frac{\alpha k}{m_f} & 0 \end{bmatrix}, b_3 = \begin{bmatrix} 0 \\ -g + \frac{\alpha k l_0}{m_b} \\ 0 \\ -g - \frac{\alpha k l_0}{m_f} \end{bmatrix}, A_4 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{\alpha k}{m_b} & 0 & \frac{\alpha k}{m_b} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, b_4 = \begin{bmatrix} 0 \\ -g + \frac{\alpha k l_0}{m_b} \\ 0 \\ 0 \end{bmatrix}$$

and

$$\mathcal{H}(x) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} x.$$

Also,

$$Q = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}, R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

When we parametrize $\mathcal{G}_i$'s by u,

$$A_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -u_1 & 0 & u_1 & 0 \\ 0 & 0 & 0 & 1 \\ u_2 & 0 & -u_2 & 0 \end{bmatrix}, b_1 = \begin{bmatrix} 0 \\ u_3 \\ 0 \\ u_4 \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -u_1 & 0 & u_1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, b_2 = \begin{bmatrix} 0 \\ u_3 \\ 0 \\ 0 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -u_5 & 0 & u_5 & 0 \\ 0 & 0 & 0 & 1 \\ u_6 & 0 & -u_6 & 0 \end{bmatrix}, b_3 = \begin{bmatrix} 0 \\ u_7 \\ 0 \\ u_8 \end{bmatrix}, A_4 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -u_5 & 0 & u_5 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, b_4 = \begin{bmatrix} 0 \\ u_7 \\ 0 \\ 0 \end{bmatrix}$$