Nonconvex Projections Arising in Bilinear Mathematical Models

by

Manish Krishan Lal

B.Sc., Kurukshetra University, 2012
M.Sc., St. Stephen's College, University of Delhi, 2014
M.Tech., Indian Institute of Technology Madras, 2017

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE COLLEGE OF GRADUATE STUDIES

(Mathematics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

December 2023

The following individuals certify that they have read, and recommend to the College of Graduate Studies for acceptance, a thesis/dissertation entitled:

submitted by Manish Krishan Lal in partial fulfilment of the requirements of the degree of Doctor of Philosophy

Xianfu Wang, Irving K. Barber Faculty of Science
Supervisor

Heinz Bauschke, Irving K. Barber Faculty of Science
Co-supervisor

John Braun, Irving K. Barber Faculty of Science
Supervisory Committee Member

Apurva Narayan, Western University
Supervisory Committee Member

Julian Cheng, School of Engineering
University Examiner

Henry Wolkowicz, University of Waterloo
External Examiner

Abstract

This thesis contributes to the study of projection operators associated with bilinear sets. Bilinear sets are not convex and appear in many applications such as deep learning, inverse problems, and other bilinear models in control and optimization.

The closed-form projection formulas for some of these bilinear sets namely crosses, hyperbolas, and hyperbolic paraboloids are provided. Along the way, a convenient presentation for the roots of cubic polynomials is highlighted and utilized further to develop more projection formulas and proximal mappings which are essential tools used in projection algorithms and proximal splitting algorithms.

The notion of Fejér monotonicity is instrumental in unifying the convergence proofs of many iterative methods, such as the Krasnoselskii-Mann iteration, the proximal point method, and the projection algorithms. In a finite-dimensional Hilbert space, the sequences generated by the proximal point algorithm enjoy directionally asymptotic properties. A comprehensive study of directionally asymptotical results of strongly convergent subsequences of Fejér monotone sequences in general Hilbert spaces is provided along with some detailed examples.

We also provide a foundational mathematical model of Elser's framework for matrix factorization.

Lay Summary

The minimum distance from a point to a line or any linear space is easily evaluated and is a well-known result in any Linear Algebra textbook. This point on the line is known as the projection point and can be expressed in a closed-form manner. However, finding closed-form projection formulas for bilinear nonconvex sets, such as conics sections and quadrics surfaces, has remained an open problem. This thesis serves as a stepping stone toward systematically addressing the challenge of deriving closed-form projections onto these bilinear sets in Hilbert spaces.

Recently, Elser introduced a feasibility-based model for embedding discrete network optimization problems in Euclidean space. To find feasible solutions, this model is complemented by splitting algorithms that make use of the aforementioned projections onto bilinear sets, such as hyperbolas and hyperbolic paraboloids. This thesis elucidates the Elser model using tensor notation for neural network training problems and explains the rise of hyperbolas and hyperbolic paraboloids during the modeling process.

Table of Contents

List of Figures

Acknowledgements

First of all, I wholeheartedly thank my supervisors Professor Heinz Bauschke and Professor Xianfu Wang. Your dedication and wisdom have not only shaped my academic path but have also enriched my life in countless ways. You are not just mentors, but integral chapters in my story, and I am profoundly grateful for your unwavering guidance, tireless encouragement, steadfast support, and the vast reservoir of knowledge you've generously shared. I am truly blessed to have had the opportunity to work under your mentorship.

Ein guter Lehrer ist wie ein Diamant: Er strahlt mit seinem Wissen und seiner Weisheit, und sein Wert ist unermesslich.

老如您，如金似，生感激不。恩如山，感激不。

I deeply thank Professor John Braun, Professor Julian Cheng, Professor Apurva Narayan, and Professor Henry Wolkowicz for serving on my thesis committee and for their constructive comments, and many suggestions.

Infinite thanks to my parents and my siblings Mohit and Kiran, four beats of my heart, who continue to support and help me through many challenges. Special thanks to my friends and colleagues.

Finally, I warmly thank the Scientific Education and Research Board (SERB), Department of Science and Technology, India and UBC for the financial support as SERB-UBC fellowship throughout the years of my doctoral studies.

I conclude by thanking my supervisors again with heartfelt quotes from my native languages, Sanskrit and Punjabi, passed down to me by my parents:

गुरोः प्रेरणां सृजति साधकः, ज्ञानं प्रकट्या शिष्यस्य । संसारे परमं साधनं, गुरोर्मूर्तिर्जगत्सखा ।।

ਗੁਰੂ ਦਾ ਧੰਨਵਾਦ ਹੈ ਜੋ ਉਸਨੂੰ ਸਿਖਾਉਂਦਾ ਹੈ, ਜੀਵਨ ਦੇ ਰਸਤੇ ਦਿਖਾਉਂਦਾ ਹੈ ਅਤੇ ਸਚਾਈ ਦੇ ਮਾਰਗ ਤੇ ਚਲਾਉਂਦਾ ਹੈ।

Great teachers inspire by illuminating knowledge for their disciples. Dem Lehrer sei Dank, er lehrt nicht nur mit Worten, sondern auch mit Taten. A teacher as a friend is an exceptional gateway to the universe of knowledge.

Dedication

To my parents and late grandparents,
whose blessings are with me eternally.

Chapter 1

Introduction

1.1 Motivation

In the fields of artificial intelligence and operations research, the concept of constraint satisfaction refers to the process of finding a solution by adhering to a set of constraints that establish conditions that variables must meet [68]. A solution is defined as a set of variable values that satisfies all the imposed constraints, essentially representing a point within the feasible region.

The origins of constraint satisfaction as a general problem can be traced back to the 1970s [50]. The development of George Dantzig's Simplex algorithm for Linear Programming in 1946 paved the way for determining feasible solutions to problems with hundreds of variables that satisfy linear constraints, offering a notable advancement in the field of mathematical optimization. In recent years, geometric algorithms such as the Ellipsoid algorithm [44, Chapter 3] have established a profound reputation for solving many NP-hard combinatorial problems up to a certain optimality.

The approaches employed in constraint satisfaction vary depending on the types of constraints involved. Such problems are typically tackled through heuristic search algorithms, particularly backtracking or local search methods. Additionally, constraint propagation techniques are utilized to simplify the problem-solving process, although most of these methods are generally incomplete, meaning they may not always provide a solution or prove unsatisfiability. Constraint propagation is often combined with search algorithms to enhance problem-solving efficiency.

A sub-class of constraint satisfaction algorithms are projection-based algorithms that utilize the modeling of constraints in a lifted product space formulation and can provide solutions to highly nonconvex discrete and combinatorial optimization problems. This modeling strategy, a general computational approach, was termed the "divide and concur" method by physicist Veit Elser. Professor Veit Elser conducts research in condensed matter physics at Cornell University and has utilized these algorithms for reconstructing particles from noisy diffraction patterns in X-ray crystallography [38, 43].

Over the years, this approach has demonstrated exceptional performance, in solving several NP-hard problems, including Sudoku, disk packing, the kissing number problem, phase retrieval, bit retrieval, protein structure prediction, diffraction signal reconstruction, boolean satisfiability problems [33], finding dominating set of queens in a chess-board, graph-coloring problems [1, 2, 33].

More recently, Elser has applied this approach to various learning problems, such as interpretable efficient data representations for nonnegative matrix factorization [36, 40], deep neural networks [40], autoencoders [40], boolean generative networks [33], and logical embeddings for large language models in natural language processing [34, 35].

In recognition of Elser's two-decade contributions, we term this approach as *Elser's framework*. Our specific focus will be on one aspect of this framework, namely the bilinear feasibility framework for learning problems [39, 40]. The numerical evidence presented in [40] is captivating and raises a series of theoretical questions that are of interest to both the mathematical optimization and deep learning communities.

We will explain this framework with a mathematical model for one toy example in Section 1.5. Before we do so, let us fix some basic concepts.

1.2 Basic notions

Throughout this thesis, we denote the set of natural numbers $\{1, 2, \ldots\}$ by $\mathbb{N}$. The real numbers, nonnegative real numbers, and strictly positive real numbers are indicated by $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{R}_{++}$, respectively. Euclidean and Hilbert spaces are the fundamental spaces we work in.

Definition 1.1. A Hilbert space, X, is a complete inner product space.

We assume that

$$X \text{ is a real Hilbert space} \tag{1.1}$$

with inner product $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ and induced norm denoted by $\|\cdot\|$. For an introductory reference on Hilbert spaces, see [47].

Let $A \colon X \rightrightarrows X$ be an arbitrary set-valued operator, i.e., $Ax \subseteq X$ $\big(\forall x \in X\big)$. The *graph* of A, denoted by $\operatorname{gra} A$, is defined as

$$\operatorname{gra} A = \big\{ \big(x, y\big) \in X \times X \bigm| y \in Ax \big\}.$$

The *domain* of A, denoted by $\operatorname{dom} A$, is defined as

$$\operatorname{dom} A = \big\{ x \in X \bigm| Ax \neq \varnothing \big\}.$$

The *range* of A, denoted by $\operatorname{ran} A$, is defined as

$$\operatorname{ran} A = \big\{ y \in X \bigm| \exists\ x \in X \text{ such that } y \in Ax \big\}.$$

For an introductory reference on set-valued operators, see [3, 8].

Definition 1.2. A subset C of X is *convex* if for all $x, y \in \mathrm{C}$ and $\lambda \in (0, 1)$,

$$\lambda x + (1 - \lambda) y \in \mathrm{C}. \tag{1.2}$$

Definition 1.3. A subset C of X is *a cone* if $\mathrm{C} = \mathbb{R}_{++}\mathrm{C}$. That is, $x \in \mathrm{C}$ and $\lambda > 0$ implies $\lambda x \in \mathrm{C}$.

A convex cone is a set that is both convex and a cone. The positive orthant and the set of semidefinite matrices both serve as examples of convex cones [8, Chapter 6]. *The conical hull* of C, cone C, is the intersection of all the cones in X containing C. If there exist two points $x, y \in C$ such that for some $\lambda \in (0,1)$, the point $\lambda x + (1-\lambda)y \notin C$, i.e., condition (1.2) fails, then C is called a set that is not convex. A simple example of a set that is not convex is the *unit sphere* in $X\backslash\{0\}$ denoted by

$$\mathsf{S} := \big\{x \in X\backslash\{0\} \;\big|\; \|x\| = 1\big\}. \tag{1.3}$$

The well-known conic sections in $\mathbb{R}^2$ and quadric surfaces in $\mathbb{R}^3$ from Euclidean geometry serve as evident examples sets that are not convex. These sets, extending beyond linear spaces, are prime candidates for the development of approximation models and algorithms in scientific computation.

1.3 Decomposition via product space

Decomposing large computational problems into simpler and smaller problems is a conventional idea in the area of scientific computing, which accounts for many physics-inspired algorithms. One of the techniques that stands out for feasibility-based optimization models is the projection-based (splitting) algorithms under decomposition paradigm via product space, see e.g., [56] and [8, Chapters 26, 28 and 30]. We will expand on this basic technique to explain our *mathematical model of Elser's framework*. Let us now define the players involved in the product space formulation.

Definition 1.4 (Product space)**.** Let $N \in \mathbb{N}$ and X_i be real Hilbert spaces, with inner products $\langle\cdot,\cdot\rangle_i$, for each $i \in \{1,\dots,N\}$. Then the product space $H := X_1 \times X_2 \times \cdots \times X_N$ is defined as

$$H = \big\{(x_i) = (x_1, x_2, \dots, x_N) \;\big|\; x_i \in X_i\big\}, \tag{1.4}$$

with inner product $\langle x, y\rangle = \sum_{i=1}^{N} \langle x_i, y_i\rangle_i\,,\ \forall x = (x_i), y = (y_i) \in H$ and norm $\|x\| = \sqrt{\sum_{i=1}^{N} \|x_i\|_i^2},\ \forall x = (x_i) \in H$.

If $X_i = X, \forall i \in \{1,\dots,N\}$, we denote H by X^N. H is also a Hilbert space.

Example 1.5. Consider two parameters $\beta \in \mathbb{R}\backslash\{0\}$ and $\xi \in \mathbb{R}\backslash\{0\}$ and additionally, consider a closed subspace $Z \subseteq X$. For $r \in \mathbb{N}$, define a product space

$$(X \times X)^r \times Z,$$

then for any

$$\big((x_1,y_1),(x_2,y_2),\dots,(x_r,y_r),z\big) \in (X \times X)^r \times Z,$$

and

$$\big((u_1,v_1),(u_2,v_2),\dots,(u_r,v_r),w\big) \in (X \times X)^r \times Z,$$

the inner product on $(X \times X)^r \times Z$ is given by,

$$\Big(\sum_{\ell=1}^{r} \langle x_\ell, u_\ell \rangle + \xi^2 \langle y_\ell, v_\ell \rangle \Big) + \beta^2 \langle z, w \rangle, \tag{1.5}$$

and the induced (ξ, β)-weighted norm on $(X \times X)^r \times Z$, is given by

$$\big\| \big((x_1, y_1), (x_2, y_2), \ldots, (x_r, y_r), z\big) \big\| := \sqrt{\sum_{\ell=1}^{r} (\|x_\ell\|^2 + \xi^2 \|y_\ell\|^2) + \beta^2 \|z\|^2}.$$

Example 1.5 is an instance of the general product space that is utilized to describe the learning models in Section 1.5, Section 1.6, and Section 1.7. Note that x, y, z could be vectors, matrices, or tensors.

Definition 1.6 (Diagonal space)**.** Let $N \in \mathbb{N}$, and X and $H = X^N$ be Hilbert spaces. The diagonal space Δ in H is the subspace

$$\Delta = \big\{ (x, \ldots, x) \in X^N \ \big| \ x \in X \big\} \subseteq H. \tag{1.6}$$

The diagonal space is a closed subspace of X^N.

We now recall the projection onto any subset C of the Hilbert space X. Recall that $A \colon X \rightrightarrows X$ denotes an arbitrary set-valued operator, i.e., $Ax \subseteq X$ $\big(\forall x \in X\big)$.

Definition 1.7. Let $C \subseteq X, x \in X$, and $z \in C$. Then z is a projection of x onto C if

$$(\forall c \in C) \quad \|x - z\| \leqslant \|x - c\|.$$

In other words, z is a closest point to $x \in C$. The associated operator $P_C : X \rightrightarrows C : x \mapsto \big\{ z \in C \ \big| \ z \text{ is a projection of } x \text{ onto } C \big\}$ is called the projection operator.

The projection onto a closed convex set in Hilbert space is unique. We mention this as a fact from [8, Theorem 3.16].

Fact 1.8. Let C be a nonempty *closed convex* subset of X and let $x \in X$. The projection of x onto C is the *unique point* in C denoted by $P_C(x)$ that satisfies

$$\|x - z\| = \inf_{c \in C} \ \|c - x\|, \quad \text{where} \quad z = P_C(x). \tag{1.7}$$

If $P_C(x) = \{z\}$ we have written $z = P_C(x)$ rather than $\{z\} = P_C(x)$, which is a slight abuse of notation.

Definition 1.9 (Two set feasibility formulation for finitely many constraints)**.** Let $I = \{1, \ldots, N\}$ and $(\forall i \in I)$ C_i be closed subsets of X. The multi-set feasibility problem asks to find a point in the intersection of the sets, i.e.,

$$\text{Find} \quad x \in \bigcap_{i \in I} C_i. \tag{1.8}$$

Now set $C := \bigtimes_i C_i \subseteq H := X^N$ and $\Delta = \big\{(x,\ldots,x) \in X^N \bigm| x \in X\big\} \subseteq H$. Then (1.8) is equivalent to the two-set feasibility problem in H:

$$\text{Find} \quad x \in X \quad \text{such that} \quad (x,\ldots,x) \in C \cap \Delta. \tag{1.9}$$

Feasibility algorithms that aim to solve (1.8) via (1.9) require the projections onto C and Δ.

These projections are given in the following result and are clear from Definition 1.7.

Fact 1.10. *Using the notations above, we have*

(i) the projection onto the product set, $C = C_1 \times C_2 \times \cdots \times C_N$ is:

$$P_C(x_1, x_2, \ldots, x_N) = P_{C_1}(x_1) \times P_{C_2}(x_2) \times \cdots \times P_{C_N}(x_N).$$

(ii) the projection onto the diagonal set, $\Delta = \big\{(x, x, \ldots, x) \in X^N \bigm| x \in X\big\}$ is:

$$P_C(x_1, x_2, \ldots, x_N) = (\bar{x}, \bar{x}, \ldots, \bar{x}),$$

where $\bar{x} = (1/N)\sum_{i=1}^{N} x_i$.

Remark 1.11. If $(\forall i \in I)$ C_i are *closed and convex*, then the projection onto the product set in Fact 1.10(i), is given by $(P_{C_1}(x_1), P_{C_2}(x_2), \ldots, P_{C_N}(x_N))$. We refer to a fact from [56, Lemma 1.1].

1.4 Matrix factorization

To later introduce Elser's framework in a mathematically precise manner, we require notations that are based on the standard definitions from Linear Algebra and fundamentals of tensors [46].

For $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{M} \in \mathbb{R}^{m\times n}$, we use $\mathbf{v}_i$ to denote the ith entry of the vector $\mathbf{v}$ and $\mathbf{M}_{i,j}$ to denote the (i,j)th entry of the matrix $\mathbf{M}$, where $i \in [m]$, $j \in [n]$ and $[n] = \{1, 2, \ldots, n\}$. We denote the ith row of $\mathbf{M}$ by $\mathbf{M}_{i,:} \in \mathbb{R}^{1\times n}$ and the jth column of $\mathbf{M}$ by $\mathbf{M}_{:,j} \in \mathbb{R}^{m\times 1}$. By convention, a vector $\mathbf{v} \in \mathbb{R}^m$ will always denote a column vector. A tensor $\boldsymbol{\mathcal{T}}$ is a multi-dimensional array and is called of order-p if $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{d_1\times d_2\times\cdots\times d_p}$ where $\boldsymbol{\mathcal{T}}_{i_1,i_2,\ldots,i_p} \in \mathbb{R}$ for each $i_1 \in [d_1], i_2 \in [d_2], \ldots, i_p \in [d_p]$. By convention, an order-1 tensor is a vector, an order-2 tensor is a matrix, and an order-3 tensor, $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{d_1\times d_2\times d_3}$, is a cube tensor. Fixing all but one of the indices of a tensor gives a vector, which is called a *fiber* of the tensor. There are three ways of doing this for an order-3, leading to the following three kinds of fibers:

$$\begin{aligned}
\text{mode-1 (column) fiber}: &\ \boldsymbol{\mathcal{T}}_{:,i_2,i_3} \in \mathbb{R}^{d_1},\\
\text{mode-2 (row) fiber}: &\ \boldsymbol{\mathcal{T}}_{i_1,:,i_3} \in \mathbb{R}^{d_2},\\
\text{mode-3 (tube) fiber}: &\ \boldsymbol{\mathcal{T}}_{i_1,i_2,:} \in \mathbb{R}^{d_3}.
\end{aligned}$$

A *slice* of an order-3 tensor $\boldsymbol{\mathcal{T}}$ is obtained by taking a slice in one direction along the cube. A slice is obtained by fixing one of the indices of an order-3 and letting the two others *free*. There are three ways of doing this for an order-3, leading to the following three kinds of slices:

$$\begin{aligned}\text{Horizontal slices}: &\ \boldsymbol{\mathcal{T}}_{i_1,:,:} \in \mathbb{R}^{d_2\times d_3},\\ \text{Lateral slices}: &\ \boldsymbol{\mathcal{T}}_{:,i_2,:} \in \mathbb{R}^{d_1\times d_3},\\ \text{Frontal slices}: &\ \boldsymbol{\mathcal{T}}_{:,:,i_3} \in \mathbb{R}^{d_1\times d_2}.\end{aligned}$$

As we saw in the definition of slices, we can fix one index of an order-3 tensor to get a matrix, however that is only a part of the full tensor. We can represent the entire tensor as matrices by flattening it out using the process of *matricization*.

Definition 1.12 (Matricization)**.** Given a tensor $\boldsymbol{\mathcal{T}}$, the process of *matricization* reshapes the tensor by flattening it into a matrix by taking all slices along one direction and stacking them together. For an order-3 tensor, we can have three modes of matricization based on which slices we stack together. Concretely,

$$\boldsymbol{\mathcal{T}}_{(1)} \in \mathbb{R}^{d_1\times d_2 d_3}, \boldsymbol{\mathcal{T}}_{(2)} \in \mathbb{R}^{d_2\times d_1 d_3}, \boldsymbol{\mathcal{T}}_{(3)} \in \mathbb{R}^{d_3\times d_1 d_2}.$$

Definition 1.13 (Vectorization)**.** Given a tensor $\boldsymbol{\mathcal{T}}$, a *vectorization* reshapes the tensor by flattening it into a vector. This is done by first matricization of the tensor along the first mode, and then stacking the columns of the resulting matrix to obtain a vector. Specifically, $\operatorname{vec}(\boldsymbol{\mathcal{T}}) = \operatorname{vec}\left(\boldsymbol{\mathcal{T}}_{(1)}\right)$, where the vectorization of a matrix is defined by:

$$\operatorname{vec}\left(\begin{bmatrix} | & | & | & \cdots & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 & \cdots & \mathbf{a}_n \\ | & | & | & \cdots & | \end{bmatrix}\right) = \begin{bmatrix} | \\ \mathbf{a}_1 \\ | \\ \vdots \\ | \\ \mathbf{a}_n \\ | \end{bmatrix}.$$

Definition 1.14 (Matrix Product)**.** For $\mathbf{W} \in \mathbb{R}^{m\times r}$, $\mathbf{X} \in \mathbb{R}^{r\times n}$, the *matrix product* $\mathbf{WX} \in \mathbb{R}^{m\times n}$ is defined by:

$$(\mathbf{WX})_{i,j} = \sum_{k=1}^{r} \mathbf{W}_{i,k}\mathbf{X}_{k,j} \quad \forall i \in [m],\ \forall j \in [n]. \tag{1.10}$$

We can view each matrix $\mathbf{Y} \in \mathbb{R}^{m\times n}$ as either a stack of row vectors or a stack of column vectors, that is:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_{:,1} & \mathbf{Y}_{:,2} & \cdots & \mathbf{Y}_{:,n} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{1,:} \\ \\ \mathbf{Y}_{2,:} \\ \vdots \\ \mathbf{Y}_{m,:} \end{bmatrix}.$$

Then for the matrix product, we have the inner product view:

$$(\mathbf{W}\mathbf{X})_{i,j} = \left(\begin{bmatrix} \mathbf{W}_{1,:} \\ \\ \mathbf{W}_{2,:} \\ \vdots \\ \mathbf{W}_{m,:} \end{bmatrix} \begin{bmatrix} \mathbf{X}_{:,1} & \mathbf{X}_{:,2} & \cdots & \mathbf{X}_{:,n} \end{bmatrix} \right)_{i,j} = \langle \mathbf{W}_{i,:}^\top, \mathbf{X}_{:,j} \rangle. \qquad (1.11)$$

Definition 1.15 (Matrix factorization or matrix decomposition)**.** Given a matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$, and an integer $r \in \mathbb{N}$, the pair $(\mathbf{W}, \mathbf{X})$ is an *r-factorization*, if $\exists\ \mathbf{W} \in \mathbb{R}^{m \times r}$ and $\exists\ \mathbf{X} \in \mathbb{R}^{r \times n}$ such that $\mathbf{Y} = \mathbf{W}\mathbf{X}$.

Moreover, we say r-factorization $(\mathbf{W}, \mathbf{X})$ is a r-rank factorization if $r = \operatorname{rank}(\mathbf{Y})$.

The following fact, from [59, Page 128, Theorem 3.6.2] on different characterizations of rank, establishes when a r-factorization $(\mathbf{W}, \mathbf{X})$ is a r-rank factorization.

Fact 1.16. *For a nonzero matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$, all the following are equivalent:*

(i) $\operatorname{rank}(\mathbf{Y}) \leqslant r$.

(ii) $\exists\ \mathbf{W} \in \mathbb{R}^{m \times r}$, $\exists\ \mathbf{X} \in \mathbb{R}^{r \times n}$ *such that:* $\mathbf{Y} = \mathbf{W}\mathbf{X}$.

(iii) $\exists\ \mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r \in \mathbb{R}^m$, $\exists\ \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_r \in \mathbb{R}^n$ *such that:*

$$\mathbf{Y} = \sum_{k=1}^{r} \mathbf{w}_k \mathbf{x}_k^\top.$$

Corollary 1.17. *Let $\mathbf{Y} \in \mathbb{R}^{m \times n} \backslash \{\mathbf{0}\}$ and $r \geqslant 1$:*

(i) If $r < \operatorname{rank}(\mathbf{Y})$, then there is no r-factorization.

(ii) If $r \geqslant \operatorname{rank}(\mathbf{Y})$, then there are infinitely many r-factorizations.

Thus, every non-null matrix has an r-rank factorization. A null matrix cannot have a r-rank factorization since there cannot be a matrix with 0 rows or 0 columns. The r-rank factorization of a matrix is not unique, since $\bar{\mathbf{W}} = \mathbf{W}\mathbf{M}$ and $\bar{\mathbf{X}} = \mathbf{M}^{-1}\mathbf{X}$ gives another matrix factorization for any invertible $r \times r$ matrix $\mathbf{M}$.

Problem 1.18. *Finding a r-rank factorization of a matrix $\mathbf{Y} \in \mathbb{R}^{m\times n}$ is defined as finding matrices $\mathbf{W} \in \mathbb{R}^{m\times r}$, and $\mathbf{X} \in \mathbb{R}^{r\times n}$ such that:*

$$\mathbf{Y}_{m\times n} = \mathbf{W}_{m\times r}\mathbf{X}_{r\times n}. \tag{1.12}$$

Remark 1.19. In other words, the goal in Problem 1.18 is to express data vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_n$, *i.e.,* columns of $\mathbf{Y}$ as mixtures of a set of feature vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_r$, *i.e.,* columns of $\mathbf{W}$. If the data vectors lie in $\mathbb{R}^m$, then in terms of the $m \times r$ matrix $\mathbf{W}$ of feature vectors, we seek a representation of the data as $\mathbf{y}_1 = \mathbf{W}\mathbf{x}_1, \mathbf{y}_2 = \mathbf{W}\mathbf{x}_2, \ldots, \mathbf{y}_n = \mathbf{W}\mathbf{x}_n$ where the representation vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ lie in $\mathbb{R}^r$.

In the context of machine learning, we now rewrite the r-rank matrix factorization as a feasibility problem.

Problem 1.20.

$$\begin{aligned} \text{Find} \quad & (\mathbf{W}, \mathbf{X}) \\ \text{subject to} \quad & \mathbf{Y} = \mathbf{W}\mathbf{X}, \end{aligned}$$

where $\mathbf{Y} \in \mathbb{R}^{m\times n}$, $\mathbf{W} \in \mathbb{R}^{m\times r}$ and $\mathbf{X} \in \mathbb{R}^{r\times n}$.

The decomposition (1.12) can be seen using inner product view (1.11) as

$$\underbrace{\begin{pmatrix} \times & \times & \times & \times \\ w_{2,1} & w_{2,2} & w_{2,3} & w_{2,4} \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{pmatrix}}_{\text{Feature matrix } \mathbf{W}} \underbrace{\begin{pmatrix} \times & \times & x_{1,3} & \times \\ \times & \times & x_{2,3} & \times \\ \times & \times & x_{3,3} & \times \\ \times & \times & x_{4,3} & \times \end{pmatrix}}_{\text{Representation matrix } \mathbf{X}} = \underbrace{\begin{pmatrix} \times & \times & \times & \times \\ \times & \times & y_{2,3} & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{pmatrix}}_{\text{Data matrix } \mathbf{Y}}, \tag{1.13}$$

such that for each scalar entry, say $\gamma_{2,3}$, the vectors $\mathbf{w}_2 = (w_{2,1}, w_{2,2}, w_{2,3}, w_{2,4})$, and $\mathbf{x}_3 = (x_{1,3}, x_{2,3}, x_{3,3}, x_{4,3})$ satisfies the following bilinear constraint

$$C_{y_{2,3}} := \left\{(\mathbf{x}_3, \mathbf{w}_2) \in \mathbb{R}^r \times \mathbb{R}^r \;\middle|\; \langle \mathbf{x}_3, \mathbf{w}_2 \rangle = \mathbf{w}_2^T \mathbf{x}_3 = y_{2,3}\right\}. \tag{1.14}$$

All the $C_{y_{i,j}}$ corresponding to each scalar entry $y_{i,j}$ of the matrix $\mathbf{Y}$ can be seen as hyperbolas in the product space $\mathbb{R}^r \times \mathbb{R}^r$ and there are as many hyperbolas as the number of entries of the matrix $\mathbf{Y}$.

1.5 Elser's framework

To reformulate the feasibility Problem 1.20, into another feasibility problem that relies on these hyperbolas, we need to highlight the crust of the *Elser's* *framework*, i.e., *the projections splits over both the data and the architecture and can be computed simultaneously.* Toward that goal, we consider a single (hidden) layer network reformulation, of the Problem 1.18 as described in Figure 1.1, such that each scalar $y_{i,j}$ in (1.13) reflects a bilinear interaction between vectors $\mathbf{x}$ and $\mathbf{w}$, depicted in Figure 1.2.

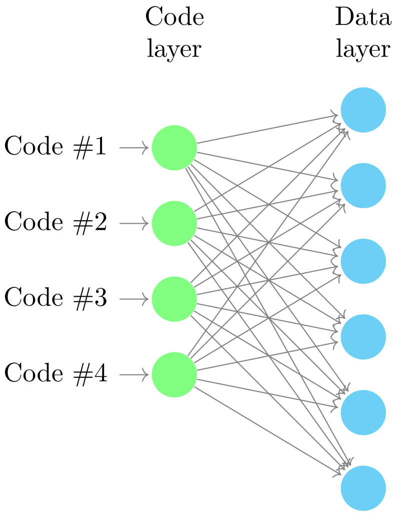

Figure 1.1: A network (instantiation) architecture for matrix-factorization: of rank 4, *i.e.,* 4 code vectors and 6 bilinear neurons [40].

Elser termed this hidden layer as code layer [40, Section 4], where the nodes were called code nodes. Let C denote its index set. We will use index $\ell = 1, \ldots, |C|$ to denote each code node. The diagram depicts how each $\mathbf{y}_k$ in the right layer of nodes is generated from the hidden variables $\mathbf{x}_k$ in the left layer of nodes, i.e., code layer feed-forwards information to a *data layer* of *data nodes*, with an index set denoted by D. We will use index $i = 1, \ldots, |D|$ for each data node. Note that $|D| = m$. Let K represent the collection of indices of columns of $\mathbf{Y}$, so $|K| = n$. We call these data items. Thus this single-layered network is tasked to learn (or find) the weights $w[\ell \to i]$, i.e., column vectors of $\mathbf{W}$ such that for each data item $\mathbf{y}_k$, $\forall k \in K$, there exists a corresponding code vector $\mathbf{x}_k$ that gives a close approximation to $\mathbf{y}_k$, *i.e.,* $\mathbf{y}_k \sim \mathbf{W}\mathbf{x}_k$, see Remark 1.19.

In the context of machine learning, a neuron holds a scalar value.

Definition 1.21 (Bilinear neuron)**.** A neuron characterized by bilinear information geometry is termed a bilinear neuron.

The equation (1.13) suggests that each entry of the matrix $\mathbf{Y}$ corresponds to a bilinear neuron. Mathematically, throughout the thesis, we will consistently refer to these neurons as the sets of bilinear constraints. In the context of matrix factorization problems, a bilinear neuron is represented as Figure 1.2.

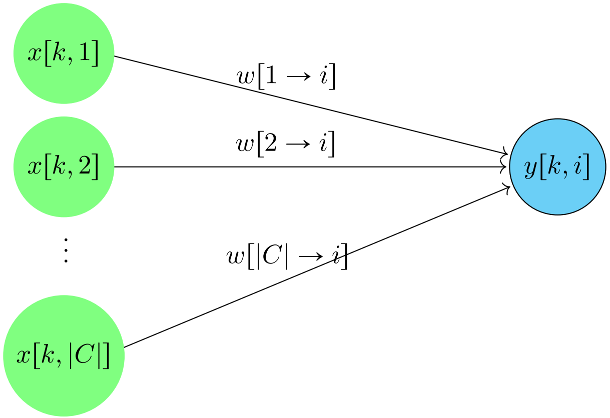

Figure 1.2: for each $k \in K$, we have localized constraint $x[k,:]w[:\to i] = y[k,i]$, also named as bilinear neuron.

Now, while we satisfying the constraints on localized neurons Figure 1.2, we need to adjust the weights w which are not allowed to be free to be different for each data item. At the same time, the x we change to correct y, we also need to make sure that different x's will contribute to all other data nodes. Thus, we split weights on the code nodes and the code vectors on the edges by using the divide and concur approach:

by splitting over the data Figure 1.3: *i.e.,* replacing $w[\ell \to i]$ with $w[k,\ell \to i]$

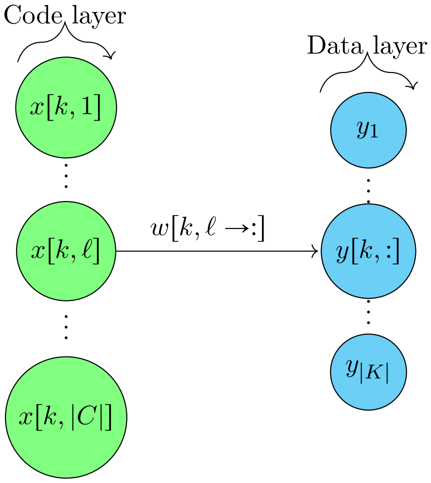

Figure 1.3: Splitting over data

and by splitting over the architecture Figure 1.4: In addition to learning weights, the network must also learn the code that goes with each data item. Thus, for each $k \in K$, the code node variables will hold the representation vectors $\mathbf{x}_k$ on code nodes $\ell \in C$. So, this splitting is done by replacing $x[k, \ell]$ with $x[k, \ell \to i]$.

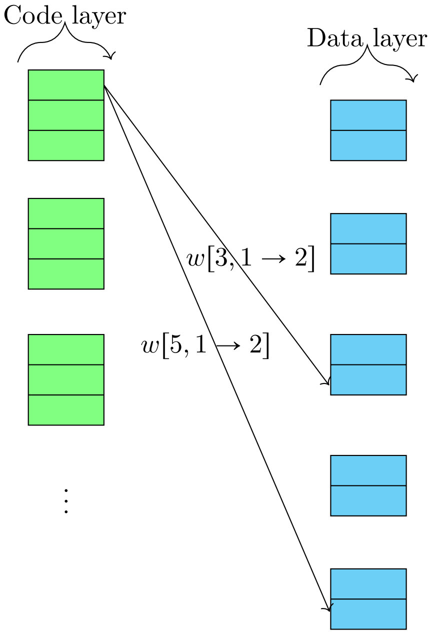

Figure 1.4: Splitting over architecture

Thus, the variables in this constraint formulation are: a weight variable $w[k, \ell \to i]$ and node variable $x[k, \ell \to i]$ for each data item $k \in K$ and edge $\ell \to i$ of the network. Thus, the splitting of data means that we create a product space of dimension $m \times n$ with mn hyperbola (data) constraints, rewritten more explicitly (in comparison to (1.14)) as

$$\forall\ k \in K, i \in D\colon\ \sum_{\ell=1}^{r} x[k, \ell \to i]\ w[k, \ell \to i] = y[k, i], \tag{1.15}$$

where k is a data index (network instantiation), i labels the receiving neuron, and the sum is over neurons r whose outputs $x[k, \ell \to i]$ are incident on i. To formalize (1.15) using tensor notations, we define

$$\boldsymbol{\mathcal{T}}^w_{k,\ell,i} := w[k, \ell \to i], \quad \boldsymbol{\mathcal{T}}^x_{k,\ell,i} := x[k, \ell \to i].$$

Thus, the resulting reshaping from splitting across data and architecture, as depicted in Figure 1.3 and Figure 1.4, has led to a pair of tensors

$$(\boldsymbol{\mathcal{T}}^w, \boldsymbol{\mathcal{T}}^x) \in (\mathbb{R}^{|K| \times |C| \times |D|})^2, \tag{1.16}$$

which[1] can be further used to rewrite the the matrix variable pair $(\mathbf{W}, \mathbf{X})$ of Problem 1.20 as the mode-2 fibers $\boldsymbol{\mathcal{T}}^w_{k,:,i} \in \mathbb{R}^r$ and $\boldsymbol{\mathcal{T}}^x_{k,:,i} \in \mathbb{R}^r$ for each $i \in [m]$ and $k \in [n]$, and the mn-many hyperbolas are given by

$$C_{k,i} = \left\{ (\boldsymbol{\mathcal{T}}^w_{k,:,i}, \boldsymbol{\mathcal{T}}^x_{k,:,i}) \in \mathbb{R}^r \times \mathbb{R}^r \mid \left\langle \boldsymbol{\mathcal{T}}^w_{k,:,i}, \boldsymbol{\mathcal{T}}^x_{k,:,i} \right\rangle = \sum_{\ell=1}^{r} \boldsymbol{\mathcal{T}}^w_{k,\ell,i} \boldsymbol{\mathcal{T}}^x_{k,\ell,i} = y_{k,i} \right\}.$$

Following Elser [40, Equations 14(a,b)], we define the consensus variables as

$$\mathbf{W}_A = \boldsymbol{\mathcal{T}}^w_{k,:,:}, \quad \mathbf{X}_A = \boldsymbol{\mathcal{T}}^x_{:,:,i},$$

where the solution to Problem 1.20 is

$$\mathbf{Y} = \mathbf{W}_A \mathbf{X}_A.$$

Thus, the splitting over both the data and architecture has led to reformulating the Problem 1.20 into a constraint satisfaction problem that relies on hyperbolas and is given as

Problem 1.22.

A constraints

$$\forall\, k \in [n] : \ \left\{ (\boldsymbol{\mathcal{T}}^w_{k,:,:}, \boldsymbol{\mathcal{T}}^w_{k,:,:}, \ldots, \boldsymbol{\mathcal{T}}^w_{k,:,:}) \mid \boldsymbol{\mathcal{T}}^w_{k,:,:} = \mathbf{W}_A \right\}, \tag{1.17a}$$

$$\forall\, i \in [m] : \ \left\{ (\boldsymbol{\mathcal{T}}^x_{:,:,i}, \boldsymbol{\mathcal{T}}^x_{:,:,i}, \ldots, \boldsymbol{\mathcal{T}}^x_{:,:,i}) \mid \boldsymbol{\mathcal{T}}^x_{:,:,i} = \mathbf{X}_A \right\}, \tag{1.17b}$$

B constraints

$$\forall\, k \in [n], i \in [m] : \ \left\langle \boldsymbol{\mathcal{T}}^w_{k,:,i}, \boldsymbol{\mathcal{T}}^x_{k,:,i} \right\rangle = y_{k,i}. \tag{1.17c}$$

In context of machine learning, the training is done by the projection algorithms, a diversion from conventional wisdom of using gradient-based and coordinate-wise algorithms. The Problem 1.22 is a two-set feasibility problem, so alternating projection can be applied where the projections are computed in parallel. The main advantage of this modelling is the independence of both local and global constraints, i.e., by consigning the consensus constraints (1.17a) and (1.17b) to set A which are inactive when the projection P_B is performed such that the neuron-input constraint (1.15) of set B is able to work with local (otherwise unconstrained) variables and vice-versa. One can also interpret the problem formulation Problem 1.22 as implicit matrix factorization: where the data space is approximated by bilinear(e.g. hyperbolas) sets and linear subspaces. The solution to the Problem 1.22 are given by $\mathbf{Y} = \mathbf{W}_A \mathbf{X}_A$.

1.6 Multiset and two-set feasibility frameworks

Let $H := \mathbb{R}^{n \times r \times m}$. We realize H in two different ways (by reshaping) given as

$$H \cong \underbrace{\mathbb{R}^{m \times r} \times \cdots \times \mathbb{R}^{m \times r}}_{n \text{ times}} \cong \underbrace{\mathbb{R}^{r \times n} \times \cdots \times \mathbb{R}^{r \times n}}_{m \text{ times}},$$

[1] Recall $|K| = n, |C| = r, |D| = m$, then $(\boldsymbol{\mathcal{T}}^w, \boldsymbol{\mathcal{T}}^x) \in \mathbb{R}^{n \times r \times m} \times \mathbb{R}^{n \times r \times m}$.

where the inner product on H is defined by taking the product between all entries of the two tensors $(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}) \in H \times H$ and summing them up. Concretely,

$$\begin{aligned}\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle &= \sum_{ijk} \boldsymbol{\mathcal{A}}_{ijk} \boldsymbol{\mathcal{B}}_{ijk} \\ &= \langle \operatorname{vec}(\boldsymbol{\mathcal{A}}), \operatorname{vec}(\boldsymbol{\mathcal{B}}) \rangle.\end{aligned}$$

From this, we can define the Frobenius norm of a tensor as:

$$\begin{aligned}||\boldsymbol{\mathcal{A}}||_F^2 &= \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{A}} \rangle \\ &= ||\operatorname{vec}(\boldsymbol{\mathcal{A}})||_2^2 \\ &= ||\boldsymbol{\mathcal{A}}_{(1)}||_F^2 = ||\boldsymbol{\mathcal{A}}_{(2)}||_F^2 = ||\boldsymbol{\mathcal{A}}_{(3)}||_F^2.\end{aligned}$$

For $\boldsymbol{\mathcal{A}} = (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n)$ and $\boldsymbol{\mathcal{B}} = (\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_n)$ where $\mathbf{A}_i, \mathbf{B}_i \in \mathbb{R}^{m \times r}$, for every $i = 1, 2, \ldots, n$, the inner product on H is defined as $\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle = \sum_{i=1}^n \langle \mathbf{A}_i, \mathbf{B}_i \rangle$ where $\langle \mathbf{A}_i, \mathbf{B}_i \rangle = \operatorname{tr}(\mathbf{A}_i^T \mathbf{B}_i)$ with tr being the trace.

In this section, we replace the role of index k with index j in the context of Section 1.5.

Consider two diagonal sets

$$\Delta_1 := \{\underbrace{(\mathbf{W}_A, \mathbf{W}_A, \ldots, \mathbf{W}_A)}_{n \text{ times}} \in (\mathbb{R}^{m \times r})^n \cong H \mid \mathbf{W}_A \in \mathbb{R}^{m \times r}\}; \tag{1.18}$$

and

$$\Delta_2 := \{\underbrace{(\mathbf{X}_A, \mathbf{X}_A, \ldots, \mathbf{X}_A)}_{m \text{ times}} \in (\mathbb{R}^{r \times n})^m \cong H \mid \mathbf{X}_A \in \mathbb{R}^{r \times n}\}. \tag{1.19}$$

Theorem 1.23 (Projection onto diagonal space Δ_1 and Δ_2)**.** *For any tensor $\boldsymbol{\mathcal{T}} \in H$ when reshaped as*

$$(\boldsymbol{\mathcal{T}}_{1,:,:}, \boldsymbol{\mathcal{T}}_{2,:,:}, \ldots, \boldsymbol{\mathcal{T}}_{n,:,:}) \in H,$$

then the projection onto the set (1.18) *is given by*

$$P_{\Delta_1}(\boldsymbol{\mathcal{T}}) = \underbrace{(\mathbf{W}_{\boldsymbol{\mathcal{T}}}, \ldots, \mathbf{W}_{\boldsymbol{\mathcal{T}}})}_{n \text{ times}} \quad \textit{where} \quad \mathbf{W}_{\boldsymbol{\mathcal{T}}} = (1/n) \sum_{j=1}^n \boldsymbol{\mathcal{T}}_{j,:,:} \in \mathbb{R}^{m \times r}. \tag{1.20}$$

Similarly, for any tensor $\tilde{\boldsymbol{\mathcal{T}}} \in H$ when reshaped as

$$(\tilde{\boldsymbol{\mathcal{T}}}_{:,:,1}, \tilde{\boldsymbol{\mathcal{T}}}_{:,:,2}, \ldots, \tilde{\boldsymbol{\mathcal{T}}}_{:,:,m}) \in H,$$

then the projection onto the set (1.19) *is given by*

$$P_{\Delta_2}(\tilde{\boldsymbol{\mathcal{T}}}) = \underbrace{(\mathbf{X}_{\tilde{\boldsymbol{\mathcal{T}}}}, \ldots, \mathbf{X}_{\tilde{\boldsymbol{\mathcal{T}}}})}_{m \text{ times}} \quad \textit{where} \quad \mathbf{X}_{\tilde{\boldsymbol{\mathcal{T}}}} = (1/m) \sum_{i=1}^m \tilde{\boldsymbol{\mathcal{T}}}_{:,:,i} \in \mathbb{R}^{r \times n}. \tag{1.21}$$

Proof. Clear from Fact 1.10(ii). □

We now summarize and give two algorithmic frameworks, see e.g., [53] for Problem 1.22 for our new modeling description.

Problem 1.24 (Multiset feasibility framework)**.** *In the product Hilbert space $H \times H$, consider the set*

$$S_0 := \Delta_1 \times \Delta_2 \subseteq H^2,$$

which represents the consensus constraint (1.17a) *and* (1.17b) *of Problem 1.22. Let us denote the mn bilinear constraints $S_{i,j}$ that appeared in* (1.17c) *of Problem 1.22 as*

$$(\forall i \in [m], j \in [n]) \quad S_{i,j} := \big\{(\boldsymbol{\mathcal{T}}^w, \boldsymbol{\mathcal{T}}^x) \in H^2 \;\big|\; \big\langle \boldsymbol{\mathcal{T}}^w_{j,:,i}, \boldsymbol{\mathcal{T}}^x_{j,:,i} \big\rangle = y_{i,j} \big\}. \tag{1.22}$$

Thus, the Problem 1.18 can written in multiset-feasibility framework of $mn+1$ constraints such that $(\mathbf{W}_A, \mathbf{X}_A)$ solves Problem 1.18 if and only if

$$\big(\underbrace{(\mathbf{W}_A, \mathbf{W}_A, \ldots, \mathbf{W}_A)}_{n\ times}, \underbrace{(\mathbf{X}_A, \mathbf{X}_A, \ldots, \mathbf{X}_A)}_{m\ times}\big) \in S_0 \cap \bigcap_{i,j} S_{i,j}.$$

Remark 1.25. The projection onto the set S_0 is obtained in Theorem 1.23. The projections onto $S_{i,j}$ are discussed in Chapter 3.

Using Definition 1.9, Problem 1.24 can be converted to a two-set feasibility problem as follows:

Problem 1.26 (Two-set feasibility framework)**.** *We are given the $mn+1$ subsets $S_0, S_{1,1}, \ldots, S_{m,n}$ of $H \times H$ from the multiset feasibility framework of Problem 1.24. We convert this to a two-set feasibility problem residing in $(H \times H)^{mn+1}$ using the sets*

$$S := S_0 \times \bigtimes_{i,j} S_{i,j} \subseteq (H \times H)^{mn+1},$$

and

$$\Delta := \big\{\underbrace{\big((h_1, h_2), \ldots, (h_1, h_2)\big)}_{(mn+1)\ times} \in (H \times H)^{mn+1} \;\big|\; (h_1, h_2) \in H \times H\big\}.$$

Remark 1.27. The projections onto S and Δ can be found using the projection onto the set S_0 from Theorem 1.23, the projections onto $S_{i,j}$, which are discussed in Chapter 3, and Fact 1.10.

1.7 Towards multilayer perceptrons

Having just taken the journey of problem reformulations of Problem 1.18, we have justified *the rise of hyperbolas* which motivates this thesis. The other important sets that are relevant to this thesis are *hyperbolic paraboloids* that show up while reformulating the neural network training problems in deep learning. Now, we give a brief account of the rise of *canonical saddle surfaces or hyperbolic paraboloids* in deep learning using Problem 1.18. To this end we consider deep matrix factorization problem, which asks to decompose a data matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$ approximately as

$$\mathbf{Y} \;\approx\; \mathbf{W}_1 \mathbf{W}_2 \cdots \mathbf{W}_L \mathbf{X}_L, \tag{1.23}$$

where L is the number of layers, $\mathbf{W}_l \in \mathbb{R}^{d_{l-1} \times d_l}$ for $l = 1, \ldots, L$ with $d_0 = m$, and $\mathbf{X}_L \in \mathbb{R}^{d_L \times n}$. The approximation in (1.23) corresponds to *successive factorizations* of $\mathbf{Y}$:

$$\begin{aligned} \mathbf{Y} &\approx \mathbf{W}_1 \mathbf{X}_1, \\ \mathbf{X}_1 &\approx \mathbf{W}_2 \mathbf{X}_2, \\ &\vdots \\ \mathbf{X}_{L-1} &\approx \mathbf{W}_L \mathbf{X}_L, \end{aligned} \tag{1.24}$$

where $\mathbf{X}_l \in \mathbb{R}^{d_l \times n}$ for all l.

In deep learning, the data matrix is input at the input layer, but in the hidden layers $l = 2, \ldots, L-1$, so the bilinear sets $S_{i,j}$ arising in (1.22) will no longer have $y_{i,j}$ as *constants*; rather, they will be considered *variables.* This yields bilinear neurons associated with hidden nodes and class nodes, represented by hyperbolic paraboloids given by

$$C^h_{i,j} = \big\{ (\boldsymbol{\mathcal{W}}_h, \boldsymbol{\mathcal{X}}_h, y^h_{i,j}) \in H \times H \times \mathbb{R} \;\big|\; \langle \boldsymbol{\mathcal{W}}_h, \boldsymbol{\mathcal{X}}_h \rangle = y^h_{i,j} \big\},$$

and

$$C^c_{i,j} = \big\{ (\boldsymbol{\mathcal{W}}_c, \boldsymbol{\mathcal{X}}_c, y^c_{i,j}) \in H \times H \times \mathbb{R} \;\big|\; \langle \boldsymbol{\mathcal{W}}_c, \boldsymbol{\mathcal{X}}_c \rangle = y^c_{i,j} \big\},$$

where h and c index corresponds to hidden codes and class nodes in Figure 1.5. Note that $(\boldsymbol{\mathcal{W}}_h, \boldsymbol{\mathcal{X}}_h)$ and $(\boldsymbol{\mathcal{W}}_c, \boldsymbol{\mathcal{X}}_c)$ are tensor variables associated to hidden layer and code layer respectively.

Note that in Figure 1.5, we visualize the architecture of *binary classification problem* where the data has to be classified into two classes.

1.8 Outline of the thesis

In this chapter, we collect a few results on basic notions and notations. We go through several reformulations of the matrix factorization problem. Along the way, we introduce feasibility frameworks including Elser's mathematical model [40]. This sets the stage for the rise of crosses, hyperbolas, and hyperbolic paraboloids. The remainder of this thesis is organized as follows. In Chapter 2, we discuss the closed-form projection onto the cross. In chapters Chapter 3 and Chapter 4, we discuss all possible cases of projections onto hyperbolas and hyperbolic paraboloids respectively. The projection operators are multi-valued, and they involve spheres. One of the key features revealed during the process

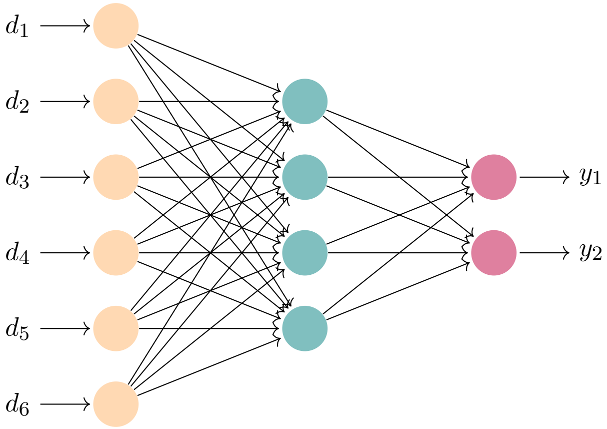

Figure 1.5: Binary Classification for data batch size $|K| = 6$ and 2 class nodes

of finding the projection is the need to find the root expressions of low-degree polynomials up to degree six. More specifically, quadratic, quartic, cubic, and quintic show up in the case of crosses, hyperbolas, and hyperbolic paraboloids respectively. We then dive deeper into the study of cubics in Chapter 5, and we present a novel and convenient presentation of the roots of a cubic polynomial.

In Chapter 6, we extend Prof. Rockafellar's result which states that the sequences generated by the proximal point algorithm enjoy directionally asymptotic properties in a finite-dimensional Hilbert space. In this chapter, we study directionally asymptotical results of strongly convergent subsequences of Fejér monotone sequences in general Hilbert spaces and also provide examples to show that the sets of directionally asymptotic cluster points can be large and that weak convergence is needed in infinite-dimensional spaces.

In Chapter 7, we summarize the contributions of this thesis and highlight some of the open problems that originated from the two areas of research we have proposed in this thesis namely

1. Closed-form projection formulas onto the generalized sets of bilinear forms.
2. A mathematical depiction of Elser's learning framework using tensor notations in the context of the theory of operator splitting algorithms.

Chapter 2

The projection onto the cross

This chapter is based on the article [12] titled "The projection onto the cross" which appeared in Set-Valued and Variational Analysis 30(3) (2023), 997–1009. In simple words, a cross is a set of pairs of orthogonal vectors in Hilbert space. If the underlying space is the real line, one can visualize it as the union of the horizontal and vertical axes in the Euclidean plane. Crosses are nonconvex sets, yet they turn out to be proximinal (i.e., they always admit projections). A complete case study of explicit formulas for the projection onto the cross is developed here.
Let us consider the set C defined by

$$C := \big\{(x,y) \in X \times X \;\big|\; \langle x,y\rangle = 0\big\}, \tag{2.1}$$

where $X \times X$ denotes the product Hilbert space with inner product

$$\big\langle (x,y),(u,v)\big\rangle = \langle x,u\rangle + \langle y,v\rangle,$$

where $(x,y),(u,v) \in X \times X$. This set plays a role in optimization although it does not seem to have a name that is universally used. We follow Kruger, Luke, and Thao's convention and refer to C as *cross* (see [48, Example 2(a) on page 291]). Here is a selection of situations in which crosses are used.

- When $X = \mathbb{R}$, then

  $$C = \big\{(x,y) \in \mathbb{R}^2 \;\big|\; xy = 0\big\} \tag{2.2}$$

  is of interest in the study of set regularity and feasibility problems; see, e.g., Kruger et al.'s [48, Example 2(a) on page 291].

- When $X = \mathbb{R}$, then C is also known as the *switching cone* in the study of mathematical programs by Liang and Ye; see, e.g., [51, equation (20) in Section 4].

- When $X = \mathbb{R}$ and one considers the *(rectangular) hyperbola* defined by

  $$xy = \alpha, \tag{2.3}$$

  where $\alpha > 0$, then C arises as the asymptotic case when $\alpha \to 0^+$.

- The set C is a special case of $C_\gamma := \big\{(x,y) \in X \times X \;\big|\; \langle x,y\rangle = \gamma\big\}$, where X is finite-dimensional and $\gamma \in \mathbb{R}$ is fixed: indeed, $C = C_0$. Sets of the form C_γ are considered in nonnegative matrix factorization and also in deep learning — see [40, Sections 4.1–4.2 and Appendix B]. In that paper, Elser discusses also certain (but not all) cases of projecting onto C_γ and computing projections numerically. He refers to C_γ as a *bilinear constraint* set.

- The study of conic optimization problems may lead to the set C. Recall that for a set $K \subseteq X$, the polar cone of K is $K^\ominus := \big\{x \in X \;\big|\; \sup\langle K, x\rangle \leqslant 0\big\}$ the *dual cone* of K is $K^\oplus = -K^\ominus$, see [8, Definition 6.22]. If K is a nonempty closed cone in X and $K^\oplus$ is its dual cone, then the set
$$C \cap (K \times K^\oplus) \tag{2.4}$$
is called the *conic complementarity set*; see, e.g., Busseti, Moursi, and Boyd's [23, Section 2] and [41].

In [40], the author is particularly interested in the *projection* (nearest point mapping) associated with the cross in the context of algorithms.

Classically, the most famous condition *sufficient for uniqueness of the projection* for a closed set is *convexity.* Unfortunately, the cross C is far from being convex. Now, we highlight some of the properties of the set cross.

2.1 General properties of cross C

Recall that for any set $S \subseteq X$, the intersection of all the convex sets containing S is called the *convex hull* of S, i.e., the smallest convex subset of X containing S, and is denoted by conv S, see [8, Definition 3.3].

Lemma 2.1 (convex hull of C). *We have* $\operatorname{conv} C = X \times X$.

Proof. Clearly, $X \times \{0\} \subseteq C$ and $\{0\} \times X \subseteq C$. Let $(x,y) \in X \times X$. Then $(2x,0) \in C$ and $(0,2y) \in C$; hence, $(x,y) = \frac{1}{2}(2x,0) + \frac{1}{2}(0,2y) \in \operatorname{conv} C$ and we are done. □

Given $x \in X$, its distance from S is given by $d_S(x) := \inf_{y\in S} \|x-y\|$. If there exists $z \in S$ with $d_S(x) = \|x-z\|$, we say that x has a nearest point in S.

Definition 2.2 (Proximinal set [20])**.** A set $S \subseteq X$ is called proximinal (sometimes proximal) if for each $x \in X\backslash S$, the set of best approximations to x from S, i.e.,
$$P_S(x) = \big\{y \in S \;\big|\; \|y-x\| = d_S(x)\big\},$$
is non-empty.

Because C defined in (2.1) is not convex, the question arises whether C at least admits projections everywhere, i.e., whether C is *proximinal*, see Definition 2.2.

An often employed condition *sufficient* for the existence of projections is weak closedness of the set (see, e.g., [8, Proposition 3.14]). When X is finite-dimensional, weak closedness *characterizes* proximinality and is of course also equivalent to ordinary closedness (see, e.g., [8, Corollary 3.15]).

Before we discuss the case when X is infinite-dimensional, recall that *the closure* of a set S is denoted by $\overline{\mathrm{S}}$. The smallest linear subspace of X containing S is span S. A sequence $(x_n)_{n\in\mathbb{N}}$ in X *converges strongly* to a point x, written $x_n \to x$, if

$$\lim_{n\to\infty} \|x_n - x\| = 0.$$

A sequence $(x_n)_{n\in\mathbb{N}}$ in X *converges weakly* to a point x, written $x_n \rightharpoonup x$, if for every $u \in X$,

$$\lim_{n\to\infty} \langle x_n, u\rangle = \langle x, u\rangle .$$

Moreover, a subset S of X is *weakly closed* if every weak limit of every weakly convergent sequence in S is also in S. Likewise, a subset S of X is *weakly sequentially closed* if every weak limit of every weakly convergent sequence in S is also in S. See e.g., [8, Page 35].

When X is infinite-dimensional, then C is far from being weakly closed as the next result illustrates:

Lemma 2.3 (**weak (sequential) closure of C**)**.** *The set C defined in* (2.1) *is closed. If X is finite-dimensional, then C is proximinal. If X is infinite-dimensional, then C is not weakly (sequentially) closed; in fact, the weak (sequential) closure of C is equal to $X \times X$.*

Proof. The continuity of the inner product immediately yields the closedness of C. So we assume that X is infinite-dimensional and we let $(u_i)_{i\in I}$ be an orthonormal family in X such that $\overline{\operatorname{span}}\{u_i\}_{i\in I} = X$. Next, let J be a countably infinite subset of I, say $J = \{i_n\}_{n\in\mathbb{N}}$ for some sequence $(i_n)_{n\in\mathbb{N}}$ in I with pairwise distinct terms. Set $(\forall n \in \mathbb{N})$ $e_n := u_{i_n}$. Then $e_n \rightharpoonup 0$ by [8, Example 2.32]. Set $S_J := \operatorname{span}\{u_j\}_{j\in J} = \operatorname{span}\{e_n\}_{n\in\mathbb{N}}$, and let $(x, y) \in S_J \times S_J$. Now set $\zeta := \langle x, y\rangle$, and $(\forall n \in \mathbb{N})$ $(x_n, y_n) := (x + \zeta e_n, y - e_n)$. Then there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$, we have $e_n \in \{x, y\}^{\perp}$ and thus

$$\langle x_n, y_n\rangle = \langle x + \zeta e_n, y - e_n\rangle = \langle x, y\rangle - \zeta \langle e_n, e_n\rangle = 0. \tag{2.5}$$

Consequently, the sequence $(x_n, y_n)_{n\geqslant N} \in C$. On the other hand,

$$(x_n, y_n) \rightharpoonup (x, y).$$

Altogether, $S_J \times S_J$ lies in the weak sequential closure of C. Because $S_J \times S_J$ is convex, it follows from [8, Theorem 3.34] that its weak sequential closure $\overline{S_J} \times \overline{S_J}$ is also contained in the weak sequential closure of C. This is true for every countably infinite subset J of I. Finally, let $(x, y) \in X \times X$. Then there exists a countably infinite subset J of I such that $(x, y) \in \overline{S_J} \times \overline{S_J}$. Therefore, the weak (sequential) closure of C is equal to $X \times X$. In particular, C is neither weakly sequentially closed nor weakly closed. □

After these explanations, we are now ready to state the main contribution of this chapter: *We will show that C is in fact proximinal and we will also provide an explicit formula for the projection onto C.*

The remainder of the chapter is organized as follows. In Section 2.2, we collect a few results that are known but will help in subsequent sections. Various auxiliary results are obtained in Section 2.3 to make the proof of the main result painless. The main result (Theorem 2.22) is proved in Section 2.4. The notation we employ is standard and follows largely [8].

2.2 Known facts

In this section, we record some definitions and results which will make the proofs given in subsequent sections more clear. We start with the definition of Fréchet differentiability on a Hilbert space X, see e.g. [18, Appendix A.5] and local minimizer of a function [8, Defniniton 11.3].

Definition 2.4. Let U be an open set in X. A function $U \to \mathbb{R}$ is said to be Fréchet differentiable at $x \in U$ if there exists a vector $v \in X$ satisfying

$$\lim_{z \to 0} \frac{f(x+z) - f(x) - \langle v, z\rangle}{\|z\|} = 0.$$

We say f is Fréchet differentiable on U if f is Fréchet differentiable at each $x \in U$.

Definition 2.5. If there exists $\rho \in \mathbb{R}_{++}$ such that x is a minimizer of $f : X \to \mathbb{R}$ over a ball $\mathbb{B}(x;\rho)$, then x is a local minimizer of $f : X \to \mathbb{R}$ defined on a Hilbert space X.

We recall the following fact from [18, Proposition 4.1.1] and [52, Theorem 9.3.1 on page 243].

Fact 2.6. *Let $f\colon X \to \mathbb{R}$ and $h\colon X \to \mathbb{R}$ be continuously Fréchet differentiable. Consider the problem*

$$\textit{minimize}\ \ f(x)\ \ \textit{subject to}\ \ h(x) = 0. \tag{2.6}$$

If $x^* \in X$ is a local minimizer of (2.6) and $\nabla h(x^*) \neq 0$, then there exists a unique $\lambda^* \in \mathbb{R}$ such that

$$\nabla f(x^*) + \lambda \nabla h(x^*) = 0. \tag{2.7}$$

Proof. If X is finite-dimensional, then this follows from [18, Proposition 4.1.1]. If X is infinite-dimensional, then use [52, Theorem 9.3.1 on page 243]. ☐

Lemma 2.7. *If $\lambda \in \mathbb{R} \smallsetminus \{-1, 1\}$, then*

$$\begin{pmatrix} \mathrm{Id} & \lambda\,\mathrm{Id} \\ \lambda\,\mathrm{Id} & \mathrm{Id} \end{pmatrix}^{-1} = \frac{1}{1-\lambda^2}\begin{pmatrix} \mathrm{Id} & -\lambda\,\mathrm{Id} \\ -\lambda\,\mathrm{Id} & \mathrm{Id} \end{pmatrix}, \tag{2.8}$$

where the block matrices are interpreted as linear operators on $X \times X$.

Proof. The result follows by a direct verification. □

The next result provides a parametrization of the unit sphere in $\mathbb{R}^n$. (See also [19].)

Fact 2.8 (parametrization of the sphere). *Suppose that $X = \mathbb{R}^n$, where $n \geqslant 2$, and let $\rho > 0$. Then every point*

$$x = (x_1, x_2, \ldots, x_n) \in \rho\mathbb{S} \subseteq \mathbb{R}^n \tag{2.9}$$

is uniquely described by its spherical coordinates $\theta_1, \ldots, \theta_{n-2}$ *in* $[0, \pi]$ *and* $\theta_{n-1} \in [0, 2\pi[$ *via*[2]

$$(\forall i \in \{1, 2, \ldots, n\}) \quad x_i = \begin{cases} \rho\cos(\theta_i)\prod_{j=1}^{i-1}\sin(\theta_j), & \text{if } i \leqslant n-2; \\ \rho\cos(\theta_{n-1})\prod_{j=1}^{n-2}\sin(\theta_j), & \text{if } i = n-1; \\ \rho\sin(\theta_{n-1})\prod_{j=1}^{n-2}\sin(\theta_j), & \text{if } i = n. \end{cases} \tag{2.10}$$

2.3 Auxillary results

This section lays out the preparatory work for our main result. We consider a point $(x_0, y_0) \in X \times X$. Recall our aim, which is to compute $P_C(x_0, y_0)$. The following result will be useful later. Recall that the *orthogonal complement* of a subset $S \subseteq X$ is denoted by $S^\perp$, i.e., $S^\perp = \big\{x \in H \;\big|\; (\forall s \in S)\ \langle s, x\rangle = 0\big\}$.

Lemma 2.9. *We have*

$$C = \bigcup_{U \text{ is a closed linear subspace of } X} U \times U^\perp; \tag{2.11}$$

moreover,

$$C = \bigcup_{U \text{ is a linear subspace of } X \text{ with } \dim U \leqslant 1} U \times U^\perp. \tag{2.12}$$

Consequently, if $(x_0, y_0) \in X \times X$, then

$$P_C(x_0, y_0) \subseteq \bigcup_{U \text{ is a closed linear subspace of } X} \big\{(P_U x_0, P_{U^\perp} y_0)\big\}. \tag{2.13}$$

Proof. Let $(x, y) \in X \times X$. If $(x, y) \in C$, i.e., $x \perp y$, and we set $U = \mathbb{R}x$, then $(x, y) \in U \times U^\perp$ and U is a (closed) linear subspace of X with $\dim U \leqslant 1$. Conversely, if U is a closed linear subspace of X and $(x, y) \in U \times U^\perp$, then $x \perp y$ and so $(x, y) \in C$. Altogether, we have verified (2.11) and (2.12). The "Consequently" part now follows from (2.11). □

The following example illustrates that the inclusion (2.13) may be strict.

Example 2.10. Suppose that $X = \mathbb{R}$, and set $(x_0, y_0) := (1, 0)$ and $U := \{0\}$. Then $(x_0, y_0) \in C$ and thus $P_C(x_0, y_0) = \{(x_0, y_0)\} = \{(1, 0)\}$. However, $U^\perp = \mathbb{R}$ and $(P_U x_0, P_{U^\perp} y_0) = (0, 0) \notin P_C(x_0, y_0)$.

[2] Recall the empty product convention which sets such products equal to 1.

We also note that when $(x_0, y_0) \in C$, then[3] $P_C(x_0, y_0) = (x_0, y_0)$. Thus, for the remainder of this section, we focus on the case when $(x_0, y_0) \notin C$, i.e.,

$$\langle x_0, y_0 \rangle \neq 0; \text{ consequently, } x_0 \neq 0 \text{ and } y_0 \neq 0. \tag{2.14}$$

To determine $P_C(x_0, y_0)$, we introduce the objective and constraint functions

$$f(x, y) := \tfrac{1}{2}\|x - x_0\|^2 + \tfrac{1}{2}\|y - y_0\|^2 \quad \text{and} \quad h(x, y) := \langle x, y \rangle, \tag{2.15}$$

which are obviously continuously Fréchet differentiable on $X \times X$. Indeed, for every $(x, y) \in X \times X$, we have

$$\nabla f(x, y) = (x - x_0, y - y_0) \quad \text{and} \quad \nabla h(x, y) = (y, x). \tag{2.16}$$

Moreover, the points in $P_C(x_0, y_0)$ are precisely the solutions to the following optimization problem:

$$\text{minimize} \quad f(x, y) \quad \text{subject to} \quad h(x, y) = 0. \tag{2.17}$$

Indeed, $C = \big\{(x, y) \in X \times X \bigm| h(x, y) = 0\big\}$ while the optimal value of (2.17) is $\frac{1}{2}d_C^2(x_0, y_0) = \inf_{(x,y)\in C} \frac{1}{2}\|(x, y) - (x_0, y_0)\|^2$.

Proposition 2.11. $(0, 0)$ *is never a solution to* (2.17)*. Consequently,* $P_C^{-1}(0, 0) = \{(0, 0)\}$*, where* C *is defined in* (2.1)*.*

Proof. Suppose to the contrary that $(0, 0)$ solves (2.17). Then the optimal value of (2.17) is $\frac{1}{2}\|x_0\|^2 + \frac{1}{2}\|y_0\|^2$ and both terms in this sum are positive (by (2.14)). Note that $(x_0, 0) \in C$. But $f(x_0, 0) = \frac{1}{2}\|x_0\|^2 < \frac{1}{2}\|x_0\|^2 + \frac{1}{2}\|y_0\|^2 = f(0, 0)$ because $y_0 \neq 0$. But this contradicts the minimality of $f(0, 0)$. Hence $(0, 0)$ cannot be optimal. The "Consequently" part follows. □

Corollary 2.12. *Suppose that* $(x, y) \in X \times X$ *solves* (2.17)*. Then* $\nabla h(x, y) \neq (0, 0)$*.*

Proof. By hypothesis, $(x, y) \in P_C(x_0, y_0)$. Suppose to the contrary that

$$\nabla h(x, y) = (0, 0).$$

Then, by (2.16), $(x, y) = (0, 0)$. Now Proposition 2.11 shows that (x, y) cannot be a solution to (2.17) which is absurd. □

Theorem 2.13. *Suppose that* $(x, y) \in X \times X$ *solves* (2.17)*. Then there exists a unique* $\lambda \in \mathbb{R}$ *such that*

$$x + \lambda y = x_0 \quad \text{and} \quad y + \lambda x = y_0. \tag{2.18}$$

[3] We should technically write $P_C(x_0, y_0) = \{(x_0, y_0)\}$; however, for convenience and readability, we will identify singleton sets with the vectors they contain.

Proof. By Fact 2.6, there exists a unique $\lambda \in \mathbb{R}$ such that $\nabla f(x,y) + \lambda \nabla h(x,y) = (0,0)$. Using (2.16), this turns into $(x - x_0, y - y_0) + \lambda(y,x) = (0,0)$, i.e., (2.18). □

Proposition 2.14. *Suppose that $(x,y) \in X \times X$ and $\lambda \in \mathbb{R}$ satisfy*

$$x + \lambda y = x_0 \quad \textit{and} \quad y + \lambda x = y_0. \tag{2.19}$$

Then the following hold:

(i) If $x_0 \neq y_0$, then $\lambda \neq 1$.

(ii) If $x_0 \neq -y_0$, then $\lambda \neq -1$.

Proof. (i): We prove the contrapositive and thus assume that $\lambda = 1$. Then, by (2.19), $x_0 = x + (1)y = x + y = y + x = y + (1)x = y_0$. (ii): Argue similarly to the proof of (i). □

Proposition 2.15. *Suppose that $(x,y) \in X \times X$ and $\lambda \in \mathbb{R} \smallsetminus \{-1,1\}$ satisfy*

$$x + \lambda y = x_0 \quad \textit{and} \quad y + \lambda x = y_0. \tag{2.20}$$

Then

$$x = \frac{1}{1-\lambda^2}(x_0 - \lambda y_0) \quad \textit{and} \quad y = \frac{1}{1-\lambda^2}(y_0 - \lambda x_0). \tag{2.21}$$

Proof. Combine (2.20) with Lemma 2.7. □

Corollary 2.16. *Suppose that $x_0 \neq \pm y_0$ and $(x,y) \in X \times X$ solves (2.17). Then there exists a unique $\lambda \in \mathbb{R} \smallsetminus \{-1,1\}$ such that $x + \lambda y = x_0$, $y + \lambda x = y_0$, and*

$$x = \frac{1}{1-\lambda^2}(x_0 - \lambda y_0) \quad \textit{and} \quad y = \frac{1}{1-\lambda^2}(y_0 - \lambda x_0). \tag{2.22}$$

Proof. The existence and uniqueness of $\lambda \in \mathbb{R}$ such that $x + \lambda y = x_0$ and $y + \lambda x = y_0$ follows from Theorem 2.13. Next, Proposition 2.14 implies that $\lambda \neq \pm 1$. Finally, apply Proposition 2.15. □

Proposition 2.17. *Suppose $\lambda \in \mathbb{R} \smallsetminus \{\pm 1\}$, and set*

$$x := \frac{1}{1-\lambda^2}(x_0 - \lambda y_0) \quad \textit{and} \quad y := \frac{1}{1-\lambda^2}(y_0 - \lambda x_0). \tag{2.23}$$

Then the following hold for (x,y) defined in (2.23):

(i) $x + \lambda y = x_0$ and $y + \lambda x = y_0$.

(ii) The objective function f defined in (2.15) evaluated at (x,y) is

$$f(x,y) = \frac{\lambda^2}{2(1-\lambda^2)^2}\big((1+\lambda^2)(\|x_0\|^2 + \|y_0\|^2) - 4\lambda\langle x_0, y_0\rangle\big). \tag{2.24}$$

(iii) $\langle x,y\rangle = 0$ if and only if $(1+\lambda^2)\langle x_0, y_0\rangle = \lambda(\|x_0\|^2 + \|y_0\|^2)$.

(iv) If $\langle x,y\rangle = 0$, then

$$f(x,y) = \tfrac{1}{2}\lambda\langle x_0, y_0\rangle. \tag{2.25}$$

Proof. (i): This is an easy algebraic verification. (ii): For convenience, set

$$\tau := \|x_0 - \lambda y_0\|^2 + \|y_0 - \lambda x_0\|^2 = (1-\lambda^2)^2\big(\|x\|^2 + \|y\|^2\big). \tag{2.26}$$

Then

$$\begin{align}
\tau &= \|x_0 - \lambda y_0\|^2 + \|y_0 - \lambda x_0\|^2 \tag{2.27a}\\
&= \|x_0\|^2 - 2\lambda\langle x_0, y_0\rangle + \lambda^2\|y_0\|^2 + \|y_0\|^2 - 2\lambda\langle y_0, x_0\rangle + \lambda^2\|x_0\|^2 \tag{2.27b}\\
&= (1+\lambda^2)(\|x_0\|^2 + \|y_0\|^2) - 4\lambda\langle x_0, y_0\rangle. \tag{2.27c}
\end{align}$$

Using (2.15), (2.23), (2.26), and (2.27), we obtain

$$\begin{align}
f(x,y) &= \tfrac{1}{2}\|x - x_0\|^2 + \tfrac{1}{2}\|y - y_0\|^2 \tag{2.28a}\\
&= \frac{1}{2}\left\|\frac{x_0 - \lambda y_0}{1-\lambda^2} - x_0\right\|^2 + \frac{1}{2}\left\|\frac{y_0 - \lambda x_0}{1-\lambda^2} - y_0\right\|^2 \tag{2.28b}\\
&= \frac{1}{2(1-\lambda^2)^2}\big(\|x_0 - \lambda y_0 - (1-\lambda^2)x_0\|^2 + \|y_0 - \lambda x_0 - (1-\lambda^2)y_0\|^2\big) \tag{2.28c}\\
&= \frac{\lambda^2}{2(1-\lambda^2)^2}\big(\|\lambda x_0 - y_0\|^2 + \|\lambda y_0 - x_0\|^2\big) \tag{2.28d}\\
&= \frac{\lambda^2}{2(1-\lambda^2)^2}\big((1+\lambda^2)(\|x_0\|^2 + \|y_0\|^2) - 4\lambda\langle x_0, y_0\rangle\big). \tag{2.28e}
\end{align}$$

(iii): Because $1-\lambda^2 \neq 0$ and using (2.23), we have the following equivalences:

$$\begin{align}
\langle x, y\rangle = 0 &\Leftrightarrow \langle x_0 - \lambda y_0, y_0 - \lambda x_0\rangle = 0 \tag{2.29a}\\
&\Leftrightarrow \langle x_0, y_0\rangle - \lambda\|x_0\|^2 - \lambda\|y_0\|^2 + \lambda^2\langle x_0, y_0\rangle = 0 \tag{2.29b}\\
&\Leftrightarrow (1+\lambda^2)\langle x_0, y_0\rangle = \lambda(\|x_0\|^2 + \|y_0\|^2). \tag{2.29c}
\end{align}$$

(iv): Suppose that $\langle x, y\rangle = 0$. By (iii),

$$\lambda\langle x_0, y_0\rangle = \frac{\lambda^2}{1+\lambda^2}\big(\|x_0\|^2 + \|y_0\|^2\big). \tag{2.30}$$

It thus follows from (2.30) that

$$\begin{align}
&\big(1+\lambda^2\big)\big(\|x_0\|^2 + \|y_0\|^2\big) - 4\lambda\langle x_0, y_0\rangle \tag{2.31a}\\
&= \big(1+\lambda^2\big)\big(\|x_0\|^2 + \|y_0\|^2\big) - 4\frac{\lambda^2}{1+\lambda^2}\big(\|x_0\|^2 + \|y_0\|^2\big) \tag{2.31b}\\
&= \frac{\|x_0\|^2 + \|y_0\|^2}{1+\lambda^2}\big((1+\lambda^2)^2 - 4\lambda^2\big) \tag{2.31c}\\
&= \frac{\|x_0\|^2 + \|y_0\|^2}{1+\lambda^2}\big(1 + \lambda^4 + 2\lambda^2 - 4\lambda^2\big) \tag{2.31d}
\end{align}$$

$$= \frac{\|x_0\|^2 + \|y_0\|^2}{1+\lambda^2}\big(1+\lambda^4 - 2\lambda^2\big) \tag{2.31e}$$

$$= \frac{\|x_0\|^2 + \|y_0\|^2}{1+\lambda^2}\big(1-\lambda^2\big)^2. \tag{2.31f}$$

Using (ii), (2.31), and (2.30), we conclude that

$$f(x,y) = \frac{\lambda^2}{2(1-\lambda^2)^2}\frac{\|x_0\|^2 + \|y_0\|^2}{1+\lambda^2}\big(1-\lambda^2\big)^2 \tag{2.32a}$$

$$= \frac{\lambda^2}{1+\lambda^2}\frac{\|x_0\|^2 + \|y_0\|^2}{2} \tag{2.32b}$$

$$= \tfrac{1}{2}\lambda\left\langle x_0, y_0\right\rangle. \tag{2.32c}$$

The proof is complete. □

Proposition 2.18. *Consider the equation*

$$\left\langle x_0, y_0\right\rangle\lambda^2 - (\|x_0\|^2 + \|y_0\|^2)\lambda + \left\langle x_0, y_0\right\rangle = 0, \tag{2.33}$$

which is a quadratic polynomial with respect to the variable λ. Then (2.33) *has (possibly distinct) real roots*

$$\lambda_\pm := \frac{\|x_0\|^2 + \|y_0\|^2 \pm \sqrt{(\|x_0\|^2 + \|y_0\|^2)^2 - 4\left\langle x_0, y_0\right\rangle^2}}{2\left\langle x_0, y_0\right\rangle} \tag{2.34a}$$

$$= \frac{\|x_0\|^2 + \|y_0\|^2 \pm \|x_0+y_0\|\|x_0-y_0\|}{2\left\langle x_0, y_0\right\rangle} \tag{2.34b}$$

which satisfy $\lambda_+\lambda_- = 1$. Moreover, $\lambda_+ \neq \lambda_- \Leftrightarrow x_0 \neq \pm y_0$,

$$\lambda_+\left\langle x_0, y_0\right\rangle \geqslant \lambda_-\left\langle x_0, y_0\right\rangle > 0, \tag{2.35}$$

and the left inequality is strict if and only if $x_0 \neq \pm y_0$.

Proof. First note that $\|x_0 \mp y_0\|^2 \geqslant 0 \Leftrightarrow \|x_0\|^2 + \|y_0\|^2 \geqslant \pm 2\left\langle x_0, y_0\right\rangle \Leftrightarrow \|x_0\|^2 + \|y_0\|^2 \geqslant 2|\left\langle x_0, y_0\right\rangle|$ which makes the discriminant of (2.33) nonnegative, and equal to $0 \Leftrightarrow x_0 = \pm y_0$. Hence, the roots of (2.33) are real, and distinct if and only if $x_0 \neq \pm y_0$. Second, Vieta's formulas give $\lambda_+\lambda_- = \left\langle x_0, y_0\right\rangle / \left\langle x_0, y_0\right\rangle = 1$, so both λ_+ and λ_- are nonzero and they have the same sign. Next, the quadratic formula yields

$$\lambda_\pm = \frac{\|x_0\|^2 + \|y_0\|^2 \pm \sqrt{(\|x_0\|^2 + \|y_0\|^2)^2 - 4\left\langle x_0, y_0\right\rangle^2}}{2\left\langle x_0, y_0\right\rangle} \tag{2.36a}$$

$$= \frac{\|x_0\|^2 + \|y_0\|^2 \pm \sqrt{\big(\|x_0\|^2 + \|y_0\|^2 + 2\left\langle x_0, y_0\right\rangle\big)\big(\|x_0\|^2 + \|y_0\|^2 - 2\left\langle x_0, y_0\right\rangle\big)}}{2\left\langle x_0, y_0\right\rangle} \tag{2.36b}$$

$$= \frac{\|x_0\|^2 + \|y_0\|^2 \pm \|x_0+y_0\|\|x_0-y_0\|}{2\left\langle x_0, y_0\right\rangle}, \tag{2.36c}$$

which yields (2.34). Clearly, $\lambda_+ \langle x_0, y_0\rangle > 0$ and $\lambda_+ \langle x_0, y_0\rangle \geqslant \lambda_- \langle x_0, y_0\rangle$. Finally, as observed above, λ_+ and λ_- have the same sign; hence, $\lambda_- \langle x_0, y_0\rangle$ is positive as well. □

Proposition 2.19. *Suppose that* $(x_0, y_0) \in X \times X$ *satisfies* $\langle x_0, y_0\rangle = 0$*. Then*

$$P_C(x_0, y_0) = (x_0, y_0). \tag{2.37}$$

Proof. This is clear because $(x_0, y_0) \in C$ by definition of C (see (2.1)). □

Proposition 2.20. *Suppose that* $(x_0, y_0) \in X \times X$ *satisfies* $\langle x_0, y_0\rangle \neq 0$ *and* $x_0 \neq \pm y_0$*. Then*

$$\lambda := \frac{\|x_0\|^2 + \|y_0\|^2 - \|x_0 + y_0\|\|x_0 - y_0\|}{2\langle x_0, y_0\rangle} \neq \pm 1, \tag{2.38}$$

$$P_C(x_0, y_0) = \frac{1}{1-\lambda^2}\big(x_0 - \lambda y_0, y_0 - \lambda x_0\big) \tag{2.39}$$

is a singleton*, and*

$$\tfrac{1}{2}d_C^2(x_0, y_0) = \tfrac{1}{2}\lambda\langle x_0, y_0\rangle = \frac{\|x_0\|^2 + \|y_0\|^2 - \|x_0 + y_0\|\|x_0 - y_0\|}{4}. \tag{2.40}$$

Proof. Assume that

$$(x, y) \in P_C(x_0, y_0). \tag{2.41}$$

(We will show that $P_C(x_0, y_0) \neq \varnothing$ later in this proof.) Then (x, y) solves (2.17). From Corollary 2.16, there exists a unique $\lambda \in \mathbb{R} \smallsetminus \{\pm 1\}$ such that $x + \lambda y = x_0$, $y + \lambda x = y_0$, and

$$x = \frac{1}{1-\lambda^2}(x_0 - \lambda y_0) \quad \text{and} \quad y = \frac{1}{1-\lambda^2}(y_0 - \lambda x_0). \tag{2.42}$$

Because $(x, y) \in C$, Proposition 2.17(iii) shows that of the quadratic equation

$$\lambda \text{ is a solution of } (1 + \mu^2)\langle x_0, y_0\rangle = \mu\big(\|x_0\|^2 + \|y_0\|^2\big), \tag{2.43}$$

which is a quadratic equation in the real variable μ. Set

$$\lambda_\pm := \frac{\|x_0\|^2 + \|y_0\|^2 \pm \|x_0 + y_0\|\|x_0 - y_0\|}{2\langle x_0, y_0\rangle}, \tag{2.44}$$

which are the two (possibly distinct) roots of the quadratic equation in (2.43). By (2.43) and Proposition 2.18, we deduce that $\lambda \in \{\lambda_-, \lambda_+\}$ and that $\lambda_+\lambda_- = 1$. Because $x_0 \neq \pm y_0$, Proposition 2.18 also yields $\lambda_+ \neq \lambda_-$ and

$$\lambda_+\langle x_0, y_0\rangle > \lambda_-\langle x_0, y_0\rangle. \tag{2.45}$$

Hence neither λ_+ nor λ_- is equal to ± 1. Now we (well) define

$$x_\pm = \frac{1}{1-\lambda_\pm^2}(x_0 - \lambda_\pm y_0) \quad \text{and} \quad y_\pm = \frac{1}{1-\lambda_\pm^2}(y_0 - \lambda_\pm x_0). \tag{2.46}$$

In view of (2.42), $x \in \{x_-, x_+\}$. Because $\lambda_\pm$ solve the quadratic equation in (2.43), we deduce from Proposition 2.17(iii) that $\langle x_+, y_+\rangle = \langle x_-, y_-\rangle = 0$, i.e., (x_+, y_+) and (x_-, y_-) both belong to C. Recalling the definition of f from (2.15), we note that Proposition 2.17(iv) and (2.45) result in $f(x_+, y_+) > f(x_-, y_-)$. On the other hand, we know that (x, y) is either (x_+, y_+) or (x_-, y_-). Altogether, because we've assumed at the beginning of the proof in (2.41) that (x, y) is a minimizer of f, we conclude that $(x, y) = (x_-, y_-)$. This yields (2.39). Moreover, (2.40) follows because

$$f(x,y) = f(x_-, y_-) = \frac{\lambda_- \langle x_0, y_0\rangle}{2} = \frac{\|x_0\|^2 + \|y_0\|^2 - \|x_0 + y_0\|\|x_0 - y_0\|}{4} \tag{2.47}$$

by (2.44).

It remains now to show that $P_C(x_0, y_0) \neq \varnothing$. Let U be a closed linear subspace of X. In view of (2.13), (2.47), and (2.15) it suffices to show that $f(x,y) \overset{?}{\leqslant} f(P_U x_0, P_{U^\perp} y_0)$, i.e.,

$$\frac{\|x_0\|^2 + \|y_0\|^2 - \|x_0 + y_0\|\|x_0 - y_0\|}{4} \overset{?}{\leqslant} \frac{1}{2}\|P_U x_0 - x_0\|^2 + \frac{1}{2}\|P_{U^\perp} y_0 - y_0\|^2. \tag{2.48}$$

To this end, recall that the reflector $R_U := P_U - P_{U^\perp}$ is a linear isometry. Using Cauchy-Schwarz, we thus estimate

$$\begin{aligned}
&\|x_0 + y_0\|\|x_0 - y_0\| \\
&\quad = \|x_0 + y_0\|\|R_U(x_0 - y_0)\| \\
&\quad \geqslant \langle x_0 + y_0, R_U(x_0 - y_0)\rangle \\
&\quad = \langle x_0 + y_0, P_U(x_0 - y_0) - P_{U^\perp}(x_0 - y_0)\rangle \\
&\quad = \langle x_0 + y_0, P_U(x_0 - y_0)\rangle + \langle x_0 + y_0, P_{U^\perp}(y_0 - x_0)\rangle \\
&\quad = \langle P_U(x_0 + y_0), P_U(x_0 - y_0)\rangle + \langle P_{U^\perp}(x_0 + y_0), P_{U^\perp}(y_0 - x_0)\rangle \\
&\quad = \langle P_U x_0 + P_U y_0, P_U x_0 - P_U y_0\rangle + \langle P_{U^\perp} x_0 + P_{U^\perp} y_0, P_{U^\perp} y_0 - P_{U^\perp} x_0\rangle \\
&\quad = \|P_U x_0\|^2 - \|P_U y_0\|^2 + \|P_{U^\perp} y_0\|^2 - \|P_{U^\perp} x_0\|^2.
\end{aligned}$$

This implies

$$\|P_U x_0\|^2 + \|P_{U^\perp} y_0\|^2 - \|x_0 + y_0\|\|x_0 - y_0\| \leqslant \|P_U y_0\|^2 + \|P_{U^\perp} x_0\|^2. \tag{2.50}$$

Adding to this $\|P_{U^\perp} x_0\|^2 + \|P_U y_0\|^2$ yields

$$\|x_0\|^2 + \|y_0\|^2 - \|x_0 + y_0\|\|x_0 - y_0\| \leqslant 2\|P_U y_0\|^2 + 2\|P_{U^\perp} x_0\|^2. \tag{2.51}$$

Finally, dividing (2.51) by 4 gives (2.48) and we are done. □

Proposition 2.21. *Suppose that $(x_0, y_0) \in X \times X$ satisfies $x_0 = \pm y_0$. Then*

$$P_C(x_0, y_0) = \bigcup_{U \text{ is a closed subspace of } X} \big\{(P_U x_0, P_{U^\perp} y_0)\big\}; \tag{2.52}$$

this can also be written as

$$P_C(x_0, y_0) = \big\{(0, y_0)\big\} \cup \Big\{(0, y_0) + \big(\langle u, x_0\rangle u, -\langle u, y_0\rangle u\big) \;\Big|\; u \in \mathbb{S}\Big\}, \tag{2.53}$$

where $\mathbb{S} = \big\{z \in X \;\big|\; \|z\| = 1\big\}$ *is the unit sphere of* X*.*

Proof. In view of Lemma 2.9, let U be an arbitrary closed subspace of X. Then

$$P_{U\times U^\perp}(x_0, y_0) = \big(P_U x_0, P_{U^\perp} y_0\big). \tag{2.54}$$

Let f be defined as in (2.15). Using the fact that $P_U + P_{U^\perp} = \mathrm{Id}$ in (2.55b), the assumption that $x_0 = \pm y_0$ in (2.55c), and the linearity of P_U in (2.55d), we see that

$$\begin{aligned} f\big(P_U x_0, P_{U^\perp} y_0\big) &= \tfrac{1}{2}\|x_0 - P_U x_0\|^2 + \tfrac{1}{2}\|y_0 - P_{U^\perp} y_0\|^2 && \text{(2.55a)}\\ &= \tfrac{1}{2}\|P_{U^\perp} x_0\|^2 + \tfrac{1}{2}\|P_U y_0\|^2 && \text{(2.55b)}\\ &= \tfrac{1}{2}\|P_{U^\perp} x_0\|^2 + \tfrac{1}{2}\|P_U(\pm x_0)\|^2 && \text{(2.55c)}\\ &= \tfrac{1}{2}\|P_{U^\perp} x_0\|^2 + \tfrac{1}{2}\|P_U x_0\|^2 && \text{(2.55d)}\\ &= \tfrac{1}{2}\|x_0\|^2 && \text{(2.55e)}\\ &= \tfrac{1}{4}\|x_0\|^2 + \tfrac{1}{4}\|y_0\|^2 && \text{(2.55f)} \end{aligned}$$

is *independent* of U! This proves (2.52).

We now tackle (2.53) via (2.12). So let U be a linear subspace of X with $\dim U \leqslant 1$.

If $U = \{0\}$, then $U^\perp = X$ and $(P_U x_0, P_{U^\perp} y_0) = (0, y_0)$ which is the first term on the right side of (2.53).

Now assume that $\dim U = 1$, say $U = \mathbb{R}u$, where $u \in \mathbb{S}$. Then $\|u\| = 1$, $P_U x_0 = \langle u, x_0\rangle u$ and $P_{U^\perp} y_0 = y_0 - \langle u, y_0\rangle u$. Hence

$$\big(P_U x_0, P_{U^\perp} y_0\big) = (0, y_0) + \big(\langle u, x_0\rangle u, -\langle u, y_0\rangle u\big) \tag{2.56}$$

which yields the second term on the right side of (2.53). □

2.4 Closed-form projection formula onto the cross

Let us summarize our work in one convenient theorem.

Theorem 2.22 (main result). *The set C defined in* (2.1) *is proximinal. Let* $(x_0, y_0) \in X \times X$*. Then* exactly one *of the following three cases occurs.*

(i) $\langle x_0, y_0\rangle = 0$, $\frac{1}{2}d_C^2(x_0, y_0) = 0$*, and*

$$P_C(x_0, y_0) = (x_0, y_0). \tag{2.57}$$

(ii) $\langle x_0, y_0\rangle \neq 0$, $x_0 \neq \pm y_0$,

$$P_C(x_0, y_0) = \frac{1}{1-\lambda^2}\big(x_0 - \lambda y_0, y_0 - \lambda x_0\big), \tag{2.58}$$

and

$$\tfrac{1}{2}d_C^2(x_0, y_0) = \tfrac{1}{2}\lambda\left\langle x_0, y_0\right\rangle = \frac{\|x_0\|^2 + \|y_0\|^2 - \|x_0 + y_0\|\|x_0 - y_0\|}{4}, \tag{2.59}$$

where

$$\lambda := \frac{\|x_0\|^2 + \|y_0\|^2 - \|x_0 + y_0\|\|x_0 - y_0\|}{2\left\langle x_0, y_0\right\rangle} \neq \pm 1. \tag{2.60}$$

(iii) $\langle x_0, y_0\rangle \neq 0$, $x_0 = \pm y_0$, $\tfrac{1}{2}d_C^2(x_0, y_0) = \tfrac{1}{4}(\|x_0\|^2 + \|y_0\|^2)$, *and*

$$P_C(x_0, y_0) = \big\{(0, y_0)\big\} \cup \Big\{(0, y_0) + \big(\langle u, x_0\rangle u, -\langle u, y_0\rangle u\big) \;\Big|\; u \in \mathbb{S}\Big\}$$

is not a singleton, where $\mathbb{S} = \big\{z \in X \;\big|\; \|z\| = 1\big\}$ *is the unit sphere of* X*.*

Proof. Combine Proposition 2.19, Proposition 2.20, and Proposition 2.21. □

Remark 2.23. Several results regarding Theorem 2.22 are in order.

1. **(relationship to Elser's work)** Case (ii) was considered by Elser who obtained the basic structure of (2.58) in a more general setting; see [40, equation (19)]. However, his analysis is carried out in the finite-dimensional setting. And neither was the explicit formula for λ in (2.60) presented nor the case (iii) discussed.
2. A convenient *selection* of P_C in case (iii) is $(0, y_0)$ or $(x_0, 0)$.
3. If we work in $X = \mathbb{R}^n$ and we require the complete projection in case (iii), then we may invoke Fact 2.8.
4. It is possible to subsume case (i) into case (ii) by setting $\lambda = 0$ and using (2.58) to obtain (2.57).
5. A tight *injective* parametrization in case (iii) is

   $$P_C(x_0, y_0) = \big\{(0, y_0)\big\} \uplus \biguplus_{u \in \mathbb{S} \text{ and } \langle u, x_0\rangle > 0} \Big\{(0, y_0) + \big(\langle u, x_0\rangle u, -\langle u, y_0\rangle u\big)\Big\},$$

   where "$\uplus$" denotes disjoint union. Hence the *cardinality* of $P_C(x_0, y_0)$ is the same as the cardinality of $\mathbb{S}$.

When $X = \mathbb{R}$, then Theorem 2.22 simplifies to the following:

Example 2.24. Suppose that $X = \mathbb{R}$. Then

$$C = \big\{(x, y) \in \mathbb{R}^2 \;\big|\; xy = 0\big\} = (\mathbb{R} \times \{0\}) \cup (\{0\} \times \mathbb{R}). \tag{2.61}$$

Let $(x_0, y_0) \in \mathbb{R}^2$. Then exactly one of the following cases holds.

1. $|x_0| \neq |y_0|$ and

$$P_C(x_0, y_0) = \begin{cases} (0, y_0), & \text{if } |x_0| < |y_0|; \\ (x_0, 0), & \text{if } |x_0| > |y_0|. \end{cases} \tag{2.62}$$

2. $|x_0| = |y_0|$ and
$$P_C(x_0, y_0) = \big\{(x_0, 0), (0, y_0)\big\}. \tag{2.63}$$

For an illustration, see Figure 2.1.

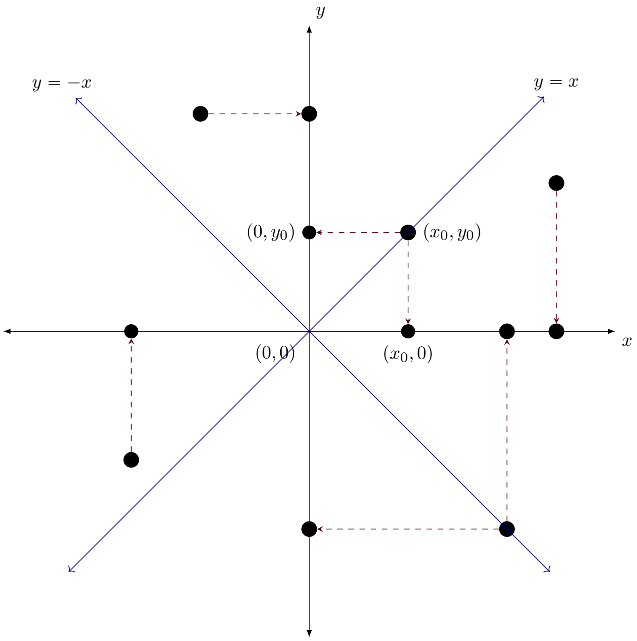

Figure 2.1: Projecting onto the cross C when $X = \mathbb{R}$ (see Example 2.24).

2.5 The alternative Wolkowicz approach

I am grateful to the external examiner, Prof. Henry Wolkowicz, who suggested an alternative approach via [57] (see also [65] and [66]) to the main result of this chapter. Let me outline this approach now.

Fact 2.25. *(Pong–Wolkowicz, see [57, Theorem 2.3 and Theorem 2.1(ii)]) Suppose the underlying space is $Z = \mathbb{R}^n$ on which we consider the* objective function
$$f(x) = x^T Ax - 2a^T x, \tag{2.64}$$
where $A \in \mathbb{R}^{n\times n}$ and $a \in \mathbb{R}^n$. The constraint function *is*
$$c(x) = x^T Bx, \tag{2.65}$$
where $B \in \mathbb{R}^{n\times n}$ and $B \neq 0$. The Generalized Trust Region Subproblem (GTRS) *asks to*
$$\textit{minimize } f(x) \textit{ subject to } \ell \leqslant c(x) \leqslant u, \tag{2.66}$$

where $-\infty < \ell \leqslant u < +\infty$. Recall that $B \neq 0$, and assume that (GTRS) has a feasible solution and is bounded below and that there exists $\hat{X} \in \mathbb{R}^{n\times n}$ such that $\operatorname{tr}\big(B\hat{X}\big) \in \operatorname{ri}[\ell, u]$. Then $x \in \mathbb{R}^n$ is a solution to (GTRS) if and only if there exists $\lambda \in \mathbb{R}$, written as $\lambda = \lambda_+ - \lambda_-$ where $\lambda_+ = \max\{\lambda, 0\}$ and $\lambda_- = -\min\{\lambda, 0\}$ such the following hold:

$$(A - \lambda B)x = a \tag{2.67a}$$

$$A - \lambda B \succeq 0 \tag{2.67b}$$

$$\ell \leqslant x^T Bx \leqslant u \tag{2.67c}$$

$$\lambda_+(\ell - x^T Bx) = 0 \tag{2.67d}$$

$$\lambda_-(x^T Bx - u) = 0. \tag{2.67e}$$

We note that the derivation of Fact 2.25 in [57] shows that Fact 2.25 is quite nontrivial.

We now specialize Fact 2.25 to help characterize the projection onto the cross. Indeed, let

$$A = \begin{bmatrix} \mathrm{Id} & 0 \\ 0 & \mathrm{Id} \end{bmatrix}, \quad a = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & \mathrm{Id} \\ \mathrm{Id} & 0 \end{bmatrix}, \quad \ell = 0 = u, \tag{2.68}$$

where Id is the identity matrix on $\mathbb{R}^n$. Then the objective function turns into

$$f(x, y) = [x^T \ y^T]A \begin{bmatrix} x \\ y \end{bmatrix} - 2a^T \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.69a}$$

$$= [x^T \ y^T] \begin{bmatrix} \mathrm{Id} & 0 \\ 0 & \mathrm{Id} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} - 2[x_0^T \ y_0^T] \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.69b}$$

$$= x^T x + y^T y - 2x_0^T x - 2y_0^T y \tag{2.69c}$$

$$= \|x - x_0\|^2 + \|y - y_0\|^2 - \|x_0\|^2 - \|y_0\|^2 \tag{2.69d}$$

$$= \|(x, y) - (x_0, y_0)\|^2 - \|(x_0, y_0)\|^2. \tag{2.69e}$$

On the other hand,

$$g([x \ y]^T) = [x^T \ y^T]B \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.70a}$$

$$= [x^T \ y^T] \begin{bmatrix} 0 & \mathrm{Id} \\ \mathrm{Id} & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.70b}$$

$$= 2x^T y. \tag{2.70c}$$

Because $\ell = 0 = u$, it is clear that

$$(x, y) \text{ solves GTRS in this case} \Leftrightarrow (x, y) \in P_C(x_0, y_0),$$

which is precisely the topic of this chapter. (We could also work with $B/2$, which would have reduced $g(x, y)$ to $x^T y$ but B works because $\ell = 0 = u$. We avoid the fractions that occur when working with $B/2$.)

Because $\ell = u$, we see that Fact 2.25 turns into the following:

$$(x, y) \text{ is in projection onto the cross of the point } (x_0, y_0) \tag{2.71}$$

if and only if

$$(A - \lambda B) \begin{bmatrix} x \\ y \end{bmatrix} = a, \quad A - \lambda B \succeq 0, \quad 2x^T y = 0 \tag{2.72}$$

and this is also equivalent to

$$\begin{bmatrix} \mathrm{Id} & -\lambda \mathrm{Id} \\ -\lambda \mathrm{Id} & \mathrm{Id} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}, \quad \begin{bmatrix} \mathrm{Id} & -\lambda \mathrm{Id} \\ -\lambda \mathrm{Id} & \mathrm{Id} \end{bmatrix} \succeq 0, \quad x^T y = 0. \tag{2.73}$$

Note that

$$\ker \begin{bmatrix} \mathrm{Id} & -\lambda \mathrm{Id} \\ -\lambda \mathrm{Id} & \mathrm{Id} \end{bmatrix} = \begin{cases} \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}, & \text{if } |\lambda| \neq 1; \\ \left\{ \begin{bmatrix} z \\ z \end{bmatrix} \middle| \, z \in Z \right\}, & \text{if } \lambda = 1; \\ \left\{ \begin{bmatrix} z \\ -z \end{bmatrix} \middle| \, z \in Z \right\}, & \text{if } \lambda = -1; \end{cases} \tag{2.74}$$

and that

$$\begin{bmatrix} \mathrm{Id} & -\lambda \mathrm{Id} \\ -\lambda \mathrm{Id} & \mathrm{Id} \end{bmatrix}^{\dagger} = \begin{cases} \frac{1}{1-\lambda^2} \begin{bmatrix} \mathrm{Id} & \lambda \mathrm{Id} \\ \lambda \mathrm{Id} & \mathrm{Id} \end{bmatrix}, & \text{if } |\lambda| \neq 1; \\ \frac{1}{4} \begin{bmatrix} \mathrm{Id} & -\mathrm{Id} \\ -\mathrm{Id} & \mathrm{Id} \end{bmatrix}, & \text{if } \lambda = 1; \\ \frac{1}{4} \begin{bmatrix} \mathrm{Id} & \mathrm{Id} \\ \mathrm{Id} & \mathrm{Id} \end{bmatrix}, & \text{if } \lambda = -1. \end{cases} \tag{2.75}$$

(The last formula may be verified by checking the Penrose conditions for the

Moore-Penrose inverse.) It follows that

$$\begin{bmatrix} \mathrm{Id} & -\lambda\,\mathrm{Id} \\ -\lambda\,\mathrm{Id} & \mathrm{Id} \end{bmatrix}^{\dagger} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{cases} \frac{1}{1-\lambda^2} \begin{bmatrix} x_0 + \lambda y_0 \\ y_0 + \lambda x_0 \end{bmatrix}, & \text{if } |\lambda| \neq 1; \\ \frac{1}{4} \begin{bmatrix} x_0 - y_0 \\ y_0 - x_0 \end{bmatrix}, & \text{if } \lambda = 1; \\ \frac{1}{4} \begin{bmatrix} x_0 + y_0 \\ x_0 + y_0 \end{bmatrix}, & \text{if } \lambda = -1. \end{cases} \tag{2.76}$$

Next, we note that

$$\begin{aligned} [v^T \ w^T] \begin{bmatrix} \mathrm{Id} & -\lambda\,\mathrm{Id} \\ -\lambda\,\mathrm{Id} & \mathrm{Id} \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} &= [v^T \ w^T] \begin{bmatrix} v - \lambda w \\ -\lambda v + w \end{bmatrix} && \text{(2.77a)} \\ &= v^T(v - \lambda w) + w^T(-\lambda v + w) && \text{(2.77b)} \\ &= \|v\|^2 - 2\lambda \langle v, w\rangle + \|w\|^2 && \text{(2.77c)} \\ &= \|v - \lambda w\|^2 + (1 - \lambda^2)\|w\|^2 && \text{(2.77d)} \end{aligned}$$

which is nonnegative if and only if $\lambda^2 \leqslant 1$; equivalently,

$$A - \lambda B \succeq 0 \quad \Leftrightarrow \quad |\lambda| \leqslant 1. \tag{2.78}$$

Therefore,

$$\begin{aligned} & \begin{bmatrix} \mathrm{Id} & -\lambda\,\mathrm{Id} \\ -\lambda\,\mathrm{Id} & \mathrm{Id} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} && \text{(2.79a)} \\ \Leftrightarrow & \begin{bmatrix} x \\ y \end{bmatrix} \in \begin{bmatrix} \mathrm{Id} & -\lambda\,\mathrm{Id} \\ -\lambda\,\mathrm{Id} & \mathrm{Id} \end{bmatrix}^{\dagger} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \ker \begin{bmatrix} \mathrm{Id} & -\lambda\,\mathrm{Id} \\ -\lambda\,\mathrm{Id} & \mathrm{Id} \end{bmatrix} && \text{(2.79b)} \\ \Leftrightarrow & \begin{bmatrix} x \\ y \end{bmatrix} \in \begin{cases} \left\{ \frac{1}{1-\lambda^2} \begin{bmatrix} x_0 + \lambda y_0 \\ y_0 + \lambda x_0 \end{bmatrix} \right\}, & \text{if } |\lambda| \neq 1; \\ \left\{ \frac{1}{4} \begin{bmatrix} x_0 - y_0 + z \\ y_0 - x_0 + z \end{bmatrix} \middle| \, z \in Z \right\}, & \text{if } \lambda = 1; \\ \left\{ \frac{1}{4} \begin{bmatrix} x_0 + y_0 + z \\ x_0 + y_0 - z \end{bmatrix} \middle| \, z \in Z \right\}, & \text{if } \lambda = -1. \end{cases} && \text{(2.79c)} \end{aligned}$$

To sum up: we get that $(x, y) \in P_C(x_0, y_0)$ if and only if there exists $\lambda \in \mathbb{R}$ such that (2.78) holds, (2.79) holds, and (x, y) lies in the cross ($x^T y = 0$). We now discuss these conditions on λ which we know must lie in $[-1, 1]$. We shall assume that $\langle x_0, y_0\rangle \neq 0$ for otherwise we are done.

Case 1: $|\lambda| < 1$.
Then

$$\begin{aligned}
\begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{1-\lambda^2}\begin{bmatrix} x_0 + \lambda y_0 \\ y_0 + \lambda x_0 \end{bmatrix} &\in C \qquad (2.80a)\\
\Leftrightarrow \langle x_0 + \lambda y_0, y_0 + \lambda x_0 \rangle &= 0 \qquad (2.80b)\\
\Leftrightarrow (1+\lambda^2)\langle x_0, y_0\rangle + \lambda(\|x_0\|^2 + \|y_0\|^2) &= 0 \qquad (2.80c)\\
\Leftrightarrow \lambda &= \frac{-(\|x_0\|^2 + \|y_0\|^2) \pm \sqrt{(\|x_0\|^2 + \|y_0\|^2)^2 - 4\langle x_0, y_0\rangle^2}}{2\langle x_0, y_0\rangle} \qquad (2.80d)\\
\Leftrightarrow \lambda &= \frac{-(\|x_0\|^2 + \|y_0\|^2)}{2\langle x_0, y_0\rangle} \pm \\
&\quad \frac{\sqrt{(\|x_0\|^2 + \|y_0\|^2 + 2\langle x_0, y_0\rangle)(\|x_0\|^2 + \|y_0\|^2 - 2\langle x_0, y_0\rangle)}}{2\langle x_0, y_0\rangle} \qquad (2.80e)\\
\Leftrightarrow \lambda &= \frac{-(\|x_0\|^2 + \|y_0\|^2) \pm \sqrt{\|x_0 + y_0\|^2\|x_0 - y_0\|^2}}{2\langle x_0, y_0\rangle} \qquad (2.80f)\\
\Leftrightarrow \lambda &= \frac{-(\|x_0\|^2 + \|y_0\|^2) \pm \|x_0 + y_0\|\|x_0 - y_0\|}{2\langle x_0, y_0\rangle}. \qquad (2.80g)
\end{aligned}$$

Next, arguing as in Proposition 2.18 in the thesis, we see that $|\lambda| < 1$ requires $x_0 \neq \pm y_0$ and in this case we pick the λ yielding the smaller absolute value, i.e.,

$$\lambda = \frac{-(\|x_0\|^2 + \|y_0\|^2) + \|x_0 + y_0\|\|x_0 - y_0\|}{2\langle x_0, y_0\rangle}. \qquad (2.81)$$

This corresponds to Theorem 2.22(ii) (with λ replaced by $-\lambda$).

Case 2: $\lambda = 1$.
Then for every $z \in X$, we have

$$\begin{aligned}
\begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{4}\begin{bmatrix} x_0 - y_0 + z \\ y_0 - x_0 + z \end{bmatrix} \in C &\Leftrightarrow \langle x_0 - y_0 + z, y_0 - x_0 + z\rangle = 0 \qquad (2.82)\\
&\Leftrightarrow \langle z + x_0 - y_0, z - (x_0 - y_0)\rangle = 0 \qquad (2.83)\\
&\Leftrightarrow \|z\|^2 = \|x_0 - y_0\|^2 \qquad (2.84)\\
&\Leftrightarrow \|z\| = \|x_0 - y_0\|. \qquad (2.85)
\end{aligned}$$

Now again from Proposition 2.18, we know that we cannot be in *Case 1* and therefore it must be the case when $y_0 = -x_0$.

Case 3: $\lambda = -1$.
This is similar to *Case 2* and corresponds to the case when $y_0 = x_0$. We omit the details.

Let us summarize our findings in the following remark.

Remark 2.26.

1. Using the Pong-Wolkowicz result (Fact 2.25) one indeed quickly gets to cases on λ, which cuts down on some of the work. However, one still needs to untangle the cases on λ to conditions on the original data vector (x_0, y_0). Another advantage of the Pong-Wolkowicz approach is that it leads to a more concise parametrization of the multi-valued case (see *Case 2* above).
2. We expect that the Pong-Wolkowicz result will also help to find the projections in Chapter 3 and Chapter 4.
3. Finally, we expect the Pong-Wolkowicz result will help to tackle some of the open problems outlined in Section 7.2.

Chapter 3

Projections onto hyperbolas or bilinear constraints

This chapter is based on the paper [14] titled "Projections onto hyperbolas or bilinear constraint sets in Hilbert spaces" which appeared in the Journal of Global Optimization 86 (2023), 25–36, which was already cited in papers by researchers from Cornell and MIT [32, 36, 45]. Sets of bilinear constraints are important in machine learning, inverse problems, and many engineering applications. Mathematically speaking, they are hyperbolas in a product space. Here, we outline the process of finding the projections and give a complete formula for projections onto sets of bilinear constraints or hyperbolas in a general Hilbert space.

In various learning models [39, 40], Elser utilizes projections onto the *bilinear constraint* set:

$$C_\gamma := \big\{(x, y) \in X \times X \;\big|\; \langle x, y\rangle = \gamma\big\} \tag{3.1}$$

where $\gamma \in \mathbb{R}$ is a fixed constant. Up to a rotation for $\gamma \neq 0$, this is just a hyperbola or quadratic surface in $X \times X$. Finding projections onto quadratic curves or surfaces algorithmically have many practical applications; see, e.g., [25, 26, 39, 40]. If $\gamma = 0$ in C_γ, then C_0 becomes the cross in $X \times X$. The projection formula for the cross has been thoroughly investigated [12] and Chapter 2. In [39, 40], although Elser has provided some results on projections onto C_γ only when $X = \mathbb{R}^n$, complete mathematical details are not presented.

The goal of this chapter is to give a complete analysis of the projection onto the set C_γ with $\gamma \neq 0$, and extend the results to a general Hilbert space.

The remainder of the chapter is organized as follows. Section 3.1 gives some general properties of bilinear constraint sets. In Section 3.2 we focus on full mathematical details for the existence and explicit formula of projections onto hyperbolas. Using results from Section 3.2, in Section 3.3 and Section 3.4 we provide explicit formulas for projections onto sets of bilinear constraints. Our notation is standard and follows largely [8, 62].

3.1 General properties of C_γ

Theorem 3.1 (**convex hull of C_γ hyperbolas**). *Let $\gamma > 0$. Then* conv $C_\gamma = X \times X$*, where* conv *denotes the convex hull.*

Proof. Let $(x_0, y_0) \in X \times X$.
Case 1: $x_0 = y_0 \neq 0$.

1. $\langle x_0, x_0\rangle < \gamma$. Take $(x_0, x_0) + t(x_0, x_0)$ and $\langle (1+t)x_0, (1+t)x_0\rangle = \gamma$, then we have

$$(1+t)^2 = \frac{\gamma}{\|x_0\|^2} \Rightarrow t = -1 \pm \sqrt{\frac{\gamma}{\|x_0\|^2}},$$
$$t_1 = -1 + \sqrt{\frac{\gamma}{\|x_0\|^2}} > 0 \quad \text{because} \quad \frac{\gamma}{\|x_0\|^2} > 0, \quad \text{and}$$
$$t_2 = -1 - \sqrt{\frac{\gamma}{\|x_0\|^2}} < 0.$$

   Then, $(x_0, x_0) = \frac{-t_2}{t_1 - t_2}\big((x_0, x_0) + t_1(x_0, x_0)\big) + \frac{t_1}{t_1 - t_2}\big((x_0, x_0) + t_2(x_0, x_0)\big)$ and

$$(x_0, x_0) + t_i(x_0, x_0) \in C_\gamma \quad \text{for} \quad i = 1, 2.$$

2. $\langle x_0, x_0\rangle > \gamma$. Take $(x_0, x_0) + t(x_0, -x_0)$ and $\langle (1+t)x_0, (1-t)x_0\rangle = \gamma$, then we have

$$1 - t^2 = \frac{\gamma}{\|x_0\|^2} \Rightarrow t = \pm\sqrt{1 - \frac{\gamma}{\|x_0\|^2}},$$
$$t_1 = \sqrt{1 - \frac{\gamma}{\|x_0\|^2}} > 0, \quad \text{and} \quad t_2 = -\sqrt{1 - \frac{\gamma}{\|x_0\|^2}} < 0.$$

   Then, $(x_0, x_0) = \frac{-t_2}{t_1 - t_2}\big((x_0, x_0) + t_1(x_0, -x_0)\big) + \frac{t_1}{t_1 - t_2}\big((x_0, x_0) + t_2(x_0, -x_0)\big)$ and

$$(x_0, -x_0) + t_i(x_0, -x_0) \in C_\gamma \quad \text{for} \quad i = 1, 2.$$

Case 2: $x_0 = -y_0 \neq 0$. Consider $(x_0, -x_0) + t(x_0, x_0)$ and $\langle (1+t)x_0, (-1+t)x_0\rangle = \gamma$, then we have

$$(t^2 - 1)\|x_0\|^2 = \gamma \Rightarrow t = \pm\sqrt{\frac{\gamma}{\|x_0\|^2} + 1},$$
$$t_1 = \sqrt{\frac{\gamma}{\|x_0\|^2} + 1} > 0, \quad \text{and} \quad t_2 = -\sqrt{\frac{\gamma}{\|x_0\|^2} + 1} < 0.$$

Then,

$$(x_0, -x_0) = \frac{-t_2}{t_1 - t_2}\big((x_0, -x_0) + t_1(x_0, x_0)\big) + \frac{t_1}{t_1 - t_2}\big((x_0, -x_0) + t_2(x_0, x_0)\big).$$

Case 3: $x_0 = y_0 = 0$. Take any $v \neq 0$. Consider (tv, tv) with $\langle tv, tv\rangle = \gamma$, then we have

$$t^2\|v\|^2 = \gamma \Rightarrow t = \pm\sqrt{\frac{\gamma}{\|v\|^2}},$$
$$t_1 = \sqrt{\frac{\gamma}{\|v\|^2}} > 0, \quad \text{and} \quad t_2 = -\sqrt{\frac{\gamma}{\|v\|^2}} < 0.$$

Then,

$$(0,0) = \frac{-t_2}{t_1-t_2}\big(t_1v, t_1v\big) + \frac{t_1}{t_1-t_2}\big(t_2v, t_2v\big).$$

Case 4: $x_0 \neq y_0, x_0 \neq -y_0$.

1. $\langle x_0, y_0\rangle > \gamma$. Let $v = \frac{x_0-y_0}{\|x_0-y_0\|} \neq 0$. Consider $(x_0, y_0) + t(v, -v)$ and $\langle x_0 + tv, y_0 - tv\rangle = \gamma$. Then

$$\begin{aligned}\langle x_0, y_0\rangle + t\langle v, -x_0\rangle + t\langle v, y_0\rangle - t^2 &= \gamma,\\ \langle x_0, y_0\rangle + t\langle v, y_0 - x_0\rangle - t^2 &= \gamma,\\ \langle x_0, y_0\rangle - \gamma + t\langle v, y_0 - x_0\rangle - t^2 &= 0,\\ t^2 + \|x_0 - y_0\|t - (\langle x_0, y_0\rangle - \gamma) &= 0.\end{aligned}$$

   We have

$$\begin{aligned}t &= \frac{-\|y_0 - x_0\| \pm \sqrt{\|y_0 - x_0\|^2 + 4(\langle x_0, y_0\rangle - \gamma)}}{2},\\ &= \frac{-\|y_0 - x_0\| \pm \sqrt{\|y_0 + x_0\|^2 - 4\gamma}}{2},\end{aligned}$$

   where

$$t_2 = \frac{-\|y_0 - x_0\| - \sqrt{\|y_0 + x_0\|^2 - 4\gamma}}{2} < 0,$$

   and note that

$$\begin{aligned}\|y_0 + x_0\|^2 - 4\gamma &> \|y_0 + x_0\|^2 - 4\langle x_0, y_0\rangle\\ &= \|y_0 - x_0\|^2.\end{aligned}$$

   So

$$t_1 = \frac{-\|y_0 - x_0\| + \sqrt{\|y_0 + x_0\|^2 - 4\gamma}}{2} > 0.$$

   Then,

$$(x_0, y_0) = \frac{-t_2}{t_1-t_2}\big((x_0, y_0) + t_1(v, -v)\big) + \frac{t_1}{t_1-t_2}\big((x_0, y_0) + t_2(v, -v)\big).$$

2. $\langle x_0, y_0\rangle < \gamma$. Let $v = \frac{x_0+y_0}{\|x_0+y_0\|} \neq 0$. Consider $(x_0, y_0) + t(v, v)$ and $\langle x_0 + tv, y_0 - tv\rangle = \gamma$. Then,

$$\begin{aligned}\langle x_0, y_0\rangle + t\langle x_0, v\rangle + t\langle y_0, v\rangle + t^2 &= \gamma,\\ \langle x_0, y_0\rangle + t\langle x_0 + y_0, v\rangle + t^2 &= \gamma,\\ \langle x_0, y_0\rangle + t\|x_0 + y_0\| + t^2 &= \gamma,\\ t^2 + \|x_0 + y_0\|t + (\langle x_0, y_0\rangle - \gamma) &= 0.\end{aligned}$$

We have

$$\begin{aligned} t &= \frac{-\|x_0+y_0\| \pm \sqrt{\|x_0+y_0\|^2 - 4\langle x_0, y_0\rangle + 4\gamma}}{2}, \\ &= \frac{-\|x_0+y_0\| \pm \sqrt{\|x_0-y_0\|^2 + 4\gamma}}{2}. \end{aligned}$$

Then,

$$t_1 = \frac{-\|x_0+y_0\| + \sqrt{\|x_0-y_0\|^2 + 4\gamma}}{2}.$$

Note that

$$\begin{aligned} \|x_0-y_0\|^2 + 4\gamma &> \|x_0-y_0\|^2 + 4\langle x_0, y_0\rangle \\ &= \|x_0+y_0\|^2. \end{aligned}$$

So $t_1 > 0$, and

$$t_2 = \frac{-\|x_0+y_0\| - \sqrt{\|x_0-y_0\|^2 + 4\gamma}}{2} < 0.$$

Then,

$$(x_0, y_0) = \frac{-t_2}{t_1 - t_2}\big((x_0, y_0) + t_1(v, v)\big) + \frac{t_1}{t_1 - t_2}\big((x_0, y_0) + t_2(v, v)\big).$$

□

In an infinite-dimensional space, the existence of projection onto a set often requires the weak closedness of the set. Our next result says that although the set C_γ is norm closed it is not weakly closed in $X \times X$.

Proposition 3.2. *The set C_γ is closed in the norm topology but not closed in the weak topology in $X \times X$. In fact, $\overline{C_\gamma}^{\text{weak}} = X \times X$.*

Proof. Evidently, C_γ is norm closed. We show that $\overline{C_\gamma}^{\text{weak}} = X \times X$. Let $(x, y) \in X \times X$. Consider $S = \operatorname{span}\{x, y\}$. The orthogonal decomposition theorem gives $X = S \bigoplus S^\perp$. Because X is infinite-dimensional and S is at most two-dimensional, $S^\perp$ is infinite dimensional, so any orthonormal base of $S^\perp$ must have a sequence $(e_n)_{n\in\mathbb{N}}$ which converges weakly to 0, i.e., $e_n \rightharpoonup 0$; see, e.g., [47]. Also $e_n \perp x$ and $e_n \perp y$. Set $\xi = \gamma - \langle x, y\rangle$. Then

$$\begin{aligned} \langle x - \xi e_n, y - e_n\rangle &= \langle x, y\rangle - \langle x, e_n\rangle - \xi\langle e_n, y\rangle + \xi\langle e_n, e_n\rangle && (3.2) \\ &= \langle x, y\rangle + \xi = \gamma, && (3.3) \end{aligned}$$

so $(x - \xi e_n, y - e_n) \in C_\gamma$. Since $(x - \xi e_n, y - e_n) \rightharpoonup (x, y)$, we have $(x, y) \in \overline{C_\gamma}^{\text{weak}}$. Because $(x, y) \in X \times X$ was arbitrary, we conclude that $X \times X \subseteq \overline{C}^{\text{weak}}$. □

Proposition 3.2 indicates that finding P_{C_γ}, i.e., projections onto C_γ, might be complicated in a general Hilbert space. This seemingly difficult issue can be completely avoided by utilizing the structures of the optimization problem.

Recall that a function $f : X \to \left]-\infty, +\infty\right] = \mathbb{R} \cup \{+\infty\}$ is said to be *convex* if its domain, $\operatorname{dom} f = \{x \in X \mid f(x) < +\infty\}$, is a convex set and $\forall x, y \in X$, and $0 < \lambda < 1$, we have

$$f\big(\lambda x + \big(1 - \lambda\big)y\big) \leqslant \lambda f\big(x\big) + (1 - \lambda)f\big(y\big). \tag{3.4}$$

Let S be a convex and closed subset of X. Then the *indicator* function of S at x defined as

$$\iota_{\mathrm{S}}\big(x\big) := \begin{cases} 0, & \text{if } x \in \mathrm{S}; \\ +\infty, & \text{otherwise.} \end{cases} \tag{3.5}$$

is convex and lower semicontinuous.

Our next result says that when $\gamma \neq 0$, locally around the set C_γ the projection onto the set is always single-valued.

Proposition 3.3. *Let $\gamma \neq 0$, and $C_\gamma = \big\{(x, y) \in X \times X \;\big|\; \langle x, y\rangle = \gamma\big\}$. For every $(x, y) \in C_\gamma$, then the following hold:*

(i) C_γ is prox-regular at (x, y).

(ii) There exists a neighborhood of (x, y) on which the projection mapping onto C_γ is monotone and Lipschitz continuous.

Proof. (i): Write $\iota_{C_\gamma}(x, y) = \iota_{\{\gamma\}}(h(x, y))$ where $h(x, y) = \langle x, y\rangle$. Let $(x, y) \in C_\gamma$. Then $\nabla h(x, y) = (y, x) \neq (0, 0)$ because of $\gamma \neq 0$. Being a composition of a convex function $\iota_{\{\gamma\}}$ and a twice differentiable function h that is qualified (see [17, Line 3 on page 4]) at (x, y), [17, Proposition 2.4] shows that ι_{C_γ} is prox-regular at (x, y), so is C_γ.

(ii): Apply (i) and [17, Proposition 4.4]. □

Observe that Proposition 3.3 also follows from [62, Proposition 13.32] and [62, Exercise 13.38] when $X = \mathbb{R}^n$; and that when $\gamma = 0$, C_0 is not prox-regular at $(0, 0)$. Although the projection exists locally around C_γ, it is still not clear for the global existence.

3.2 Projections onto hyperbolas

For ease of analysis, we start with

$$C_\gamma := \big\{(x, y) \in X \times X \;\big|\; h(x, y) := \langle x, y\rangle - \gamma = 0\big\}$$

where $\gamma > 0$. Our goal is to find the projection formula $P_{C_\gamma}(x_0, y_0)$ for every $(x_0, y_0) \in X \times X$. That is,

$$\text{minimize} \quad f(x, y) := \|x - x_0\|^2 + \|y - y_0\|^2 \quad \text{subject to} \quad (x, y) \in C_\gamma. \tag{P}$$

3.2.1 Auxiliary problems and existence of projections

To determine the projection operator P_{C_γ} of the set C_γ, we shall introduce two equivalently reformulated problems. First, for every $(u_0, v_0) \in X \times X$, we solve the problem

$$\text{minimize} \quad f_1(u,v) := \|u-u_0\|^2 + \|v-v_0\|^2 \quad \text{subject to} \quad (u,v) \in \tilde{C}_1, \qquad (\tilde{P}_1)$$

where $\tilde{C}_1 := \big\{(u,v) \in X \times X \;\big|\; h_1(u,v) := \|u\|^2 - \|v\|^2 - 2 = 0\big\}$. Next, for every $(\tilde{u}_0, \tilde{v}_0) \in X \times X$ we solve the problem

$$\text{minimize} \quad f_\gamma(\tilde{u},\tilde{v}) := \|\tilde{u}-\tilde{u}_0\|^2 + \|\tilde{v}-\tilde{v}_0\|^2 \quad \text{subject to} \quad (\tilde{u},\tilde{v}) \in \tilde{C}_\gamma, \qquad (\tilde{P}_\gamma)$$

where $\tilde{C}_\gamma := \big\{(\tilde{u},\tilde{v}) \in X \times X \;\big|\; h_\gamma(\tilde{u},\tilde{v}) := \|\tilde{u}\|^2 - \|\tilde{v}\|^2 - 2\gamma = 0\big\}$. Both $\tilde{C}_1$ and $\tilde{C}_\gamma$ are hyperbolas. $(\tilde{P}_1)$ and $(\tilde{P}_\gamma)$ solve for projections $P_{\tilde{C}_1}$ and $P_{\tilde{C}_\gamma}$ respectively, and their connections to P_{C_γ} are given by the Proposition 3.5 below. Recall

Definition 3.4 (Rotation with an angle ϕ)**.** A change of coordinates $(u,v) \in X \times X$, by rotation through an angle ϕ, is defined by

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\phi \operatorname{Id} & -\sin\phi \operatorname{Id} \\ \sin\phi \operatorname{Id} & \cos\phi \operatorname{Id} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}.$$

Put

$$A_\phi := \begin{bmatrix} \cos\phi \operatorname{Id} & -\sin\phi \operatorname{Id} \\ \sin\phi \operatorname{Id} & \cos\phi \operatorname{Id} \end{bmatrix}.$$

Then $A_\phi A_{-\phi} = \operatorname{Id} = A_{-\phi} A_\phi$ and $A_\phi^{-1} = A_{-\phi}$. The relationships among $P_{C_\gamma}, P_{\tilde{C}_\gamma}$ and $P_{\tilde{C}_1}$ are summarized below.

Proposition 3.5. *The following hold:*
(i) $P_{\tilde{C}_\gamma} = \sqrt{\gamma} P_{\tilde{C}_1}(\operatorname{Id}/\sqrt{\gamma})$.
(ii) $P_{C_\gamma} = A_{\pi/4} P_{\tilde{C}_\gamma} A_{-\pi/4}$.

Proof. (i): $(\tilde{P}_1)$ is equivalent to $(\tilde{P}_\gamma)$ by a change of variables of scaling $\operatorname{Id}/\sqrt{\gamma}$. Indeed, using

$$\begin{bmatrix} u \\ v \end{bmatrix} = \frac{1}{\sqrt{\gamma}} \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix}, \text{ and } \begin{bmatrix} u_0 \\ v_0 \end{bmatrix} = \frac{1}{\sqrt{\gamma}} \begin{bmatrix} \tilde{u}_0 \\ \tilde{v}_0 \end{bmatrix}$$

$h_\gamma(\tilde{u},\tilde{v}) = 0$ becomes $h_1(u,v) = 0$, and $f_\gamma(\tilde{u},\tilde{v})$ becomes $\gamma f_1(u,v) = \gamma(\|u - u_0\|^2 + \|v-v_0\|^2)$.

(ii): (P) is equivalent to $(\tilde{P}_\gamma)$ by a change of variables of rational $A_{\pi/4}$. Indeed, with

$$\begin{bmatrix} x \\ y \end{bmatrix} = A_{\pi/4} \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix}, \text{ and } \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = A_{\pi/4} \begin{bmatrix} \tilde{u}_0 \\ \tilde{v}_0 \end{bmatrix}$$

the objective $f(x, y) = \|x - x_0\|^2 + \|y - y_0\|^2 = \|(x, y) - (x_0, y_0)\|^2$ can be rewritten as

$$\begin{aligned} f(x,y) &= \left\| A_{\frac{\pi}{4}} \begin{bmatrix} \tilde{u} - \tilde{u}_0 \\ \tilde{v} - \tilde{v}_0 \end{bmatrix} \right\|^2 = \begin{bmatrix} \tilde{u} - \tilde{u}_0 \\ \tilde{v} - \tilde{v}_0 \end{bmatrix}^T A_{\frac{\pi}{4}}^T A_{\frac{\pi}{4}} \begin{bmatrix} \tilde{u} - \tilde{u}_0 \\ \tilde{v} - \tilde{v}_0 \end{bmatrix} \\ &= \|\tilde{u} - \tilde{u}_0\|^2 + \|\tilde{v} - \tilde{v}_0\|^2 , \end{aligned}$$

and $h(x, y) = \langle x, y\rangle - \gamma = 0$ becomes $h_\gamma(\tilde{u}, \tilde{v}) = \|\tilde{u}\|^2 - \|\tilde{v}\|^2 - 2\gamma = 0$. □

Recall that a function $f : X \to [-\infty, +\infty]$ is *coercive* if

$$\lim_{\|x\| \to +\infty} f(x) = +\infty. \tag{3.6}$$

In view of Proposition 3.5(ii), to see that $P_{C_\gamma}(x, y) \neq \varnothing$ for every $(x, y) \in X \times X$, the following observation is crucial.

Proposition 3.6. *For every $(\tilde{u}_0, \tilde{v}_0) \in X \times X$, the minimization problem*

$$\begin{aligned} &\textit{minimize} && f_\gamma(\tilde{u}, \tilde{v}) = \|\tilde{u} - \tilde{u}_0\|^2 + \|\tilde{v} - \tilde{v}_0\|^2 && (3.7) \\ &\textit{subject to} && h_\gamma(\tilde{u}, \tilde{v}) = \|\tilde{u}\|^2 - \|\tilde{v}\|^2 - 2\gamma = 0 && (3.8) \end{aligned}$$

always has a solution, i.e., $P_{\tilde{C}_\gamma}(\tilde{u}_0, \tilde{v}_0) \neq \varnothing$.

Proof. We shall illustrate only the case $\tilde{u}_0 \neq 0, \tilde{v}_0 \neq 0$, since the other cases are similar. We claim that the optimization problem is essentially 2-dimensional. To this end, we expand

$$f_\gamma(\tilde{u}, \tilde{v}) = \underbrace{\|\tilde{u}\|^2 - 2\langle \tilde{u}, \tilde{u}_0\rangle + \|\tilde{u}_0\|^2} + \underbrace{\|\tilde{v}\|^2 - 2\langle \tilde{v}, \tilde{v}_0\rangle + \|\tilde{v}_0\|^2}. \tag{3.9}$$

The constraint

$$h_\gamma(\tilde{u}, \tilde{v}) := \|\tilde{u}\|^2 - \|\tilde{v}\|^2 - 2\gamma = 0$$

means that only the norms $\|\tilde{u}\|$ and $\|\tilde{v}\|$ matter. With $\|\tilde{u}\|$ fixed, the Cauchy-Schwarz inequality in a Hilbert space, see, e.g., [47], shows that $\tilde{u} \mapsto \langle \tilde{u}, \tilde{u}_0\rangle$ will be larger so that the first underlined part in f_γ will be smaller when $\tilde{u}$ and $\tilde{u}_0$ are positively co-linear, i.e, $\tilde{u} = \alpha\tilde{u}_0$ for some $\alpha \geqslant 0$. Similarly, for fixed $\|v\|$ the second underlined part in f_γ will be smaller when $\tilde{v} = \beta\tilde{v}_0$ for some $\beta \geqslant 0$. It follows that the optimization problem given by (3.7)-(3.8) is equivalent to

$$\begin{aligned} &\text{minimize} && g(\alpha, \beta) := (1 - \alpha)^2\|\tilde{u}_0\|^2 + (1 - \beta)^2\|\tilde{v}_0\|^2 && (3.10) \\ &\text{subject to} && g_1(\alpha, \beta) := \alpha^2 \|\tilde{u}_0\|^2 - \beta^2 \|\tilde{v}_0\|^2 - 2\gamma = 0, \alpha \geqslant 0, \beta \geqslant 0. && (3.11) \end{aligned}$$

Because $g : \mathbb{R}^2 \to \mathbb{R}$ is continuous and coercive, and $g_1 : \mathbb{R}^2 \to \mathbb{R}$ is continuous, we conclude that the optimization problem given by (3.10)-(3.11) has a solution. □

We are now ready for the investigation of projections onto $\tilde{C}_1, \tilde{C}_\gamma$ and C_γ.

3.2.2 Finding the projection $P_{\tilde{C}_1}$

Note that

$$\nabla f_1(u,v) = (2(u-u_0), 2(v-v_0)) \quad \text{and} \quad \nabla h_1(u,v) = (2u, -2v). \tag{3.12}$$

By [18, Proposition 4.1.1], every solution of $(\tilde{P}_1)$ satisfies the necessary conditions, i.e., the Karush-Kuhn-Tucker (KKT) system of $(\tilde{P}_1)$, given by

$$(1+\lambda)u = u_0, \tag{3.13}$$
$$(1-\lambda)v = v_0, \tag{3.14}$$
$$\|u\|^2 - \|v\|^2 - 2 = 0. \tag{3.15}$$

Lemma 3.7. *Let $u_0 \neq 0, v_0 \neq 0$. If (u,v,λ) verifies (3.13)–(3.15) and (u,v) is an optimal solution of problem $(\tilde{P}_1)$, then the Lagrange multiplier λ satisfies $|\lambda| < 1$.*

Proof. The constraint set
$$\|u\|^2 - \|v\|^2 = 2,$$
has a special structure: replacing u by $-u$, or v by $-v$ the constraint is still verified. Also consider

$$f(u,v) = \|u-u_0\|^2 + \|v-v_0\|^2 = \|u_0\|^2 - 2\langle u_0, u\rangle + \|u\|^2 + \|v_0\|^2 - 2\langle v_0, v\rangle + \|v\|^2,$$

Given u_0 and v_0, for fixed $\|u\|$ and $\|v\|$, $f(u,v)$ becomes smaller if one choose $\langle u_0, u\rangle \geqslant 0$, and $\langle v_0, v\rangle \geqslant 0$. Indeed, one can do so by replacing u by $-u$ or v by $-v$ if needed. Now by (3.13) and (3.14),

$$(1+\lambda)\langle u_0, u\rangle = \|u_0\|^2, \text{ and } (1-\lambda)\langle v_0, v\rangle = \|v_0\|^2.$$

Because $u_0 \neq 0, v_0 \neq 0$, we have $\langle u_0, u\rangle > 0, \langle v_0, v\rangle > 0$ so that $1+\lambda > 0, 1-\lambda > 0$. Hence $|\lambda| < 1$. □

Proposition 3.8. *Let $u_0 \neq 0, v_0 \neq 0$. Define $p := \|u_0\|^2 - \|v_0\|^2$ and $q := \|u_0\|^2 + \|v_0\|^2$. Suppose that $u, v \in X$ and $\lambda \in \mathbb{R}$ verify (3.13)-(3.15) and (u,v) is an optimal solution to $(\tilde{P}_1)$. Then the following hold:*

(i) $u = \frac{u_0}{(1+\lambda)}$, $v = \frac{v_0}{(1-\lambda)}$, *and* $(1-\lambda)^2\|u_0\|^2 - (1+\lambda)^2\|v_0\|^2 = 2(1-\lambda^2)^2$.

(ii) The objective function has $f_1(u,v) = \lambda^2\left(\frac{\|u_0\|^2}{(1+\lambda)^2} + \frac{\|v_0\|^2}{(1-\lambda)^2}\right)$.

(iii) λ *is the unique solution of*

$$H(\lambda) := \frac{(\lambda^2+1)p - 2\lambda q}{2(1-\lambda^2)^2} - 1 = 0 \tag{3.16}$$

in $]-1, 1[$.

Proof. (i): Because $u_0 \neq 0, v_0 \neq 0$, we obtain $|\lambda| < 1$, $u = \frac{u_0}{(1+\lambda)}$ and $v = \frac{v_0}{(1-\lambda)}$ by Lemma 3.7 and (3.13)-(3.14). Then we have the following equivalences

$$\begin{aligned} \|u\|^2 - \|v\|^2 = 2 \Leftrightarrow \frac{\|u_0\|^2}{(1+\lambda)^2} - \frac{\|v_0\|^2}{(1-\lambda)^2} = 2 && (3.17)\\ \Leftrightarrow (1-\lambda)^2 \|u_0\|^2 - (1+\lambda)^2 \|v_0\|^2 = 2(1-\lambda)^2(1+\lambda)^2 && (3.18)\\ \Leftrightarrow (1-\lambda)^2 \|u_0\|^2 - (1+\lambda)^2 \|v_0\|^2 = 2(1-\lambda^2)^2. && (3.19) \end{aligned}$$

(ii): Substitute $u = \frac{u_0}{(1+\lambda)}$ and $v = \frac{v_0}{(1-\lambda)}$ in f_1.
(iii): By (i), we have

$$(1-\lambda)^2 \|u_0\|^2 - (1+\lambda)^2 \|v_0\|^2 = 2(1-\lambda^2)^2. \tag{3.20}$$

Since

$$\begin{aligned} &(1-\lambda)^2 \|u_0\|^2 - (1+\lambda)^2 \|v_0\|^2 && (3.21)\\ &= \big(1+\lambda^2-2\lambda\big)\|u_0\|^2 - \big(1+\lambda^2+2\lambda\big)\|v_0\|^2 && (3.22)\\ &= \lambda^2\big(\|u_0\|^2 - \|v_0\|^2\big) + \big(\|u_0\|^2 - \|v_0\|^2\big) - 2\lambda\big(\|u_0\|^2 + \|v_0\|^2\big) && (3.23)\\ &= \lambda^2 p + p - 2\lambda q, && (3.24) \end{aligned}$$

using (3.24) on left side of (3.20), we obtain $\lambda^2 p + p - 2\lambda q = 2(1-\lambda^2)^2$, a univariate quartic equation in λ, equivalently,

$$H(\lambda) := \frac{(\lambda^2+1)p - 2\lambda q}{2(1-\lambda^2)^2} - 1 = 0. \tag{3.25}$$

We show that (3.25) has a unique solution in $]-1,1[$. Because $u_0 \neq 0, v_0 \neq 0$, we know that $q > |p|$, and using it along-with $|\lambda| < 1$ from Lemma 3.7, we get $p\lambda \leqslant |p||\lambda| \leqslant q|\lambda|$, and

$$\begin{aligned} H'(\lambda) &= \frac{1}{(1-\lambda^2)^3}\big(-q(1+3\lambda^2) + p(\lambda^3+3\lambda)\big) && (3.26)\\ &\leqslant \frac{1}{(1-\lambda^2)^3}\big(-q(1+3\lambda^2) + |p||\lambda|(\lambda^2+3)\big) && (3.27)\\ &\leqslant \frac{1}{(1-\lambda^2)^3}\big(-q(1+3\lambda^2) + q|\lambda|(\lambda^2+3)\big) && (3.28)\\ &= \frac{q}{(1-\lambda^2)^3}\big((|\lambda|-1)^3\big) && (3.29)\\ &= \frac{-q}{((1-|\lambda|)(1+|\lambda|))^3}\big((1-|\lambda|)^3\big) && (3.30)\\ &= \frac{-q}{(1+|\lambda|)^3}. && (3.31) \end{aligned}$$

Since $q = \|u_0\|^2 + \|v_0\|^2 > 0$ and $(1+|\lambda|)^3 > 0$, we have $H'(\lambda) < 0$ on $]-1,1[$, which implies $H(\lambda)$ is strictly decreasing on $]-1,1[$. Notice that $\lambda = \pm 1$ are vertical asymptotes, so that $\lim_{\lambda\to 1^-} H(\lambda) = -\infty$ and $\lim_{\lambda\to -1^+} H(\lambda) = +\infty$. Since H is continuous, strictly decreasing, and its range is $]-\infty,\infty[$, we conclude that $H(\lambda) = 0$ has a unique zero in $]-1,1[$. □

Remark 3.9. As indicated in [40], an approximate solution to (3.16) can be found by the root finding methods such as the Bisection method, Newton's method, or a combined version. See also [22, 37].

Theorem 3.10. *Let $u_0, v_0 \in X$, $u_0 \neq 0, v_0 \neq 0$. Then $P_{\tilde{C}_1}(u_0, v_0)$ is a singleton, and*

$$P_{\tilde{C}_1}(u_0,v_0) = \left\{ \begin{bmatrix} \frac{u_0}{1+\lambda} \\ \\ \frac{v_0}{1-\lambda} \end{bmatrix} \right\},$$

in which λ is the unique root of $H(\lambda) = 0$ in $]-1,1[$ where

$$H(\lambda) = \frac{(\lambda^2+1)p - 2\lambda q}{2(1-\lambda^2)^2} - 1, \text{ and}$$
$$p = \|u_0\|^2 - \|v_0\|^2, \quad q = \|u_0\|^2 + \|v_0\|^2.$$

Proof. Apply Proposition 3.8. □

Theorem 3.11. *Let $u_0, v_0 \in X$ with $u_0 = 0$ or $v_0 = 0$.*

(i) When $u_0 = 0$, we have

$$P_{\tilde{C}_1}(0,v_0) = \left\{ \begin{bmatrix} u \\ \frac{v_0}{2} \end{bmatrix} \;\middle|\; \|u\|^2 = 2 + \frac{\|v_0\|^2}{4} \right\}. \tag{3.32}$$

(ii) a) When $v_0 = 0$ and $\|u_0\| \geqslant 2\sqrt{2}$, we have

$$P_{\tilde{C}_1}(u_0,0) = \left\{ \begin{bmatrix} \frac{u_0}{2} \\ v \end{bmatrix} \;\middle|\; \|v\|^2 = \frac{\|u_0\|^2}{4} - 2 \right\}. \tag{3.33}$$

b) When $v_0 = 0$ and $0 < \|u_0\| < 2\sqrt{2}$, we have

$$P_{\tilde{C}_1}(u_0,0) = \left\{ \begin{bmatrix} \sqrt{2}\frac{u_0}{\|u_0\|} \\ 0 \end{bmatrix} \right\}. \tag{3.34}$$

Proof. By (3.13)–(3.15), $(u,v) \in P_{\tilde{C}_1}(u_0,v_0)$ satisfies:

$$(1+\lambda)u = u_0, \tag{3.35}$$
$$(1-\lambda)v = v_0, \tag{3.36}$$
$$\|u\|^2 - \|v\|^2 = 2, \tag{3.37}$$

for some $\lambda \in \mathbb{R}$. Note that (3.37) yields $\|u\|^2 \geqslant 2 > 0$. Hence $u \neq 0$.

(i): $u_0 = 0$. Then (3.35) yields $(1+\lambda)u = 0$. Because $u \neq 0$, we have $1+\lambda = 0$, i.e., $\lambda = -1$. By (3.36), $2v = v_0 \Rightarrow v = \frac{v_0}{2}$ and then by (3.37),

$$\|u\|^2 - \|\frac{v_0}{2}\|^2 = 2 \Rightarrow \|u\|^2 = 2 + \frac{\|v_0\|^2}{4}.$$

So, $(u,v) = \left(u, \frac{v_0}{2}\right)$ and $\|u\|^2 = 2 + \frac{\|v_0\|^2}{4}$. This gives the formula (3.32), and we also have

$$\begin{aligned} f(u,v) &= \|u-u_0\|^2 + \|v-v_0\|^2 = \|u\|^2 + \left\|\frac{v_0}{2}\right\|^2 \\ &= 2 + \frac{\|v_0\|^2}{4} + \frac{\|v_0\|^2}{4} \\ &= 2 + \frac{\|v_0\|^2}{2}. \end{aligned}$$

(ii): $u_0 \neq 0$, but $v_0 = 0$, which implies $(1-\lambda)v = 0$ by (3.36). We consider two cases:

Case 1: $1-\lambda = 0$, i.e., $\lambda = 1$. By (3.35), $2u = u_0 \Rightarrow u = \frac{u_0}{2}$ and then by (3.37), $\|v\|^2 = \frac{\|u_0\|^2}{4} - 2 \Rightarrow \frac{\|u_0\|^2}{4} - 2 \geqslant 0 \Rightarrow \|u_0\|^2 \geqslant 8$. Thus, $(u,v) = \left(\frac{u_0}{2}, v\right)$ with $\|v\|^2 = \frac{\|u_0\|^2}{4} - 2$, where the objective is

$$f(u,v) = \|u-u_0\|^2 + \|v-v_0\|^2 = \frac{\|u_0\|^2}{4} + \frac{\|u_0\|^2}{4} - 2 = \frac{\|u_0\|^2}{2} - 2.$$

Note that Case 1 needs $\|u_0\| \geqslant 2\sqrt{2}$.

Case 2: $v = 0$. Then $\|u\|^2 = 2 \Rightarrow \|u\| = \sqrt{2}$ by (3.37). By (3.35), $|1+\lambda|\|u\| = \|u_0\| \Rightarrow |1+\lambda| = \frac{\|u_0\|}{\sqrt{2}} \Rightarrow 1+\lambda = \pm\frac{\|u_0\|}{\sqrt{2}}$.

Subcase 1: $1+\lambda = -\frac{\|u_0\|}{\sqrt{2}}$. By (3.35), $u = \frac{u_0}{(-\frac{\|u_0\|}{\sqrt{2}})} = -\frac{\sqrt{2}u_0}{\|u_0\|}$. Then

$$f(u,0) = \|u-u_0\|^2 + \|0-v_0\|^2 = \left\|-\frac{\sqrt{2}u_0}{\|u_0\|} - u_0\right\|^2 = (\|u_0\| + \sqrt{2})^2.$$

Subcase 2: $1+\lambda = \frac{\|u_0\|}{\sqrt{2}}$. By (3.35), $u = \frac{\sqrt{2}u_0}{\|u_0\|}$. Then

$$f(u,0) = \|u-u_0\|^2 + \|0-v_0\|^2 = \left\|u_0 - \frac{\sqrt{2}u_0}{\|u_0\|}\right\|^2 = (\|u_0\| - \sqrt{2})^2.$$

Comparing Subcase 1 and Subcase 2, we obtain $|\|u_0\| - \sqrt{2}| < \|u_0\| + \sqrt{2}$ because $\|u_0\| \neq 0$. That is, Subcase 2 gives a smaller value at $(u,0)$ with $u = \sqrt{2}\frac{u_0}{\|u_0\|}$. We need to compare it to Case 1 whenever it happens. Note that

$$\frac{\|u_0\|^2}{2} - 2 < (\|u_0\| - \sqrt{2})^2 \quad \text{whenever} \quad \|u_0\| \neq 2\sqrt{2}.$$

Indeed, we have

$$\begin{aligned}
&\frac{\|u_0\|^2}{2} - 2 < \|u_0\|^2 - 2\sqrt{2}\|u_0\| + 2 \\
\iff &\frac{\|u_0\|^2}{2} - 2\sqrt{2}\|u_0\| + 4 > 0 \\
\iff &(\|u_0\| - 2\sqrt{2})^2 > 0,
\end{aligned}$$

which always holds if $\|u_0\| \neq 2\sqrt{2}$. Hence, the nearest points are given by

$$\left(\frac{u_0}{2}, v\right) \quad \text{with} \quad \|v\|^2 = \frac{\|u_0\|^2}{4} - 2, \quad \text{when} \quad \|u_0\| \neq 2\sqrt{2} \tag{3.38}$$

$$\left(\frac{u_0}{2}, 0\right) = \left(\sqrt{2}\frac{u_0}{\|u_0\|}, 0\right), \quad \text{when} \quad \|u_0\| = 2\sqrt{2}. \tag{3.39}$$

However, Case 1 occurs only when $\|u_0\| \geqslant 2\sqrt{2}$. Hence, when $\|u_0\| \geqslant 2\sqrt{2}$ the nearest points are given by

$$\left(\frac{u_0}{2}, v\right) \quad \text{with} \quad \|v\|^2 = \frac{\|u_0\|^2}{4} - 2.$$

When $\|u_0\| < 2\sqrt{2}$, Case 1 is impossible. Then we only need to compare Subcase 1 and Subcase 2. Hence, the nearest point is

$$\left(\sqrt{2}\frac{u_0}{\|u_0\|}, 0\right) \quad \text{when} \quad \|u_0\| < 2\sqrt{2}.$$

Finally, since $\|u_0\| = 2\sqrt{2} \Rightarrow \|v\|^2 = 2 - 2 = 0$, (3.38) gives $\left(\frac{u_0}{2}, 0\right)$. Also $\frac{\sqrt{2}}{\|u_0\|} = \frac{1}{2}$, therefore $\sqrt{2}\frac{u_0}{\|u_0\|} = \frac{1}{2}u_0$. It follows that $\left(\frac{u_0}{2}, 0\right) = \left(\sqrt{2}\frac{u_0}{\|u_0\|}, 0\right)$ when $\|u_0\| = 2\sqrt{2}$, which implies that (3.39) can be obtained from (3.38) when $\|u_0\| = 2\sqrt{2}$. Hence formulas (3.33) and (3.34) hold. □

3.2.3 Finding the projection $P_{\tilde{C}_\gamma}$

$P_{\tilde{C}_\gamma}$ can be found via $P_{\tilde{C}_1}$.

Theorem 3.12. *Let $\gamma > 0, \tilde{u}_0, \tilde{v}_0 \in X$, and*

$$\tilde{C}_\gamma = \left\{(\tilde{u}, \tilde{v}) \in X \times X \;\middle|\; \|\tilde{u}\|^2 - \|\tilde{v}\|^2 = 2\gamma\right\}.$$

Then the following hold:

(i) When $\tilde{u}_0 \neq 0$ and $\tilde{v}_0 \neq 0$, we have

$$P_{\tilde{C}_\gamma}(\tilde{u}_0, \tilde{v}_0) = \left\{ \begin{bmatrix} \frac{\tilde{u}_0}{1+\lambda} \\ \frac{\tilde{v}_0}{1-\lambda} \end{bmatrix} \right\},$$

in which λ is the unique root of $H(\lambda) = 0$ in $]-1, 1[$, where

$$H(\lambda) = \frac{(\lambda^2+1)p - 2\lambda q}{2(1-\lambda^2)^2} - \gamma,\ p = \|\tilde{u}_0\|^2 - \|\tilde{v}_0\|^2\,,\ q = \|\tilde{u}_0\|^2 + \|\tilde{v}_0\|^2\,.$$

(ii) When $\tilde{u}_0 = 0$, we have

$$P_{\tilde{C}_\gamma}(0, \tilde{v}_0) = \left\{ \begin{bmatrix} \tilde{u} \\ \frac{\tilde{v}_0}{2} \end{bmatrix} \;\middle|\; \|\tilde{u}\|^2 = 2\gamma + \frac{\|\tilde{v}_0\|^2}{4} \right\}. \tag{3.40}$$

(iii) a) When $\tilde{v}_0 = 0$ and $\|\tilde{u}_0\| \geqslant 2\sqrt{2\gamma}$, we have

$$P_{\tilde{C}_\gamma}(\tilde{u}_0, 0) = \left\{ \begin{bmatrix} \frac{\tilde{u}_0}{2} \\ \tilde{v} \end{bmatrix} \;\middle|\; \|\tilde{v}\|^2 = \frac{\|\tilde{u}_0\|^2}{4} - 2\gamma \right\}. \tag{3.41}$$

b) When $\tilde{v}_0 = 0$ and $0 < \|\tilde{u}_0\| < 2\sqrt{2\gamma}$, we have

$$P_{\tilde{C}_\gamma}(\tilde{u}_0, 0) = \left\{ \begin{bmatrix} \sqrt{2\gamma}\frac{\tilde{u}_0}{\|\tilde{u}_0\|} \\ 0 \end{bmatrix} \right\}. \tag{3.42}$$

Proof. Proposition 3.5(i) states

$$P_{\tilde{C}_\gamma}(\tilde{u}_0, \tilde{v}_0) = \sqrt{\gamma}\, P_{\tilde{C}_1}\left(\frac{\tilde{u}_0}{\sqrt{\gamma}}, \frac{\tilde{v}_0}{\sqrt{\gamma}}\right).$$

Apply Theorem 3.10 and Theorem 3.11. □

3.3 Projections onto bilinear set C_γ when $\gamma > 0$

P_{C_γ} can be found via $P_{\tilde{C}_\gamma}$, which is the main result of this section.

Theorem 3.13. *Let $\gamma > 0$, $x_0, y_0 \in X$, and $C_\gamma = \big\{(x, y) \in X \times X \;\big|\; \langle x, y\rangle = \gamma\big\}$. Then the following hold:*

(i) When $x_0 \neq \pm y_0$, the projection is a singleton:

$$P_{C_\gamma}(x_0, y_0) = \left\{ \begin{bmatrix} \frac{x_0 - \lambda y_0}{1-\lambda^2} \\ \frac{y_0 - \lambda x_0}{1-\lambda^2} \end{bmatrix} \right\},$$

in which λ is the unique solution of $H(\lambda) = 0$ in $]-1, 1[$, where

$$H(\lambda) = \frac{(\lambda^2+1)p - 2\lambda q}{2(1-\lambda^2)^2} - \gamma,\ p = 2\langle x_0, y_0\rangle,\ \text{and } q = \|x_0\|^2 + \|y_0\|^2\,.$$

(ii) When $x_0 = -y_0$, the projection is a set:

$$P_{C_\gamma}(x_0, -x_0) = \left\{ \begin{bmatrix} \frac{x_0}{2} + \frac{\tilde{u}}{\sqrt{2}} \\ -\frac{x_0}{2} + \frac{\tilde{u}}{\sqrt{2}} \end{bmatrix} \,\middle|\, \|\tilde{u}\|^2 = 2\gamma + \frac{\|x_0\|^2}{2}, \quad \tilde{u} \in X \right\}.$$

(iii) a) When $x_0 = y_0$ and $\|x_0\| \geqslant 2\sqrt{\gamma}$, the projection is a set:

$$P_{C_\gamma}(x_0, x_0) = \left\{ \begin{bmatrix} \frac{x_0}{2} - \frac{\tilde{v}}{\sqrt{2}} \\ \frac{x_0}{2} + \frac{\tilde{v}}{\sqrt{2}} \end{bmatrix} \,\middle|\, \|\tilde{v}\|^2 = \frac{\|x_0\|^2}{2} - 2\gamma, \quad \tilde{v} \in X \right\}.$$

b) When $x_0 = y_0$ and $0 < \|x_0\| < 2\sqrt{\gamma}$, the projection is a singleton:

$$P_{C_\gamma}(x_0, x_0) = \left\{ \sqrt{\gamma} \begin{bmatrix} \frac{x_0}{\|x_0\|} \\ \frac{x_0}{\|x_0\|} \end{bmatrix} \right\}.$$

Proof. Proposition 3.5(ii) states

$$P_{C_\gamma}(x_0, y_0) = A_{\frac{\pi}{4}} P_{\tilde{C}_\gamma} A_{-\frac{\pi}{4}}(x_0, y_0). \tag{3.43}$$

It suffices to apply Theorem 3.12. Indeed, with

$$\begin{bmatrix} \tilde{u}_0 \\ \tilde{v}_0 \end{bmatrix} = A_{-\pi/4} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} \frac{x_0+y_0}{\sqrt{2}} \\ \frac{-x_0+y_0}{\sqrt{2}} \end{bmatrix}, \tag{3.44}$$

using (3.43) and (3.44) Theorem 3.12 gives:

(i): When $x_0 \neq \pm y_0$, we have $\tilde{u}_0 \neq 0, \tilde{v}_0 \neq 0$. By Theorem 3.12(i), we get $P_{\tilde{C}_\gamma}(\tilde{u}_0, \tilde{v}_0) = [\frac{\tilde{u}_0}{(1+\lambda)}, \frac{\tilde{v}_0}{(1-\lambda)}]^T$, so using Proposition 3.5(ii) we have

$$\begin{bmatrix} \frac{1}{\sqrt{2}}\,\mathrm{Id} & -\frac{1}{\sqrt{2}}\,\mathrm{Id} \\ \frac{1}{\sqrt{2}}\,\mathrm{Id} & \frac{1}{\sqrt{2}}\,\mathrm{Id} \end{bmatrix} \begin{bmatrix} \frac{1}{1+\lambda}\,\mathrm{Id} & 0 \\ 0 & \frac{1}{1-\lambda}\,\mathrm{Id} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}}\,\mathrm{Id} & \frac{1}{\sqrt{2}}\,\mathrm{Id} \\ -\frac{1}{\sqrt{2}}\,\mathrm{Id} & \frac{1}{\sqrt{2}}\,\mathrm{Id} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{1}{1-\lambda^2}\,\mathrm{Id} & \frac{-\lambda}{1-\lambda^2}\,\mathrm{Id} \\ \frac{-\lambda}{1-\lambda^2}\,\mathrm{Id} & \frac{1}{1-\lambda^2}\,\mathrm{Id} \end{bmatrix},$$

and thus,

$$P_{C_\gamma}(x_0, y_0) = \begin{bmatrix} \frac{1}{1-\lambda^2}\,\mathrm{Id} & \frac{-\lambda}{1-\lambda^2}\,\mathrm{Id} \\ \frac{-\lambda}{1-\lambda^2}\,\mathrm{Id} & \frac{1}{1-\lambda^2}\,\mathrm{Id} \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}.$$

Also, $\|p\| = \|\tilde{u}_0\|^2 - \|\tilde{v}_0\|^2 = 2\langle x_0, y_0\rangle$, and $\|q\| = \|\tilde{u}_0\|^2 + \|\tilde{v}_0\|^2 = \|x_0\|^2 + \|y_0\|^2$.

(ii): For $x_0 = -y_0$, we have $\tilde{u}_0 = 0, \tilde{v}_0 = -\sqrt{2}x_0$, so

$$\begin{aligned} P_{C_\gamma}(x_0, -x_0) &= A_{\frac{\pi}{4}}\, P_{\tilde{C}_\gamma}(0, -\sqrt{2}x_0) \\ &= \left\{ A_{\frac{\pi}{4}} \begin{bmatrix} \tilde{u} \\ \\ \frac{-x_0}{\sqrt{2}} \end{bmatrix} \,\middle|\, \|\tilde{u}\|^2 = 2\gamma + \frac{\|-\sqrt{2}x_0\|^2}{4} = 2\gamma + \frac{\|x_0\|^2}{2}, \tilde{u} \in X \right\}. \end{aligned}$$

(iii): When $x_0 \neq -y_0$, and $x_0 = y_0$, we have $x_0 \neq 0, (x_0, y_0) = (x_0, x_0)$, so that $\tilde{v}_0 = 0$ and $\tilde{u}_0 = \frac{x_0 + y_0}{\sqrt{2}} = \sqrt{2}x_0$. Then $\|\tilde{u}_0\| \geqslant 2\sqrt{2\gamma} \Leftrightarrow \|x_0\| \geqslant 2\sqrt{\gamma}$.

a) When $\|x_0\| \geqslant 2\sqrt{\gamma}$, we have

$$\begin{aligned} P_{C_\gamma}(x_0, x_0) &= A_{\frac{\pi}{4}}\, P_{\tilde{C}_\gamma}(\sqrt{2}x_0, 0) \\ &= A_{\frac{\pi}{4}} \left\{ \begin{bmatrix} \frac{x_0}{\sqrt{2}} \\ \\ \tilde{v} \end{bmatrix} \,\middle|\, \|\tilde{v}\|^2 = \frac{\|\sqrt{2}x_0\|^2}{4} - 2\gamma = \frac{\|x_0\|^2}{2} - 2\gamma, \tilde{v} \in X \right\}. \end{aligned}$$

b) When $0 < \|x_0\| < 2\sqrt{\gamma}$, we have $\|\tilde{u}_0\| < 2\sqrt{2\gamma}$, so

$$P_{C_\gamma}(x_0, x_0) = A_{\frac{\pi}{4}}\, P_{\tilde{C}_\gamma}(\sqrt{2}x_0, 0) = A_{\frac{\pi}{4}} \begin{bmatrix} \sqrt{2\gamma}\frac{\sqrt{2}x_0}{\|\sqrt{2}x_0\|} \\ \\ 0 \end{bmatrix} = A_{\frac{\pi}{4}} \begin{bmatrix} \sqrt{2\gamma}\frac{x_0}{\|x_0\|} \\ \\ 0 \end{bmatrix}.$$

□

3.4 Projections onto $\tilde{C}_\gamma$ and bilinear set C_γ when $\gamma < 0$

Recall that $\tilde{C}_\gamma$ is a hyperbola and C_γ is a bilinear constraint set. Armed with the results in Section 3.2 and Section 3.3, we can study $P_{\tilde{C}_\gamma}$ and P_{C_γ} when $\gamma < 0$. Define

$$T_1 : X \times X \to X \times X : (x, y) \mapsto (y, x), \text{ and}$$

$$T_2 : X \times X \to X \times X : (x, y) \mapsto (x, -y).$$

Theorem 3.14. *Let $\gamma < 0, \tilde{u}_0, \tilde{v}_0 \in X$, and*

$$\tilde{C}_\gamma = \big\{(u, v) \in X \times X \;\big|\; \|u\|^2 - \|v\|^2 = 2\gamma\big\}.$$

Then the following hold:

(i) When $\tilde{u}_0 \neq 0, \tilde{v}_0 \neq 0$, we have

$$P_{\tilde{C}_\gamma}(\tilde{u}_0, \tilde{v}_0) = \left\{ \begin{bmatrix} \frac{\tilde{u}_0}{1-\lambda} \\ \\ \frac{\tilde{v}_0}{1+\lambda} \end{bmatrix} \right\}$$

in which λ is the unique root of $H(\lambda) = 0$ in $]-1,1[$, where

$$H(\lambda) = \frac{(\lambda^2+1)p - 2\lambda q}{2(1-\lambda^2)^2} + \gamma,\ p = \|\tilde{v}_0\|^2 - \|\tilde{u}_0\|^2, \quad \text{and } q = \|\tilde{u}_0\|^2 + \|\tilde{v}_0\|^2.$$

(ii) When $\tilde{v}_0 = 0$, we have

$$P_{\tilde{C}_\gamma}(\tilde{u}_0, 0) = \left\{ \begin{bmatrix} \frac{\tilde{u}_0}{2} \\ \tilde{v} \end{bmatrix} \,\middle|\, \|\tilde{v}\|^2 = \frac{\|\tilde{u}_0\|^2}{4} - 2\gamma \right\}. \tag{3.45}$$

(iii) a) When $\tilde{u}_0 = 0$ and $\|\tilde{v}_0\| \geqslant 2\sqrt{2(-\gamma)}$, we have

$$P_{\tilde{C}_\gamma}(0, \tilde{v}_0) = \left\{ \begin{bmatrix} \tilde{u} \\ \frac{\tilde{v}_0}{2} \end{bmatrix} \,\middle|\, \|\tilde{u}\|^2 = \frac{\|\tilde{v}_0\|^2}{4} + 2\gamma \right\}. \tag{3.46}$$

b) When $\tilde{u}_0 = 0$ and $0 < \|\tilde{v}_0\| < 2\sqrt{2(-\gamma)}$, we have

$$P_{\tilde{C}_\gamma}(0, \tilde{v}_0) = \left\{ \begin{bmatrix} 0 \\ \sqrt{2(-\gamma)}\frac{\tilde{v}_0}{\|\tilde{v}_0\|} \end{bmatrix} \right\}. \tag{3.47}$$

Proof. Since

$$\text{minimize } \|\tilde{u} - \tilde{u}_0\|^2 + \|\tilde{v} - \tilde{v}_0\|^2 \quad \text{subject to} \quad \|\tilde{u}\|^2 - \|\tilde{v}\|^2 = 2\gamma$$

is equivalent to

$$\text{minimize } \|\tilde{v} - \tilde{v}_0\|^2 + \|\tilde{u} - \tilde{u}_0\|^2 \quad \text{subject to} \quad \|\tilde{v}\|^2 - \|\tilde{u}\|^2 = 2(-\gamma),$$

we have $P_{\tilde{C}_\gamma} = T_1 P_{\tilde{C}_{-\gamma}} T_1$. It suffices to apply Theorem 3.12. □

Theorem 3.15. *Let $\gamma < 0$, $x_0, y_0 \in X$, and $C_\gamma = \{(x,y) \in X \times X \mid \langle x, y\rangle = \gamma\}$. Then the following hold:*

(i) When $x_0 \neq \pm y_0$, the projection is a singleton:

$$P_{C_\gamma}(x_0, y_0) = \left\{ \begin{bmatrix} \frac{x_0 + \lambda y_0}{1-\lambda^2} \\ \frac{y_0 + \lambda x_0}{1-\lambda^2} \end{bmatrix} \right\}$$

in which λ is the unique solution of $H(\lambda) = 0$ in $]-1,1[$, where

$$H(\lambda) = \frac{(\lambda^2+1)p - 2\lambda q}{2(1-\lambda^2)^2} + \gamma,\ p = -2\langle x_0, y_0\rangle,\ \text{and } q = \|x_0\|^2 + \|y_0\|^2.$$

(ii) When $x_0 = y_0$, the projection is a set:

$$P_{C_\gamma}(x_0, x_0) = \left\{ \begin{bmatrix} \frac{x_0}{2} + \frac{\tilde{u}}{\sqrt{2}} \\ \frac{x_0}{2} - \frac{\tilde{u}}{\sqrt{2}} \end{bmatrix} \,\middle|\, \|\tilde{u}\|^2 = -2\gamma + \frac{\|x_0\|^2}{2}, \quad \tilde{u} \in X \right\}.$$

(iii) a) When $x_0 = -y_0$ and $\|x_0\| \geqslant 2\sqrt{-\gamma}$, the projection is a set:

$$P_{C_\gamma}(x_0, -x_0) = \left\{ \begin{bmatrix} \frac{x_0}{2} - \frac{\tilde{v}}{\sqrt{2}} \\ \\ \frac{-x_0}{2} - \frac{\tilde{v}}{\sqrt{2}} \end{bmatrix} \;\middle|\; \|\tilde{v}\|^2 = \frac{\|x_0\|^2}{2} + 2\gamma, \quad \tilde{v} \in X \right\}.$$

b) When $x_0 = -y_0$ and $0 < \|x_0\| < 2\sqrt{-\gamma}$, the projection is a singleton:

$$P_{C_1}(x_0, -x_0) = \left\{ \sqrt{-\gamma} \begin{bmatrix} \frac{x_0}{\|x_0\|} \\ \\ \frac{-x_0}{\|x_0\|} \end{bmatrix} \right\}.$$

Proof. Since

$$\text{minimize } \; \|x - x_0\|^2 + \|y - y_0\|^2 \quad \text{subject to} \quad \langle x, y\rangle = \gamma$$

is equivalent to

$$\text{minimize } \; \|x - x_0\|^2 + \|z - (-y_0)\|^2 \quad \text{subject to} \quad \langle x, z\rangle = -\gamma,$$

we have $P_{C_\gamma} = T_2 P_{C_{-\gamma}} T_2$. It remains to apply Theorem 3.13. □

Chapter 4

Projecting onto rectangular hyperbolic paraboloids

This chapter is based on the paper [13] titled "Projecting onto rectangular hyperbolic paraboloids in Hilbert space" that has been accepted in Applied Set-Valued Analysis and Optimization, 2023. In $\mathbb{R}^3$, a hyperbolic paraboloid is a classical saddle-shaped quadric surface. Recently, Elser has modeled problems arising in Deep Learning [40, Section 5 and 6] using rectangular hyperbolic paraboloids in $\mathbb{R}^3$. We provide a rigorous analysis of the associated projection. In some cases, finding this projection amounts to finding a certain root of a quintic or cubic polynomial. We also observe when the projection is not a singleton and point out connections to graphical and set convergence.

Consider $\alpha \in \mathbb{R} \smallsetminus \{0\}$ and $\beta > 0$. Define the β-weighted norm on the product space $X \times X \times \mathbb{R}$ by

$$(\forall (x, y, \gamma) \in X \times X \times \mathbb{R}) \ \|(x, y, \gamma)\| := \sqrt{\|x\|^2 + \|y\|^2 + \beta^2 |\gamma|^2}.$$

Now define the set

$$C_\alpha := \big\{(x, y, \gamma) \in X \times X \times \mathbb{R} \ \big| \ \langle x, y \rangle = \alpha\gamma\big\}. \tag{4.1}$$

The set C_α is a special bilinear constraint set with applications in optimization, and it corresponds to a rectangular (a.k.a. orthogonal) hyperbolic paraboloid in geometry [54]. Motivated by Deep Learning, Elser recently presented in [40] a formula for the projection $P_{C_\alpha}(x_0, y_0, \gamma_0)$ when $x_0 \neq \pm y_0$. However, complete mathematical justifications were not presented, and the case when $x_0 = \pm y_0$ was not considered. The goal of this chapter is to provide a complete analysis of P_{C_α} that applies to all possible cases.

The chapter is organized as follows. We collect auxiliary results in Section 4.1. Our main result is proved in Section 4.2 and Section 4.3 contains a numerical illustration. The formula for the projection onto the set C_α is presented in Section 4.4.

As usual, the distance function and projection mapping associated to C_α are denoted by

$$d_{C_\alpha}(x_0, y_0, \gamma_0) := \inf_{(x,y,\gamma) \in C_\alpha} \|(x, y, \gamma) - (x_0, y_0, \gamma_0)\|$$

and

$$P_{C_\alpha}(x_0, y_0, \gamma_0) := \operatorname{argmin}_{(x,y,\gamma) \in C_\alpha} \|(x, y, \gamma) - (x_0, y_0, \gamma_0)\|,$$

respectively. We say that $x, x_0 \in X$ are *conically dependent* if there exists $s \geqslant 0$ such that $x = sx_0$ or $x_0 = sx$.

4.1 Properties of C_α & projection in Hilbert space

We start with some elementary properties of C_α, and justify the existence of projections onto these sets.

Proposition 4.1. *The following hold:*

(i) The set C_α is closed. If X is infinite-dimensional, then C_α is not weakly closed; in fact, $\overline{C_\alpha}^{\text{weak}} = X \times X \times \mathbb{R}$.

(ii) C_α is prox-regular in $X \times X \times \mathbb{R}$. Hence, for every point $(x_0, y_0, \gamma_0) \in C_\alpha$, there exists a neighborhood such that the projection mapping P_{C_α} is single-valued.

Proof. (i): Clearly, C_α is closed. Thus assume that X is infinite-dimensional. By [14, Proposition 2.1], for every $\gamma \in \mathbb{R}$, $\overline{\{(x,y) \in X \times X |\ \langle x, y\rangle = \alpha\gamma\}}^{\text{weak}} = X \times X$. Thus,

$$\begin{aligned} X \times X \times \mathbb{R} &= \bigcup_{\gamma \in \mathbb{R}} \left(\overline{\{(x,y) \in X \times X |\ \langle x, y\rangle = \alpha\gamma\}}^{\text{weak}} \times \{\gamma\} \right) \\ &\subseteq \overline{\{(x,y,\gamma) \in X \times X \times \mathbb{R} |\ \langle x, y\rangle = \alpha\gamma\}}^{\text{weak}} \subseteq X \times X \times \mathbb{R}. \end{aligned}$$

(ii): Set $F \colon X \times X \times \mathbb{R} \to \mathbb{R} \colon (x,y,\gamma) \mapsto \langle x, y\rangle - \alpha\gamma$. Then $C_\alpha = F^{-1}(0)$ and $\nabla F(x,y,\gamma) = (y, x, -\alpha) \neq (0,0,0)$ because $\alpha \neq 0$. The prox-regularity of C_α now follows from [62, Example 6.8] when $X = \mathbb{R}^n$ or from [17, Proposition 2.4] in the general case. Finally, the single-valuedness of the projection locally around every point in C_α follows from [17, Proposition 4.4]. ☐

To study the projection onto C_α, it is convenient to introduce

$$\widetilde{C}_\alpha := \big\{(u,v,\gamma) \in X \times X \times \mathbb{R} \ \big|\ \|u\|^2 - \|v\|^2 = 2\alpha\gamma\big\}, \tag{4.2}$$

which is the standard form of a rectangular hyperbolic paraboloid. Define a linear operator $A : X \times X \times \mathbb{R} \to X \times X \times \mathbb{R}$ by sending (u, v, γ) to (x, y, γ), where

$$x = \frac{u - v}{\sqrt{2}} \quad \text{and} \quad y = \frac{u + v}{\sqrt{2}}.$$

In terms of block matrix notation, we have

$$\begin{bmatrix} x \\ y \\ \gamma \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}}\,\mathrm{Id} & -\frac{1}{\sqrt{2}}\,\mathrm{Id} & 0 \\ \frac{1}{\sqrt{2}}\,\mathrm{Id} & \frac{1}{\sqrt{2}}\,\mathrm{Id} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ \gamma \end{bmatrix} \Leftrightarrow \begin{bmatrix} u \\ v \\ \gamma \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}}\,\mathrm{Id} & \frac{1}{\sqrt{2}}\,\mathrm{Id} & 0 \\ -\frac{1}{\sqrt{2}}\,\mathrm{Id} & \frac{1}{\sqrt{2}}\,\mathrm{Id} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ \gamma \end{bmatrix}.$$

Thus, we may and do identify A with its block matrix representation

$$A = \begin{bmatrix} \frac{1}{\sqrt{2}}\,\mathrm{Id} & -\frac{1}{\sqrt{2}}\,\mathrm{Id} & 0 \\ \frac{1}{\sqrt{2}}\,\mathrm{Id} & \frac{1}{\sqrt{2}}\,\mathrm{Id} & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and we denote the adjoint of A by $A^\intercal$. Note that A corresponds to a rotation by $\pi/4$ about the γ-axis. The relationship between C_α and $\widetilde{C}_\alpha$ is summarized as follows.

Proposition 4.2. *The following hold:*

(i) A is a surjective isometry (i.e., a unitary operator): $AA^\intercal = A^\intercal A = \mathrm{Id}$.

(ii) $A\widetilde{C}_\alpha = C_\alpha$ and $\widetilde{C}_\alpha = A^\intercal C_\alpha$.

(iii) $P_{C_\alpha} = A\,P_{\widetilde{C}_\alpha} A^\intercal$.

Proof. It is straightforward to verify (i) and (ii). To show (iii), let $(x_0, y_0, \gamma_0) \in X \times X \times \mathbb{R}$. In view of (i) and (ii), we have $(x, y, \gamma) \in P_{C_\alpha}(x_0, y_0, \gamma_0)$ if and only if $(x, y, \gamma) \in C_\alpha$ and

$$\begin{aligned}\|(x, y, \gamma) - (x_0, y_0, \gamma_0)\| = d_{C_\alpha}(x_0, y_0, \gamma_0) = d_{A\widetilde{C}_\alpha}(x_0, y_0, \gamma_0) \\ = d_{\widetilde{C}_\alpha}(A^\intercal[x_0, y_0, \gamma_0]^\intercal),\end{aligned}$$

and this is equivalent to

$$\|A^\intercal[x, y, \gamma]^\intercal - A^\intercal[x_0, y_0, \gamma_0]^\intercal\| = d_{\widetilde{C}_\alpha}(A^\intercal[x_0, y_0, \gamma_0]^\intercal).$$

Since $A^\intercal[x, y, \gamma]^\intercal \in \widetilde{C}_\alpha$, this gives $A^\intercal[x, y, \gamma]^\intercal \in P_{\widetilde{C}_\alpha}(A^\intercal[x_0, y_0, \gamma_0]^\intercal)$, i.e., $[x, y, \gamma]^\intercal \in A\,P_{\widetilde{C}_\alpha}(A^\intercal[x_0, y_0, \gamma_0]^\intercal)$. The converse inclusion is proved similarly. □

Exploiting the structure of $\widetilde{C}_\alpha$ is crucial for showing the existence of

$$P_{\widetilde{C}_\alpha}(u_0, v_0, \gamma_0)$$

for every $(u_0, v_0, \gamma_0) \in X \times X \times \mathbb{R}$.

Proposition 4.3. (existence of the projection) *Let $(u_0, v_0, \gamma_0) \in X \times X \times \mathbb{R}$. Then the minimization problem*

$$\begin{aligned} &\textit{minimize} && f(u, v, \gamma) := \|u - u_0\|^2 + \|v - v_0\|^2 + \beta^2|\gamma - \gamma_0|^2 && \text{(4.3a)} \\ &\textit{subject to} && h(u, v, \gamma) := \|u\|^2 - \|v\|^2 - 2\alpha\gamma = 0 && \text{(4.3b)} \end{aligned}$$

always has a solution, i.e., $P_{C_\alpha}(u_0, v_0, \gamma_0) \neq \varnothing$. If $(u, v, \gamma) \in P_{C_\alpha}(u_0, v_0, \gamma_0)$, then u, u_0 are conically dependent, and v, v_0 are also conically dependent.

Proof. We only illustrate the case when $u_0 \neq 0, v_0 \neq 0$, since the other cases are similar. We claim that the optimization problem is essentially 3-dimensional. To this end, we expand

$$f(u,v,\gamma) = \underbrace{\|u\|^2 - 2\langle u, u_0\rangle + \|u_0\|^2}_{} + \underbrace{\|v\|^2 - 2\langle v, v_0\rangle + \|v_0\|^2}_{} + \beta^2|\gamma - \gamma_0|^2. \tag{4.4}$$

The constraint

$$h(u,v,\gamma) = \|u\|^2 - \|v\|^2 - 2\alpha\gamma = 0$$

means that for the variables u, v only the norms $\|u\|$ and $\|v\|$ matter. With $\|u\|$ fixed, the Cauchy-Schwarz inequality in Hilbert space (see, e.g., [47]), shows that $-2\langle u, u_0\rangle$ in the left underbraced part of (4.4) will be smallest when u, u_0 are conically dependent. Similarly, for fixed $\|v\|$, the second underlined part in f will be smaller when $v = tv_0$ for some $t \geqslant 0$. It follows that the optimization problem given by (4.3) is equivalent to

$$\text{minimize} \quad g(s,t,\gamma) := (1-s)^2\|u_0\|^2 + (1-t)^2\|v_0\|^2 + \beta^2|\gamma - \gamma_0|^2 \tag{4.5a}$$

$$\text{subject to} \quad g_1(s,t,\gamma) := s^2\|u_0\|^2 - t^2\|v_0\|^2 - 2\alpha\gamma = 0, \quad s \geqslant 0, t \geqslant 0, \gamma \in \mathbb{R}. \tag{4.5b}$$

Because g is continuous and coercive, and g_1 is continuous, we conclude that the optimization problem (4.5) has a solution. □

Next we provide a result on set convergence and review graphical convergence, see, e.g., [3, 62]. We shall need the *cross*

$$C := \big\{(x,y) \in X \times X \;\big|\; \langle x, y\rangle = 0\big\}, \tag{4.6}$$

which was studied in, e.g., [12], as well as

$$\tilde{C} := \big\{(u,v) \in X \times X \;\big|\; \|u\|^2 - \|v\|^2 = 0\big\}. \tag{4.7}$$

Proposition 4.4. *The following hold:*

(i) $\lim_{\alpha\to 0} \widetilde{C}_\alpha = \tilde{C} \times \mathbb{R}$.

(ii) $\lim_{\alpha\to 0} C_\alpha = C \times \mathbb{R}$.

Proof. (i): First we show that $\limsup_{\alpha\to 0} \widetilde{C}_\alpha \subseteq \tilde{C} \times \mathbb{R}$. Let $(u_\alpha, v_\alpha, \gamma_\alpha) \to (u, v, \gamma)$ and $(u_\alpha, v_\alpha, \gamma_\alpha) \in \widetilde{C}_\alpha$ with $\alpha \to 0$. Then $\|u_\alpha\|^2 - \|v_\alpha\|^2 = 2\alpha\gamma_\alpha$ gives $\|u\|^2 - \|v\|^2 = 0$ when $\alpha \to 0$, so $(u, v, \gamma) \in \tilde{C} \times \mathbb{R}$.

Next we show $\tilde{C} \times \mathbb{R} \subseteq \liminf_{\alpha\to 0} \widetilde{C}_\alpha$. Let $(u, v, \gamma) \in \tilde{C} \times \mathbb{R}$, i.e., $\|u\|^2 - \|v\|^2 = 0$ and $\gamma \in \mathbb{R}$. Let $\varepsilon > 0$. We consider three cases:

Case 1: $\gamma = 0$. Then $(u_\alpha, v_\alpha, 0) = (u, v, 0) \in \widetilde{C}_\alpha$ for every α.

Case 2: $\gamma \neq 0$ but $(u, v) = (0, 0)$. If $\alpha\gamma > 0$, take $(u_\alpha, 0, \gamma)$ with $\|u_\alpha\|^2 - 0 = \alpha\gamma$ so that $(u_\alpha, 0, \gamma) \in C_\alpha$; if $\alpha\gamma < 0$, take $(0, v_\alpha, \gamma)$ with $0 - \|v_\alpha\|^2 = \alpha\gamma$ so that $(0, v_\alpha, \gamma) \in C_\alpha$. Then

$$\|(u_\alpha, 0, \gamma) - (0, 0, \gamma)\| = \|u_\alpha\| = \sqrt{|\alpha\gamma|} < \varepsilon,$$

or

$$\|(0, v_\alpha, \gamma) - (0, 0, \gamma)\| = \|v_\alpha\| = \sqrt{|\alpha\gamma|} < \varepsilon,$$

if $|\alpha| < \varepsilon^2/|\gamma|$.

Case 3: $\gamma \neq 0$ and $(u, v) \neq (0, 0)$. Take $\alpha \in \mathbb{R}$ such that

$$|\alpha| < \min\left\{\frac{\varepsilon\|(u,v)\|}{|\gamma|}, \frac{\|(u,v)\|^2}{|\gamma|}\right\},$$

and set

$$\lambda := \frac{\alpha\gamma}{\|(u,v)\|^2}.$$

Then

$$|\lambda| = \frac{|\alpha\gamma|}{\|(u,v)\|^2} < 1.$$

Now set

$$u_\alpha := \sqrt{1+\lambda}u, \quad v_\alpha := \sqrt{1-\lambda}v.$$

Then

$$\begin{aligned}\|u_\alpha\|^2 - \|v_\alpha\|^2 &= (1+\lambda)\|u\|^2 - (1-\lambda)\|v\|^2\\ &= \lambda(\|u\|^2 + \|v\|^2) = \alpha\gamma,\end{aligned}$$

so that $(u_\alpha, v_\alpha, \gamma) \in \widetilde{C}_\alpha$ and

$$\begin{aligned}\|(u_\alpha, v_\alpha, \gamma) - (u, v, \gamma)\| &= \sqrt{(\sqrt{1+\lambda}-1)^2\|u\|^2 + (\sqrt{1-\lambda}-1)^2\|v\|^2}\\ &= \sqrt{\frac{\lambda^2}{(1+\sqrt{1+\lambda})^2}\|u\|^2 + \frac{\lambda^2}{(1+\sqrt{1-\lambda})^2}\|v\|^2}\\ &\leqslant \sqrt{\lambda^2(\|u\|^2+\|v\|^2)} = |\lambda|\|(u,v)\| < \varepsilon.\end{aligned}$$

(ii): This follows from (i) because that $C_\alpha = A\widetilde{C}_\alpha$ and $C \times \mathbb{R} = A(\tilde{C} \times \mathbb{R})$ and that A is an isometry. See also [62, Theorem 4.26]. □

Definition 4.5. (graphical limits of mappings) (See [62, Definition 5.32].) For a sequence of set-valued mappings $S^k : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, we say S^k converges graphically to S, in symbols $S^k \xrightarrow{g} S$, if for every $x \in \mathbb{R}^n$ one has

$$\bigcup_{\{x^k \to x\}} \limsup_{k\to\infty} S^k(x^k) \subseteq S(x) \subseteq \bigcup_{\{x^k \to x\}} \liminf_{k\to\infty} S^k(x^k).$$

Fact 4.6. (Rockafellar–Wets) *(See [62, Example 5.35].) For closed subsets sets S^k, S of $\mathbb{R}^n$, one has $P_{S^k} \xrightarrow{g} P_S$ if and only if $S^k \to S$.*

4.2 Projection onto a hyperbolic paraboloid

Recall that we work with rectangular hyperbolic paraboloids. So, we begin with projections onto rectangular hyperbolic paraboloids. In view of Proposition 4.2(iii), to find P_{C_α} it suffices to find $P_{\widetilde{C}_\alpha}$. That is, for every $(u_0, v_0, \gamma_0) \in X \times X \times \mathbb{R}$, we need to solve:

$$\min_{u,v,\gamma} \quad f(u,v,\gamma) := \|u-u_0\|^2 + \|v-v_0\|^2 + \beta^2|\gamma-\gamma_0|^2 \tag{4.8a}$$

$$\text{subject to} \quad h(u,v,\gamma) := \|u\|^2 - \|v\|^2 - 2\alpha\gamma = 0. \tag{4.8b}$$

Theorem 4.7. *Let $(u_0, v_0, \gamma_0) \in X \times X \times \mathbb{R}$. Then the following hold:*

(i) When $u_0 \neq 0, v_0 \neq 0$, then

$$P_{\widetilde{C}_\alpha}(u_0, v_0, \gamma_0) = \left\{ \Big(\frac{u_0}{1+\lambda}, \frac{v_0}{1-\lambda}, \gamma_0 + \frac{\lambda\alpha}{\beta^2} \Big) \right\}, \tag{4.9}$$

where the unique $\lambda \in \,]-1,1[$ solves the following (essentially) quintic equation

$$g(\lambda) := \frac{(\lambda^2+1)p - 2\lambda q}{(1-\lambda^2)^2} - \frac{2\lambda\alpha^2}{\beta^2} - 2\alpha\gamma_0 = 0, \tag{4.10}$$

and where $p := \|u_0\|^2 - \|v_0\|^2$ and $q := \|u_0\|^2 + \|v_0\|^2$.

(ii) When $u_0 = 0, v_0 \neq 0$, we have:

(a) If $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) < -\frac{\|v_0\|^2}{8}$, then

$$P_{\widetilde{C}_\alpha}(0, v_0, \gamma_0) = \left\{ \Big(0, \frac{v_0}{1-\lambda}, \gamma_0 + \frac{\lambda\alpha}{\beta^2} \Big) \right\}, \tag{4.11}$$

for a unique $\lambda \in \,]-1,1[$ that solves the (essentially) cubic equation

$$g_1(\lambda) := \frac{\|v_0\|^2}{(1-\lambda)^2} + \frac{2\lambda\alpha^2}{\beta^2} + 2\alpha\gamma_0 = 0. \tag{4.12}$$

(b) If $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) \geqslant -\frac{\|v_0\|^2}{8}$, then

$$P_{\widetilde{C}_\alpha}(0, v_0, \gamma_0) = \left\{ \Big(u, \frac{v_0}{2}, \gamma_0 - \frac{\alpha}{\beta^2} \Big) \;\middle|\; \|u\| = \sqrt{2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big) + \frac{\|v_0\|^2}{4}}, u \in X \right\}, \tag{4.13}$$

which is a singleton if and only if $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) = -\frac{\|v_0\|^2}{8}$.

(iii) When $u_0 \neq 0, v_0 = 0$, we have:

(a) If $\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) > \frac{\|u_0\|^2}{8}$*, then*

$$P_{\widetilde{C}_\alpha}(u_0, 0, \gamma_0) = \Big\{\Big(\frac{u_0}{1+\lambda}, 0, \gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big)\Big\} \tag{4.14}$$

for a unique $\lambda \in\,]-1, 1[$ *that solves the (essentially) cubic equation*

$$g_2(\lambda) := \frac{\|u_0\|^2}{(1+\lambda)^2} - \frac{2\lambda\alpha^2}{\beta^2} - 2\alpha\gamma_0 = 0. \tag{4.15}$$

(b) If $\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) \leqslant \frac{\|u_0\|^2}{8}$*, then*

$$P_{\widetilde{C}_\alpha}(u_0, 0, \gamma_0) = \Big\{\Big(\frac{u_0}{2}, v, \gamma_0 + \frac{\alpha}{\beta^2}\Big) \;\Big|\; \|v\| = \sqrt{-2\alpha\Big(\gamma_0 + \frac{\alpha}{\beta^2}\Big) + \frac{\|u_0\|^2}{4}}, v \in X\Big\}, \tag{4.16}$$

which is a singleton if and only if $\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) = \frac{\|u_0\|^2}{8}$.

(iv) When $u_0 = 0, v_0 = 0$*, we have:*

(a) If $\alpha\gamma_0 > \frac{\alpha^2}{\beta^2}$*, then the projection is the non-singleton set*

$$P_{\widetilde{C}_\alpha}(0, 0, \gamma_0) = \Big\{\Big(u, 0, \gamma_0 - \frac{\alpha}{\beta^2}\Big) \;\Big|\; \|u\| = \sqrt{2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big)},\ u \in X\Big\}. \tag{4.17}$$

(b) If $|\alpha\gamma_0| \leqslant \frac{\alpha^2}{\beta^2}$*, then*

$$P_{\widetilde{C}_\alpha}(0, 0, \gamma_0) = \big\{(0, 0, 0)\big\}. \tag{4.18}$$

(c) If $\alpha\gamma_0 < -\frac{\alpha^2}{\beta^2}$*, then the projection is the non-singleton set*

$$P_{\widetilde{C}_\alpha}(0, 0, \gamma_0) = \Big\{\Big(0, v, \gamma_0 + \frac{\alpha}{\beta^2}\Big) \;\Big|\; \|v\| = \sqrt{-2\alpha\Big(\gamma_0 + \frac{\alpha}{\beta^2}\Big)},\ v \in X\Big\}. \tag{4.19}$$

Proof. Observe that $\nabla f(u, v, \gamma) = (2(u - u_0), 2(v - v_0), 2\beta^2(\gamma - \gamma_0))$ and

$$\nabla h(u, v, \gamma) = (2u, -2v, -2\alpha).$$

Since $\alpha \neq 0$, we have $\forall (u, v, \gamma) \in X \times X \times \mathbb{R}$, $\nabla h(u, v, \gamma) \neq 0$. Using [18, Proposition 4.1.1], we obtain the following KKT optimality conditions of (4.8):

$$(1+\lambda)u = u_0, \tag{4.20a}$$
$$(1-\lambda)v = v_0, \tag{4.20b}$$
$$\beta^2(\gamma-\gamma_0) - \lambda\alpha = 0, \tag{4.20c}$$
$$\|u\|^2 - \|v\|^2 - 2\alpha\gamma = 0, \tag{4.20d}$$

where $\lambda \in \mathbb{R}$ is the Lagrange multiplier. The proofs of (i)–(iv) are presented below.

4.2.1 Case (i): $u_0 \neq 0, v_0 \neq 0$

Proof. Because $u_0 \neq 0, v_0 \neq 0$, we obtain $\lambda \neq \pm 1$. Solving (4.20a), (4.20b) and (4.20c) gives $u = \frac{u_0}{(1+\lambda)}$, $v = \frac{v_0}{(1-\lambda)}$ and $\gamma = \gamma_0 + \frac{\lambda\alpha}{\beta^2}$. By Proposition 4.3, $1+\lambda > 0$ and $1-\lambda > 0$, i.e., $\lambda \in \,]-1,1[$. Substituting u and v back into equation (4.20d), we get the (essentially) quintic equation (4.10). Using also $p < q$ and $q > 0$, we have

$$\begin{aligned}
(\forall \lambda \in \,]-1,1[)\; g'(\lambda) &= \frac{2}{(1-\lambda^2)^3}\big(-q(1+3\lambda^2)+p(\lambda^3+3\lambda)\big) - 2\frac{\alpha^2}{\beta^2}\\
&< \frac{2}{(1-\lambda^2)^3}\big(-q(1+3\lambda^2)+q(\lambda^3+3\lambda)\big) - 2\frac{\alpha^2}{\beta^2}\\
&= \frac{2q(\lambda-1)^3}{(1-\lambda^2)^3} - 2\frac{\alpha^2}{\beta^2} = \frac{-2q}{(1+\lambda)^3} - 2\frac{\alpha^2}{\beta^2}\\
&< 0;
\end{aligned}$$

hence, g is strictly decreasing. Moreover, $g(-1) = +\infty$, $g(1) = -\infty$ and g is continuous on $]-1,1[$. Thus, $g(\lambda) = 0$ has unique zero in $]-1,1[$. □

4.2.2 Case (ii): $u_0 = 0, v_0 \neq 0$

Proof. When $u_0 = 0$, the objective function is

$$f(u,v,\gamma) = \|u\|^2 + \|v-v_0\|^2 + \beta^2|\gamma-\gamma_0|^2,$$

and the KKT optimality conditions (4.20) become

$$(1+\lambda)u = 0, \tag{4.21a}$$
$$(1-\lambda)v = v_0, \tag{4.21b}$$
$$\gamma = \gamma_0 + \frac{\lambda\alpha}{\beta^2}, \tag{4.21c}$$
$$\|u\|^2 - \|v\|^2 = 2\alpha\gamma. \tag{4.21d}$$

Then (4.21a) gives

$$1+\lambda = 0 \;\text{ or }\; u = 0. \tag{4.22}$$

Because $v_0 \neq 0$, we have $1 - \lambda \neq 0$, so that

$$v = \frac{v_0}{1-\lambda}. \tag{4.23}$$

By Proposition 4.3, $\lambda < 1$.

Our analysis is divided into the following three situations:

Situation 1: $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) < -\frac{\|v_0\|^2}{8}$**.**

In view of (4.22), we analyze two cases.

Case 1: $1 + \lambda = 0$, i.e., $\lambda = -1$. By (4.23), $v = \frac{v_0}{2}$, and then (4.21d) and (4.21c) give

$$\|u\|^2 = 2\alpha\gamma + \frac{\|v_0\|^2}{4} = 2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big) + \frac{\|v_0\|^2}{4} < 0,$$

which is absurd.

Case 2: $u = 0$. By (4.21d), $-\|v\|^2 = 2\alpha\gamma$, together with (4.23) and (4.21c), we have

$$g_1(\lambda) := \frac{\|v_0\|^2}{(1-\lambda)^2} + 2\alpha\Big(\gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big) = 0.$$

As

$$g_1'(\lambda) = \frac{2\|v_0\|^2}{(1-\lambda)^3} + \frac{2\alpha^2}{\beta^2} > 0 \quad \text{on} \quad]-\infty, 1[\,,$$

g_1 is strictly increasing on $]-\infty, 1[$. Moreover, $g_1(1) = +\infty$ and

$$g_1(-1) = \frac{\|v_0\|^2}{4} + 2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big) < 0.$$

Because g_1 is strictly increasing and continuous, by the Intermediate Value Theorem, there exists a unique $\lambda \in]-1, 1[$ such that $g_1(\lambda) = 0$. Hence, the possible optimal solution is given by

$$\Big(0, \frac{v_0}{1-\lambda}, \gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big), \tag{4.24}$$

where $g_1(\lambda) = 0$ and $\lambda \in]-1, 1[$.

Combining Case 1 and Case 2, we obtain that (4.24) is the unique projection.

Situation 2: $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) > -\frac{\|v_0\|^2}{8}$**.**

In view of (4.22), we consider two cases:

Case 1: $1 + \lambda = 0$, i.e., $\lambda = -1$. By (4.23), $v = \frac{v_0}{2}$, and then (4.21d) and (4.21c) give

$$\|u\|^2 = 2\alpha\gamma + \frac{\|v_0\|^2}{4} = 2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big) + \frac{\|v_0\|^2}{4} > 0.$$

The possible optimal value is attained at

$$\Big(u, \frac{v_0}{2}, \gamma_0 - \frac{\alpha}{\beta^2}\Big) \tag{4.25}$$

with $\|u\|^2 = 2\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) + \frac{\|v_0\|^2}{4}$ such that

$$f\Big(u, \frac{v_0}{2}, \gamma_0 - \frac{\alpha}{\beta^2}\Big) = 2\alpha\gamma_0 - \frac{\alpha^2}{\beta^2} + \frac{\|v_0\|^2}{2}. \tag{4.26}$$

Case 2: $u = 0$. By (4.21d), $-\|v\|^2 = 2\alpha\gamma$, together with (4.21c), we have

$$g_1(\lambda) := \frac{\|v_0\|^2}{(1-\lambda)^2} + 2\alpha\Big(\gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big) = 0.$$

As

$$g_1'(\lambda) = \frac{2\|v_0\|^2}{(1-\lambda)^3} + \frac{2\alpha^2}{\beta^2} > 0 \quad \text{on } \left]-\infty, 1\right[,$$

g_1 is strictly increasing. Observe that

$$g_1(-1) = \frac{\|v_0\|^2}{4} + 2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big) > 0,$$

and $g_1(-\infty) = -\infty$. By the Intermediate Value Theorem, there exists a unique $\lambda \in \left]-\infty, -1\right[$ such that $g_1(\lambda) = 0$ because g_1 is strictly increasing and continuous. The possible optimal value is attained at (recall (4.23))

$$\Big(0, \frac{v_0}{1-\lambda}, \gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big)$$

with

$$f\Big(0, \frac{v_0}{1-\lambda}, \gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big) = \frac{\lambda^2\|v_0\|^2}{(1-\lambda)^2} + \frac{\lambda^2\alpha^2}{\beta^2}, \tag{4.27}$$

where λ is the unique solution of

$$g_1(\lambda) := \frac{\|v_0\|^2}{(1-\lambda)^2} + 2\alpha\Big(\gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big) = 0 \quad \text{in } \left]-\infty, -1\right[. \tag{4.28}$$

Because both Case 1 and Case 2 may occur, we have to compare possible optimal objective function values, namely, (4.26) and (4.27). We claim that Case 1 wins, i.e.,

$$2\alpha\gamma_0 - \frac{\alpha^2}{\beta^2} + \frac{\|v_0\|^2}{2} < \frac{\lambda^2\|v_0\|^2}{(1-\lambda)^2} + \frac{\lambda^2\alpha^2}{\beta^2}. \tag{4.29}$$

In view of (4.28), we have

$$0 < \frac{\|v_0\|^2}{(1-\lambda)^2} = -2\alpha\Big(\gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big), \text{ and so } \alpha\Big(\gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big) < 0. \tag{4.30}$$

To show (4.29), we shall reformulate it in equivalent forms:

$$\Big(\lambda^2 - \frac{(1-\lambda)^2}{2}\Big)\frac{\|v_0\|^2}{(1-\lambda)^2} + (1+\lambda^2)\frac{\alpha^2}{\beta^2} > 2\alpha\gamma_0,$$

which is
$$\frac{\lambda^2+2\lambda-1}{2}\Big(-2\alpha\Big(\gamma_0+\frac{\lambda\alpha}{\beta^2}\Big)\Big)+(1+\lambda^2)\frac{\alpha^2}{\beta^2}>2\alpha\gamma_0$$
by (4.30). After simplifications, this reduces to
$$\frac{\alpha^2}{\beta^2}(1+\lambda)^2(1-\lambda)>\alpha\gamma_0(1+\lambda)^2.$$
Since $\lambda+1<0$, this is equivalent to
$$\frac{\alpha^2}{\beta^2}(1-\lambda)>\alpha\gamma_0,\quad \text{i.e.,}\quad \alpha\Big(\gamma_0+\frac{\lambda\alpha}{\beta^2}\Big)<\frac{\alpha^2}{\beta^2},$$
which obviously holds because of (4.30) and $\alpha^2/\beta^2>0$.

Hence, (4.25) of Case 1 gives the optimal solution.

Situation 3:
$$\alpha\Big(\gamma_0-\frac{\alpha}{\beta^2}\Big)=-\frac{\|v_0\|^2}{8}. \tag{4.31}$$
We again consider two cases.

Case 1: $1+\lambda=0$, i.e., $\lambda=-1$. By (4.21b), $v=\frac{v_0}{2}$ and then (4.21d) and (4.21c) give
$$\|u\|^2=2\alpha\gamma+\frac{\|v_0\|^2}{4}=2\alpha\Big(\gamma_0-\frac{\alpha}{\beta^2}\Big)+\frac{\|v_0\|^2}{4}=0,$$
so $u=0$. The possible optimal value is attained at
$$\Big(0,\frac{v_0}{2},\gamma_0-\frac{\alpha}{\beta^2}\Big) \tag{4.32}$$
with
$$f\Big(0,\frac{v_0}{2},\gamma_0-\frac{\alpha}{\beta^2}\Big)=\frac{\|v_0\|^2}{4}+\frac{\alpha^2}{\beta^2}.$$
Case 2: $u=0$. By (4.21d), $-\|v\|^2=2\alpha\gamma$, together with (4.21c), we have
$$g_2(\lambda):=\frac{\|v_0\|^2}{(1-\lambda)^2}+2\alpha\Big(\gamma_0+\frac{\lambda\alpha}{\beta^2}\Big)=0.$$
By (4.31),
$$g_2(-1)=\frac{\|v_0\|^2}{4}+2\alpha\Big(\gamma_0-\frac{\alpha}{\beta^2}\Big)=0.$$
As
$$g_2'(\lambda)=\frac{2\|v_0\|^2}{(1-\lambda)^3}+\frac{2\alpha^2}{\beta^2}>0\quad\text{on}\quad]-\infty,1[\,,$$
g_2 is strictly increasing and continuous on $]-\infty,1[$, so $\lambda=-1$ is the unique solution in $]-\infty,1[$. Then the possible optimal value is attained at
$$\Big(0,\frac{v_0}{2},\gamma_0-\frac{\alpha}{\beta^2}\Big)$$

with

$$f\Big(0, \frac{v_0}{2}, \gamma_0 - \frac{\alpha}{\beta^2}\Big) = \frac{\|v_0\|^2}{4} + \frac{\alpha^2}{\beta^2}. \tag{4.33}$$

Therefore, Case 1 and Case 2 give exactly the same solution. The optimal solution is given by (4.32), and it can be recovered by (4.25), the optimal solution of Situation 2. □

4.2.3 Case (iii): $u_0 \neq 0, v_0 = 0$

Proof. The minimization problem now is

$$\text{minimize} \quad f(u, v, \gamma) = \|u_0 - u\|^2 + \|v\|^2 + \beta^2|\gamma_0 - \gamma|^2 \tag{4.34a}$$

$$\text{subject to} \quad \|u\|^2 - \|v\|^2 = 2\alpha\gamma. \tag{4.34b}$$

Rewrite it as

$$\text{minimize} \quad f(u, v, \gamma) = \|v\|^2 + \|u_0 - u\|^2 + \beta^2|\gamma_0 - \gamma|^2 \tag{4.35a}$$

$$\text{subject to} \quad \|v\|^2 - \|u\|^2 = 2(-\alpha)\gamma. \tag{4.35b}$$

Luckily, we can apply Section 4.2.2 for the point $(0, u_0, \gamma_0)$ and parameter $-\alpha$. More precisely, when $-\alpha(\gamma_0 - \frac{-\alpha}{\beta^2}) < -\frac{\|u_0\|^2}{8}$, the optimal solution to (4.35) is

$$\Big(0, \frac{u_0}{1-\tilde{\lambda}}, \gamma_0 + \frac{\tilde{\lambda}(-\alpha)}{\beta^2}\Big)$$

where $\tilde{g}_2(\tilde{\lambda}) = 0$, $\tilde{\lambda} \in \,]-1, 1[$, and

$$\tilde{g}_2(\tilde{\lambda}) = \frac{\|u_0\|^2}{(1-\tilde{\lambda})^2} + 2(-\alpha)\Big(\gamma_0 - \frac{\tilde{\lambda}\alpha}{\beta^2}\Big) = 0.$$

Put $\lambda = -\tilde{\lambda}$. Simplifications give: when $\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) > \frac{\|u_0\|^2}{8}$, the optimal solution to (4.35) is

$$\Big(0, \frac{u_0}{1+\lambda}, \gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big) \tag{4.36}$$

where $g_2(\lambda) = 0$, $\lambda \in \,]-1, 1[$, and

$$g_2(\lambda) := \tilde{g}_2(-\lambda) = \frac{\|u_0\|^2}{(1+\lambda)^2} - 2\alpha\Big(\gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big) = 0.$$

Switching the first and second components in (4.36) gives the optimal solution to (4.34).

When $-\alpha(\gamma_0 - \frac{-\alpha}{\beta^2}) \geqslant -\frac{\|u_0\|^2}{8}$, the optimal solution to (4.35) is

$$\Big(v, \frac{u_0}{2}, \gamma_0 - \frac{-\alpha}{\beta^2}\Big)$$

with

$$\|v\|^2 = 2(-\alpha)\Big(\gamma_0 - \frac{-\alpha}{\beta^2}\Big) + \frac{\|u_0\|^2}{4}.$$

That is, when $\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) \leqslant \frac{\|u_0\|^2}{8}$, the optimal solution to (4.35) is

$$\Big(v, \frac{u_0}{2}, \gamma_0 + \frac{\alpha}{\beta^2}\Big) \tag{4.37}$$

with

$$\|v\|^2 = -2\alpha\Big(\gamma_0 + \frac{\alpha}{\beta^2}\Big) + \frac{\|u_0\|^2}{4}.$$

Switching the first and second components in (4.37) gives the optimal solution to (4.34). □

4.2.4 Case (iv): $u_0 = v_0 = 0$

Proof. The objective function is $f(u, v, \gamma) = \|u\|^2 + \|v\|^2 + \beta^2|\gamma - \gamma_0|^2$, and the KKT optimality conditions (4.20) become

$$(1+\lambda)u = 0, \tag{4.38a}$$
$$(1-\lambda)v = 0, \tag{4.38b}$$
$$\gamma = \gamma_0 + \frac{\lambda\alpha}{\beta^2}, \tag{4.38c}$$
$$\|u\|^2 - \|v\|^2 = 2\alpha\gamma. \tag{4.38d}$$

We shall consider three cases:
(i) $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) > 0$; hence, $\gamma_0 - \frac{\alpha}{\beta^2} \neq 0$.
(ii) $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) = 0$; hence, $\gamma_0 - \frac{\alpha}{\beta^2} = 0$.
(iii) $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) < 0$; hence, $\gamma_0 - \frac{\alpha}{\beta^2} \neq 0$.
For each item (i)–(iii), we will apply (4.38):
Case 1: $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) > 0$. By (4.38a), we have $\lambda = -1$ or $u = 0$. We consider two subcases.

Subcase 1: $\lambda = -1$. Using (4.38b), (4.38c) and (4.38d), we obtain $v = 0$, $\gamma = \gamma_0 - \frac{\alpha}{\beta^2}$, and

$$\|u\|^2 = 2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big). \tag{4.39}$$

Therefore, the candidate for the solution is $(u, 0, \gamma_0 - \frac{\alpha}{\beta^2})$ with u given by (4.39) and its objective function value is

$$f\Big(u, 0, \gamma_0 - \frac{\alpha}{\beta^2}\Big) = 2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big) + 0 + \beta^2\Big(\frac{-\alpha}{\beta^2}\Big)^2 = 2\alpha\gamma_0 - \frac{\alpha^2}{\beta^2}. \tag{4.40}$$

Subcase 2: $u = 0$. Using (4.38b)–(4.38d), we obtain $-\|v\|^2 = 2\alpha\gamma, \gamma = \gamma_0 + \frac{\lambda\alpha}{\beta^2}$ and $(1-\lambda)v = 0$. We have to consider two further cases: $1 - \lambda = 0$ or $v = 0$.

(i) $v = 0$. We get $-(0)^2 = 2\alpha\gamma \Rightarrow \gamma = 0$ because $\alpha \neq 0$. This gives a possible solution $(0,0,0)$ with function value

$$f(0,0,0) = \|u\|^2 + \|v\|^2 + \beta^2|\gamma - \gamma_0|^2 = \beta^2\gamma_0^2. \tag{4.41}$$

(ii) $\lambda = 1$. We have $\gamma = \gamma_0 + \frac{\alpha}{\beta^2}$ and $-\|v\|^2 = 2\alpha(\gamma_0 + \frac{\alpha}{\beta^2})$. So, $0 \leqslant \|v\|^2 = -2\alpha(\gamma_0 + \frac{\alpha}{\beta^2})$. However,

$$-2\alpha\Big(\gamma_0 + \frac{\alpha}{\beta^2}\Big) = \underbrace{-2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big)}_{<0} - \frac{4\alpha^2}{\beta^2} < 0 \tag{4.42}$$

because $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) > 0$. This contradiction shows $\lambda = 1$ does not happen. We now compare objective function values (4.40) and (4.41):

$$\begin{aligned} 2\alpha\gamma_0 - \frac{\alpha^2}{\beta^2} < \beta^2\gamma_0^2 &\Leftrightarrow \beta^2\gamma_0^2 + \frac{\alpha^2}{\beta^2} - 2\alpha\gamma_0 > 0 \\ &\Leftrightarrow \Big(\beta\gamma_0 - \frac{\alpha}{\beta}\Big)^2 > 0 \\ &\Leftrightarrow \beta^2\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big)^2 > 0, \end{aligned}$$

which holds because $\gamma_0 - \frac{\alpha}{\beta^2} \neq 0$. Hence, the optimal solution is $(u, 0, \gamma_0 - \frac{\alpha}{\beta^2})$ with $\|u\| = \sqrt{2\alpha(\gamma_0 - \frac{\alpha}{\beta^2})}$. That is,

$$P_{\tilde{C}_2}(0,0,\gamma_0) = \bigg\{\Big(u, 0, \gamma_0 - \frac{\alpha}{\beta^2}\Big) \;\bigg|\; \|u\| = \sqrt{2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big)}\bigg\}.$$

Case 2: $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) = 0$; hence, $\gamma_0 - \frac{\alpha}{\beta^2} = 0$. By (4.38a), we have two subcases to consider.

Subcase 1: $\lambda = -1$. We have $v = 0$, $\gamma = \gamma_0 - \frac{\alpha}{\beta^2} = 0$, $\|u\|^2 = 2\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) = 0$. The possible solution is $(0,0,0)$.

Subcase 2: $u = 0$. We have $-\|v\|^2 = 2\alpha\gamma$ and $\gamma = \gamma_0 + \frac{\lambda\alpha}{\beta^2}$. By (4.38b), $v = 0$ or $\lambda = 1$. This requires us to consider two further cases. For $v = 0$, we get $\gamma = 0$, which gives a possible solution $(0,0,0)$. For $\lambda = 1$, we get $\gamma = \gamma_0 + \frac{\alpha}{\beta^2}$, $\|v\|^2 = -2\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) = \frac{-4\alpha^2}{\beta^2} < 0$, which is impossible, i.e., $\lambda = 1$ does not happen.

Both **Subcase 1** and **Subcase 2** give the same solution $(0,0,0)$. Therefore, we have the optimal solution is $(0,0,0)$, when $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) = 0$; equivalently, when $\gamma_0 = \frac{\alpha}{\beta^2}$.

Case 3: $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) < 0$. In view of (4.38a), we have $\lambda = -1$ or $u = 0$. We show that $\lambda = -1$ can't happen. Indeed, when $\lambda = -1$, by (4.38b)–(4.38c), we have $v = 0$, $\gamma = \gamma_0 - \frac{\alpha}{\beta^2}$, and $0 \leqslant \|u\|^2 = 2\alpha\gamma = 2\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) < 0$, which is impossible. Therefore, we consider only the case $u = 0$. Then (4.38b)–(4.38d)

yield $\|v\|^2 = -2\alpha\gamma$, $\gamma = \gamma_0 + \frac{\lambda\alpha}{\beta^2}$, and $(1-\lambda)v = 0$, which requires us to consider two further cases.
Subcase 1: $v = 0$. Then $\gamma = 0$. The possible optimal solution is $(0,0,0)$ and its objective function value is

$$f(0,0,0) = \|u\|^2 + \|v\|^2 + \beta^2|\gamma - \gamma_0|^2 = \beta^2\gamma_0^2. \tag{4.43}$$

Subcase 2: $\lambda = 1$. Then $u = 0$, $\gamma = \gamma_0 + \frac{\alpha}{\beta^2}$, and $-\|v\|^2 = 2\alpha(\gamma_0 + \frac{\alpha}{\beta^2})$. We consider three additional cases based on the sign of $\alpha(\gamma_0 + \frac{\alpha}{\beta^2})$.

(i) $\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) > 0$. This case never happens because the relation $0 \geqslant -\|v\|^2 = 2\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) > 0$ is absurd.

(ii) $\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) = 0$. As $\alpha \neq 0$, we have $\gamma_0 + \frac{\alpha}{\beta^2} = 0$. This gives $\gamma = 0, u = 0$ and $v = 0$. So the possible optimal solution is $(0,0,0)$.

(iii) $\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) < 0$. We have $\gamma_0 + \frac{\alpha}{\beta^2} \neq 0$. The possible optimal solution is $(0, v, \gamma_0 + \frac{\alpha}{\beta^2})$ with $\|v\| = \sqrt{-2\alpha(\gamma_0 + \frac{\alpha}{\beta^2})}$ and function value

$$f\Big(0, v, \gamma_0 + \frac{\alpha}{\beta^2}\Big) = -2\alpha\Big(\gamma_0 + \frac{\alpha}{\beta^2}\Big) + \beta^2\Big(\frac{\alpha}{\beta^2}\Big)^2 = -2\alpha\gamma_0 - \frac{\alpha^2}{\beta^2}. \tag{4.44}$$

Both (i) and (ii) imply that $(0,0,0)$ from **Subcase 1** is the only optimal solution, when $\alpha^2/\beta^2 > \alpha\gamma_0 \geqslant -\alpha^2/\beta^2$.

When $\alpha\gamma_0 < -\frac{\alpha^2}{\beta^2}$, both **Subcase 1** and **Subcase 2** happen. We have to compare objectives (4.43) and (4.44). We claim $f(0, v, \gamma_0 + \frac{\alpha}{\beta^2}) < f(0,0,0)$. Indeed, this is equivalent to

$$\begin{aligned} -2\alpha\gamma_0 - \frac{\alpha^2}{\beta^2} < \beta^2\gamma_0^2 &\Leftrightarrow \beta^2\gamma_0^2 + 2\alpha\gamma_0 + \frac{\alpha^2}{\beta^2} > 0 \\ &\Leftrightarrow \Big(\beta\gamma_0 + \frac{\alpha}{\beta}\Big)^2 > 0 \\ &\Leftrightarrow \beta^2\Big(\gamma_0 + \frac{\alpha}{\beta^2}\Big)^2 > 0 \end{aligned}$$

which holds because $\gamma_0 + \frac{\alpha}{\beta^2} \neq 0$. Therefore, the optimal solution is $(0, v, \gamma_0 + \frac{\alpha}{\beta^2})$ with $\|v\| = \sqrt{-2\alpha(\gamma_0 + \frac{\alpha}{\beta^2})}$, i.e.,

$$P_{\tilde{C}_2}(0,0,\gamma_0) = \left\{\Big(0, v, \gamma_0 + \frac{\alpha}{\beta^2}\Big) \;\middle|\; \|v\| = \sqrt{-2\alpha\Big(\gamma_0 + \frac{\alpha}{\beta^2}\Big)}\right\}$$

when $\alpha\gamma_0 < \frac{-\alpha^2}{\beta^2}$. □

Altogether four cases above conclude the proof of Theorem 4.7. □

4.3 A numerical example

Let us illustrate Theorem 4.7.

Example 4.8. Suppose that $X = \mathbb{R}$, $\alpha = 5$, and $\beta = 1$. Writing z instead of γ, we note that $\widetilde{C}_\alpha$ turns into the set

$$S := \big\{(x, y, z) \in \mathbb{R}^3 \;\big|\; x^2 - y^2 = 10z\big\} = \operatorname{gra}\big((x, y) \mapsto \tfrac{1}{10}(x^2 - y^2)\big).$$

Let us now compute $P_S(x_0, y_0, z_0)$ for various points.

(i) Suppose that $(x_0, y_0, z_0) = (2, -3, 4)$.
In view of Theorem 4.7(i), we set $p := |x_0|^2 - |y_0|^2 = 2^2 - (-3)^2 = -5$ and $q := |x_0|^2 + |y_0|^2 = 2^2 + (-3)^2 = 13$. Following (4.10), we consider the equation

$$\frac{(\lambda^2 + 1)p - 2\lambda q}{(1 - \lambda^2)^2} - \frac{2\lambda\alpha^2}{\beta^2} - 2\alpha\gamma_0 = -\frac{5\lambda^2 + 26\lambda + 5}{(1 - \lambda^2)^2} - 50\lambda - 40 = 0$$

which has $\lambda = -0.52416$ as its unique (approximate) root in $]-1, 1[$. Using (4.9) now yields

$$\begin{aligned} P_S(x_0, y_0, z_0) &= \left\{\Big(\frac{x_0}{1 + \lambda}, \frac{y_0}{1 - \lambda}, z_0 + \frac{\lambda\alpha}{\beta^2}\Big)\right\} \\ &= \Big\{\big(4.20311, -1.96830, 1.37919\big)\Big\}. \end{aligned}$$

This is depicted in Figure 4.1 with the green arrow.

(ii) Suppose that $(x_0, y_0, z_0) = (0, -3, 3)$.
In view of Theorem 4.7(ii), we evaluate $\alpha(z_0 - \frac{\alpha}{\beta^2}) = 5(3 - 5) = -10 < -\frac{9}{8} = -\frac{|y_0|^2}{8}$ and we are thus in case (a). In view of (4.12), we consider the equation

$$\frac{|y_0|^2}{(1 - \lambda)^2} + \frac{2\lambda\alpha^2}{\beta^2} + 2\alpha z_0 = \frac{9}{(1 - \lambda)^2} + 50\lambda + 30 = 0$$

which has $\lambda = -0.66493$ as its unique (approximate) root in $]-1, 1[$. Using (4.11) now yields

$$\begin{aligned} P_S(x_0, y_0, z_0) &= \left\{\Big(0, \frac{v_0}{1 - \lambda}, \gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big)\right\} \\ &= \Big\{\big(0, -1.80187, -0.32467\big)\Big\}. \end{aligned}$$

This is depicted in Figure 4.1 with a single blue arrow.

(iii) Suppose that $(x_0, y_0, z_0) = (0, \sqrt{32}, 6) = (0, 5.65685, 6)$.
In view of Theorem 4.7(ii), we evaluate $\alpha(z_0 - \frac{\alpha}{\beta^2}) = 5(6 - 5) = 5 > -4 = -\frac{32}{8} = -\frac{|y_0|^2}{8}$ and we are thus in case (b). We compute

$$\sqrt{2\alpha\Big(z_0 - \frac{\alpha}{\beta^2}\Big) + \frac{|y_0|^2}{4}} = \sqrt{10(6 - 5) + \frac{32}{4}} = \sqrt{18}$$

and now (4.13) yields

$$P_S(x_0,y_0,z_0) = \left\{\Big(x, \frac{y_0}{2}, z_0 - \frac{\alpha}{\beta^2}\Big) \;\middle|\; |x| = \sqrt{2\alpha\Big(z_0 - \frac{\alpha}{\beta^2}\Big) + \frac{|y_0|^2}{4}}, u \in \mathbb{R}\right\}$$
$$= \Big\{\big(\pm\sqrt{18}, \sqrt{8}, 1\big)\Big\} = \Big\{\big(\pm 4.24264, 2.82843, 1\big)\Big\}.$$

This is depicted in Figure 4.1 with double blue arrows.

(iv) Suppose that $(x_0,y_0,z_0) = (0,0,6)$.
In view of Theorem 4.7(iv), we have $\alpha z_0 = 5(6) = 30 > 25 = \frac{\alpha^2}{\beta^2}$ and we are thus in case (a). We compute

$$\sqrt{2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big)} = \sqrt{10(6-5)} = \sqrt{10}$$

and now (4.17) yields

$$P_S(x_0,y_0,z_0) = \left\{\Big(x, 0, z_0 - \frac{\alpha}{\beta^2}\Big) \;\middle|\; |x| = \sqrt{2\alpha\Big(z_0 - \frac{\alpha}{\beta^2}\Big)}, u \in \mathbb{R}\right\}$$
$$= \Big\{\big(\pm\sqrt{10}, 0, 1\big)\Big\} = \Big\{\big(\pm 3.16228, 0, 1\big)\Big\}.$$

This is depicted in Figure 4.1 with double black arrows.

(v) Suppose that $(x_0,y_0,z_0) = (0,0,4)$.
In view of Theorem 4.7(iv), we have $|\alpha z_0| = |5(4)| = 20 < 25 = \frac{\alpha^2}{\beta^2}$ and we are thus in case (b). Therefore,

$$P_S(x_0,y_0,z_0) = \big\{(0,0,0)\big\}.$$

This is depicted in Figure 4.1 with a single black arrow.

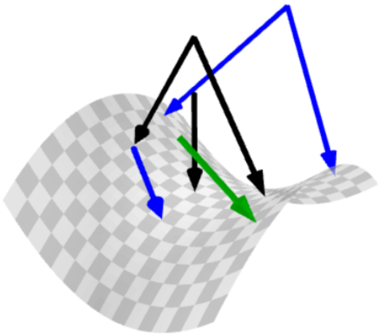

Figure 4.1: Visualization of the 5 projections from Example 4.8.

4.4 Further results

Recall that

$$C_\alpha = \big\{(x, y, \gamma) \in X \times X \times \mathbb{R} \;\big|\; \langle x, y\rangle = \alpha\gamma\big\},$$

and this is the representation more natural to use in Deep Learning (see [40]). Armed with Theorem 4.7, the projection onto C_α now can be readily obtained:

Theorem 4.9. *Let $(x_0, y_0, \gamma_0) \in X \times X \times \mathbb{R}$. Then the following hold:*

(i) If $x_0 \neq \pm y_0$, then

$$P_{C_\alpha}(x_0, y_0, \gamma_0) = \left\{\Big(\frac{x_0 - \lambda y_0}{1-\lambda^2}, \frac{y_0 - \lambda x_0}{1-\lambda^2}, \gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big)\right\}$$

for a unique $\lambda \in \left]-1, 1\right[$ that solves the (essentially) quintic equation

$$g(\lambda) := \frac{(\lambda^2+1)p - 2\lambda q}{(1-\lambda^2)^2} - \frac{2\lambda\alpha^2}{\beta^2} - 2\alpha\gamma_0 = 0,$$

where $p := 2\langle x_0, y_0\rangle$ and $q := \|x_0\|^2 + \|y_0\|^2$.

(ii) If $y_0 = -x_0 \neq 0$, then we have the following:

a) When $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) < -\frac{\|x_0\|^2}{4}$, then

$$P_{C_\alpha}(x_0, -x_0, \gamma_0) = \left\{\Big(\frac{x_0}{1-\lambda}, \frac{-x_0}{1-\lambda}, \gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big)\right\}$$

for a unique $\lambda \in \left]-1, 1\right[$ that solves

$$g_1(\lambda) := \frac{2\|x_0\|^2}{(1-\lambda)^2} + \frac{2\lambda\alpha^2}{\beta^2} + 2\alpha\gamma_0 = 0.$$

b) When $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) \geqslant -\frac{\|x_0\|^2}{4}$, then

$$P_{C_\alpha}(x_0, -x_0, \gamma_0) = \left\{\Big(\frac{x_0}{2} + \frac{u}{\sqrt{2}}, -\frac{x_0}{2} + \frac{u}{\sqrt{2}}, \gamma_0 - \frac{\alpha}{\beta^2}\Big) \;\Big|\; \|u\| = \sqrt{2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big) + \frac{\|x_0\|^2}{2}},\ u \in X\right\},$$

which is a singleton if and only if $\alpha(\gamma_0 - \frac{\alpha}{\beta^2}) = -\frac{\|x_0\|^2}{4}$.

(iii) If $y_0 = x_0 \neq 0$, then we have the following:

a) When $\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) > \frac{\|x_0\|^2}{4}$, then

$$P_{C_\alpha}(x_0, x_0, \gamma_0) = \left\{\Big(\frac{x_0}{1+\lambda}, \frac{x_0}{1+\lambda}, \gamma_0 + \frac{\lambda\alpha}{\beta^2}\Big)\right\}$$

for a unique $\lambda \in \left]-1,1\right[$ that solves the (essentially) cubic equation

$$g_2(\lambda) := \frac{2\|x_0\|^2}{(1+\lambda)^2} - \frac{2\lambda\alpha^2}{\beta^2} - 2\alpha\gamma_0 = 0.$$

b) When $\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) \leqslant \frac{\|x_0\|^2}{4}$, then

$$\begin{aligned} P_{C_\alpha}(x_0, x_0, \gamma_0) = \Big\{ \Big(\frac{x_0}{2} - \frac{v}{\sqrt{2}}, \frac{x_0}{2} + \frac{v}{\sqrt{2}}, \gamma_0 + \frac{\alpha}{\beta^2} \Big) \;\Big|\; \|v\| \\ = \sqrt{-2\alpha\Big(\gamma_0 + \frac{\alpha}{\beta^2}\Big) + \frac{\|x_0\|^2}{2}},\; v \in X \Big\}, \end{aligned}$$

which is a singleton if and only if $\alpha(\gamma_0 + \frac{\alpha}{\beta^2}) = \frac{\|x_0\|^2}{4}$.

(iv) If $x_0 = y_0 = 0$, then we have the following:

a) When $\alpha\gamma_0 > \frac{\alpha^2}{\beta^2}$, then the projection is the non-singleton set

$$P_{C_\alpha}(0,0,\gamma_0) = \Big\{ \Big(\frac{u}{\sqrt{2}}, \frac{u}{\sqrt{2}}, \gamma_0 - \frac{\alpha}{\beta^2} \Big) \;\Big|\; \|u\| = \sqrt{2\alpha\Big(\gamma_0 - \frac{\alpha}{\beta^2}\Big)},\; u \in X \Big\}.$$

b) When $|\alpha\gamma_0| \leqslant \frac{\alpha^2}{\beta^2}$, then

$$P_{C_\alpha}(0,0,\gamma_0) = \big\{(0,0,0)\big\}.$$

c) When $\alpha\gamma_0 < -\frac{\alpha^2}{\beta^2}$, then the projection is the non-singleton set

$$P_{C_\alpha}(0,0,\gamma_0) = \Big\{ \Big(-\frac{v}{\sqrt{2}}, \frac{v}{\sqrt{2}}, \gamma_0 + \frac{\alpha}{\beta^2} \Big) \;\Big|\; \|v\| = \sqrt{-2\alpha\Big(\gamma_0 + \frac{\alpha}{\beta^2}\Big)},\; v \in X \Big\}.$$

Proof. With

$$A = \begin{bmatrix} \frac{1}{\sqrt{2}}\,\mathrm{Id} & -\frac{1}{\sqrt{2}}\,\mathrm{Id} & 0 \\ \frac{1}{\sqrt{2}}\,\mathrm{Id} & \frac{1}{\sqrt{2}}\,\mathrm{Id} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

in mind, by Proposition 4.2(iii) we have

$$\begin{aligned} P_{C_\alpha}[x_0, y_0, \gamma_0]^\intercal &= A\, P_{\widetilde{C}_\alpha} A^\intercal [x_0, y_0, \gamma_0]^\intercal \\ &= A\, P_{\widetilde{C}_\alpha} \Big[\frac{x_0 + y_0}{\sqrt{2}}, \frac{-x_0 + y_0}{\sqrt{2}}, \gamma_0 \Big]^\intercal. \end{aligned}$$

Hence (i)–(iv) follow by applying Theorem 4.7. □

Remark 4.10. Theorem 4.9(i) was given in [40, Appendix B] without a rigorous mathematical justification.

It is interesting to ask what happens when $\alpha \to 0$.

Theorem 4.11. *Suppose that* $X = \mathbb{R}^n$. *Then* $P_{C_\alpha} \xrightarrow{g} P_{C\times\mathbb{R}} = P_C \times \mathrm{Id}$ *and* $P_{\tilde{C}_\alpha} \xrightarrow{g} P_{\tilde{C}\times\mathbb{R}} = P_{\tilde{C}} \times \mathrm{Id}$ *when* $\alpha \to 0$.

Proof. Apply Proposition 4.4 and Fact 4.6. □

Remark 4.12. The projection onto the cross C, P_C, has been given in [12].

Chapter 5

Real roots of real cubics and optimization

This chapter is based on the paper [15] titled "Real roots of real cubics and optimization" that has been submitted to a journal and which has already been cited by a group of researchers from China [70] and Japan [49]. The solution of the cubic equation has a century-long history; however, the usual presentation is geared towards applications in algebra and is somewhat inconvenient to use in optimization where frequently the main interest lies in real roots. In this chapter, we present the roots of the cubic in a form that makes them convenient to use and we also focus on information on the location of the real roots. Armed with this, we provide several applications in optimization where we compute Fenchel conjugates, proximal mappings and projections.

The history of solving cubic equations is rich and centuries-old; see, e.g., Confalonieri's recent book [31] on Cardano's work. Cubics do also appear in convex and nonconvex optimization. However, treatises on solving the cubic often focus on the general complex case making the results less useful to optimizers. The purpose of this chapter is two-fold. We present a largely self-contained derivation of the solution of the cubic with an emphasis on usefulness to practitioners. We do not claim the novelty of these results; however, the presentation appears to be particularly convenient for its use in convex analysis and optimization. We then turn to novel results from Section 5.4 to Section 5.8. We show how the formulas can be used to compute Fenchel conjugates and proximal mappings of some convex functions. We also discuss projections on convex and nonconvex sets.

5.1 Some facts

First, we collect some properties of polynomials that are well-known; as a reference, we recommend [58].

Fact 5.1. *Let $f(x)$ be a nonconstant complex polynomial and let $r \in \mathrm{C}$ such that $f(r) = 0$. Then the multiplicity of r is is the smallest integer k such that the kth derivative at r is nonzero: $f^{(k-1)}(r) = 0$ and $f^{(k)}(r) \neq 0$. When $k = 1$, 2, or 3, then we say that r is a simple, double, or triple root, respectively.*

Fact 5.2. (Vieta) *Suppose $f(x) = ax^3 + bx^2 + cx + d$ is a cubic polynomial (i.e., $a \neq 0$) with complex coefficients. If r_1, r_2, r_3 denote the (possibly repeated* *and complex) roots of* f*, then*

$$r_1 + r_2 + r_3 = -\frac{b}{a}, \tag{5.1a}$$

$$r_1 r_2 + r_1 r_3 + r_2 r_3 = \frac{c}{a}, \tag{5.1b}$$

$$r_1 r_2 r_3 = -\frac{d}{a}. \tag{5.1c}$$

Conversely, if r_1, r_2, r_3 *in* $\mathbb{C}$ *satisfy* (5.1)*, then they are the (possibly repeated) roots of* f*.*

Fact 5.3. *Suppose* $f(x) = ax^3 + bx^2 + cx + d$ *is a cubic polynomial (i.e.,* $a \neq 0$*) with real coefficients. Then* f *has three (possibly complex) roots (counting multiplicity). More precisely, exactly one of the following holds:*

(i) f *has exactly one real root which either is simple (and the two remaining roots are nonreal simple roots and conjugate to each other) or is a triple root.*

(ii) f *has exactly two distinct real roots: one is simple and the other double.*

(iii) f *has exactly three distinct simple real roots.*

Remark 5.4. We mention that the roots of a polynomial of a *fixed* degree depend continuously on the coefficients — see [58, Theorem 1.3.1] for a precise statement and also the other results in [58, Section 1.3].

5.2 The depressed cubic

In this section, we study the *depressed* cubic

$$g(z) := z^3 + pz + q, \quad \text{where } p \in \mathbb{R} \text{ and } q \in \mathbb{R}. \tag{5.2}$$

Note that the term "depressed" for the cubic (5.2) is by convention [31]. Depressed cubics are also called reduced cubics.

Theorem 5.5. *We have*

$$g'(z) = 3z^2 + p \ \ \textit{and} \ \ g''(z) = 6z. \tag{5.3}$$

Then 0 *is the only inflection point of* g*:* g *is strictly concave on* $\mathbb{R}_-$ *and* g *is strictly convex on* $\mathbb{R}_+$*.*

Moreover, exactly one of the following cases occurs:

(i) $p < 0$*: Set* $z_\pm := \pm\sqrt{-p/3}$*. Then* $z_- < z_+$*,* $z_\pm$ *are two distinct simple roots of* g'*,* g *is strictly increasing on* $]-\infty, z_-]$*,* g *is strictly decreasing on* $[z_-, z_+]$*,* g *is strictly increasing on* $[z_+, +\infty[$*. Moreover,*

$$g(z_-)g(z_+) = 4\Delta, \quad \textit{where } \Delta := (p/3)^3 + (q/2)^2, \tag{5.4}$$

and this case trifurcates further as follows:

(a) $\Delta > 0$*: Then* g *has exactly one real root* r*. It is simple and given by*

$$r := u_- + u_+, \quad \textit{where} \ u_\pm := \sqrt[3]{\frac{-q}{2} \pm \sqrt{\Delta}}. \tag{5.5}$$

The two remaining simple nonreal roots are

$$-\tfrac{1}{2}(u_- + u_+) \pm \mathrm{i}\tfrac{1}{2}\sqrt{3}(u_- - u_+). \tag{5.6}$$

(b) $\Delta = 0$*: If* $q > 0$ *(resp.* $q < 0$*), then* $2z_-$ *(resp.* $2z_+$*) is a simple real root while* z_+ *(resp.* z_-*) is a double root. Moreover, these cases can be combined into*[4]

$$\frac{3q}{p} = 2\sqrt[3]{\frac{-q}{2}} \ \textit{is a simple root and} \ \frac{-3q}{2p} = -\sqrt[3]{\frac{-q}{2}} \ \textit{is a double root.}$$

(c) $\Delta < 0$*: Then* g *has three simple real roots* r_-, r_0, r_+ *where* $r_- < z_- < r_0 < z_+ < r_+$*. Indeed, set*

$$\theta := \arccos \frac{-q/2}{(-p/3)^{3/2}}, \tag{5.7}$$

which lies in $]0, \pi[$*, and then define* z_0, z_1, z_2 *by*

$$z_k := 2(-p/3)^{1/2} \cos\Big(\frac{\theta + 2k\pi}{3}\Big). \tag{5.8}$$

Then $r_- = z_1$*,* $r_0 = z_2$*, and* $r_+ = z_0$*.*

(ii) $p = 0$*: Then* g' *has a double root at* 0*, and* g *is strictly increasing on* $\mathbb{R}$*. The only real root is*

$$r := (-q)^{1/3}. \tag{5.9}$$

If $q = 0$*, then* r *is a triple root. If* $q \neq 0$*, then* r *is a simple root and the remaining nonreal simple roots are* $-\frac{1}{2}r \pm \mathrm{i}\frac{1}{2}\sqrt{3}r$*.*

(iii) $p > 0$*: Then* g' *has no real root,* g *is strictly increasing on* $\mathbb{R}$*, and* g *has exactly one real root* r*. It is simple and given by*

$$r := u_- + u_+, \quad \textit{where} \ u_\pm := \sqrt[3]{\frac{-q}{2} \pm \sqrt{\Delta}} \ \textit{and} \ \Delta := (p/3)^3 + (q/2)^2. \tag{5.10}$$

Once again, the two remaining simple nonreal roots are

$$-\tfrac{1}{2}(u_- + u_+) \pm \mathrm{i}\tfrac{1}{2}\sqrt{3}(u_- - u_+). \tag{5.11}$$

Proof. Except for the formulas for the roots, all statements on g follow from standard calculus. (i)(a): Because $\Delta > 0$ and g is strictly decreasing on $[z_-, z_+]$, it follows from (5.4) that g has the same sign on $[z_-, z_+]$ and so g has no root in that interval. Now g is strictly increasing on $]-\infty, z_-]$ and on $[z_+, +\infty[$; hence,

[4] Observe that this is the case when $\Delta \to 0^+$ in (a).

g has exactly one real root r and it lies outside $[z_-, z_+]$. Note that r must be simple because the roots of g' are $z_\mp$ and $r \neq z_\mp$. Note that $u_- < u_+$. Next, $u_-^3 u_+^3 = (q/2)^2 - \Delta = -(p/3)^3$ and so

$$u_- u_+ = -p/3. \tag{5.12}$$

Also,

$$u_-^3 + u_+^3 = \frac{-q}{2} - \sqrt{\Delta} + \frac{-q}{2} + \sqrt{\Delta} = -q. \tag{5.13}$$

Hence

$$\begin{aligned} g(r) &= r^3 + pr + q \\ &= (u_- + u_+)^3 + p(u_- + u_+) + q \\ &= u_-^3 + u_+^3 + 3u_-u_+(u_- + u_+) + p(u_- + u_+) + q \\ &= \big(u_-^3 + u_+^3\big) + (3u_-u_+ + p)(u_- + u_+) + q \\ &= -q + \big(3(-p/3) + p\big)(u_- + u_+) + q \qquad \text{(using (5.12) and (5.13))} \\ &= 0 \end{aligned}$$

as claimed. Observe that we only need the properties (5.12) and (5.13) about u_-, u_+ to conclude that $u_- + u_+$ is a root of g. This observation leads us quickly to the two remaining complex roots: First, denote the primitive 3rd root of unity by ω, i.e.,

$$\omega := \exp(2\pi \mathrm{i}/3) = \cos(2\pi/3) + \mathrm{i}\sin(2\pi/3) = -\tfrac{1}{2} + \mathrm{i}\tfrac{1}{2}\sqrt{3}. \tag{5.14}$$

Then $\omega^2 = \overline{\omega} = -\frac{1}{2} - \mathrm{i}\frac{1}{2}\sqrt{3}$ and $\omega^3 = \overline{\omega}^3 = 1$. Now set

$$v_- := \omega u_- \ \text{ and } \ v_+ := \omega^2 u_+ = \overline{\omega} u_+.$$

Then $v_- v_+ (\omega u_-) = (\omega^2 u_+) = \omega^3 u_- u_+ = u_- u_+ = -p/3$ by (5.12), and $v_-^3 + v_+^3 = (\omega u_-)^3 + (\omega^2 u_+)^3 = \omega^3 u_-^3 + \omega^6 u_+^3 = u_-^3 + u_+^3 = -q$ by (5.12). Hence

$$\begin{aligned} v_- + v_+ &= \omega u_- + \overline{\omega} u_+ \\ &= \big(-\tfrac{1}{2} + \mathrm{i}\tfrac{1}{2}\sqrt{3}\big)u_- + \big(-\tfrac{1}{2} - \mathrm{i}\tfrac{1}{2}\sqrt{3}\big)u_+ \\ &= -\tfrac{1}{2}(u_- + u_+) + \mathrm{i}\tfrac{1}{2}\sqrt{3}(u_- - u_+) \end{aligned}$$

and its conjugate are the remaining simple complex roots of g.

(i)(b): From (5.4), it follows that z_- or z_+ is a root of g. In view of Fact 5.1 and $g'(z_-) = g'(z_+)$, it follows that one of z_-, z_+ is at least a double root, but not both; moreover, it cannot be a triple root because 0 is the only root of g'' and $z_- < 0 < z_+$. Hence exactly one of z_-, z_+ is a double root. To verify the remaining parts, we first define

$$r_1 := \frac{3q}{p} \ \text{ and } \ r_2 := \frac{-3q}{2p}.$$

Because $\Delta = 0$, it follows that $4p^3 + 27q^2 = 0$. Hence

$$g(r_1) = r_1^3 + pr_1 + q = \frac{27q^3}{p^3} + \frac{3pq}{p} + q = \frac{27q^3}{p^3} + 4q = \frac{q}{p^3}\big(27q^2 + 4p^3\big) = 0$$

and

$$g(r_2) = r_2^3 + pr_2 + q = \frac{-27q^3}{8p^3} + \frac{-3pq}{2p} + q = \frac{-27q^3}{8p^3} - \frac{q}{2} = \frac{-q}{8p^3}\big(27q^2 + 4p^3\big) = 0.$$

The assumption that $\Delta = 0$ readily yields

$$p = \frac{-3^{1/3}q^{2/3}}{2^{2/3}} \text{ and } |q| = \frac{2(-p)^{3/2}}{3^{3/2}}.$$

Hence

$$r_1 = 3qp^{-1} = 3q(-1)3^{-1/3}q^{-2/3}2^{2/3} = 2^{2/3}(-q)^{1/3}$$

and

$$r_2 = -3q2^{-1}p^{-1} = -3q2^{-1}(-1)3^{-1/3}q^{-2/3}2^{2/3} = -2^{-1/3}(-q)^{1/3}$$

as claimed.

If $q > 0$, then

$$r_1 = \frac{3q}{p} = \frac{3 \cdot 2(-p)^{3/2}}{3^{3/2}p} = -2(-p/3)^{1/2} = 2z_-$$

and

$$r_2 = \frac{-3q}{2p} = -\tfrac{1}{2}\frac{3q}{p} = -\tfrac{1}{2}r_1 = -\tfrac{1}{2}2z_- = z_+.$$

Similarly, if $q < 0$, then $r_1 = 2z_+$ and $r_2 = z_-$.

No matter the sign of q, we have $r_2 \in \{z_-, z_+\}$ and thus $g'(r_2) = 0$, i.e., r_2 is the double root.

(i)(c): In view of (5.4), $g(z_-)$ and $g(z_+)$ have opposite signs. Because g is strictly decreasing on $[z_-, z_+]$, it follows that $g(z_-) > 0 > g(z_+)$. Hence there is at least on real root r_0 in $]z_-, z_+[$. On the other hand, g is strictly increasing on $]-\infty, z_-]$ and on $[z_+, +\infty[$ which yields further roots r_- and r_+ as announced. Having now three real roots, they must all be simple.

Next, note that $\Delta < 0 \Leftrightarrow 0 \leqslant (q/2)^2 < -(p/3)^3 = (-p/3)^3 \Leftrightarrow 0 \leqslant (q/2)^2/(-p/3)^3 < 1 \Leftrightarrow 0 \leqslant (|q|/2)/(-p/3)^{3/2} < 1 \Leftrightarrow -1 < (-q/2)/(-p/3)^{3/2} < 1$. It follows that

$$\theta = \arccos \frac{-q/2}{(-p/3)^{3/2}} \in\,]0, \pi[\tag{5.15}$$

as claimed. For convenience, we set, for $k \in \{0, 1, 2\}$,

$$\theta_k := \frac{\theta + 2k\pi}{3}; \quad \text{hence, } z_k = 2(-p/3)^{1/2}\cos(\theta_k). \tag{5.16}$$

Recall that $0 < \theta < \pi$, which allows us to draw three conclusions:

$$\begin{aligned}0 < \theta_0 = \theta/3 < \pi/3 &\Rightarrow 1 > \cos(\theta_0) = \cos(\theta/3) > 1/2;\\ 2\pi/3 < \theta_1 = (\theta+2\pi)/3 < \pi &\Rightarrow -1/2 > \cos(\theta_1) = \cos((\theta+2\pi)/3) > -1;\\ 4\pi/3 < \theta_2 = (\theta+4\pi)/3 < 5\pi/3 &\Rightarrow -1/2 < \cos(\theta_2) = \cos((\theta+2\pi)/3) < 1/2.\end{aligned}$$

Hence $\cos(\theta_1) < \cos(\theta_2) < \cos(\theta_0)$ and thus

$$z_1 < z_2 < z_0. \tag{5.18}$$

All we need to do is to verify that each z_k is actually a root of g. To this end, observe first that the triple-angle formula for the cosine yields

$$\cos^3(\theta_k) = \frac{3\cos(\theta_k)+\cos(3\theta_k)}{4} = \frac{3\cos(\theta_k)+\cos(\theta+2k\pi)}{4} \tag{5.19a}$$

$$= \frac{3\cos(\theta_k)+\cos(\theta)}{4}. \tag{5.19b}$$

Then

$$\begin{aligned}g(z_k) &= z_k^3 + pz_k + q\\ &= 8(-p/3)^{3/2}\cos^3(\theta_k) + p2(-p/3)^{1/2}\cos(\theta_k) + q\\ &= 2(-p/3)^{3/2}\big(3\cos(\theta_k)+\cos(\theta)\big) + 2(-p/3)^{1/2}p\cos(\theta_k) + q\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(using (5.19))}\\ &= 2(-p/3)^{1/2}\cos(\theta_k)\big(3(-p/3)+p\big) + 2(-p/3)^{3/2}\cos(\theta) + q\\ &= 2(-p/3)^{3/2}\cos(\theta) + q\\ &= 2(-p/3)^{3/2}\frac{-q/2}{(-p/3)^{3/2}} + q \qquad\qquad\text{(using (5.15))}\\ &= 0,\end{aligned}$$

and this completes the proof for this case.

(ii): If $q = 0$, then $g(z) = z^3$ so $z = 0$ is the only root of g and it is of multiplicity 3. Thus we assume that $q \neq 0$. Then $g(z) = 0 \Leftrightarrow z^3 + q = 0 \Leftrightarrow z^3 = -q \Rightarrow z = (-q)^{1/3} \neq 0$. Because g is strictly increasing on $\mathbb{R}$, $r := (-q)^{1/3}$ is the only real root of g. Because g' has only one real root, namely 0, it follows that $g'(r) \neq 0$ and so r is a simple root. Denoting again by ω the primitive 3rd root of unity (see (5.14)), it is clear that the remaining complex (simple) roots are ωr and $\overline{\omega} r$ as claimed.

(iii): Note that $\Delta \geqslant (p/3)^3 > 0$ because $p > 0$. The fact that r is a root is shown exactly as in (a). It is simple because g' has no real roots, and r is unique because g is strictly increasing. The complex roots are derived exactly as in (a). □

We now provide a concise version of Theorem 5.5:

Corollary 5.6. (trichotomy) *Set* $\Delta := (p/3)^3 + (q/2)^2$*. Then exactly one of the following holds:*

(i) $p = 0$ *or* $\Delta > 0$*: Then* g *has exactly one real root and it is given by*

$$\sqrt[3]{\frac{-q}{2} + \sqrt{\Delta}} + \sqrt[3]{\frac{-q}{2} - \sqrt{\Delta}}. \tag{5.20}$$

(ii) $p < 0$ *and* $\Delta = 0$*: Then* g *has exactly two real roots which are given by*

$$\frac{3q}{p} = 2\sqrt[3]{\frac{-q}{2}} \quad \textit{and} \quad \frac{-3q}{2p} = -\sqrt[3]{\frac{-q}{2}}. \tag{5.21}$$

(iii) $\Delta < 0$*: Then* g *has exactly three real roots* z_0, z_1, z_2 *which are given by*

$$z_k := 2(-p/3)^{1/2} \cos\left(\frac{\theta + 2k\pi}{3}\right), \quad \textit{where } \theta := \arccos \frac{-q/2}{(-p/3)^{3/2}}, \tag{5.22}$$

and where $z_1 < z_2 < z_0$*.*

5.3 The general cubic

In this section, we turn to the general cubic

$$f(x) := ax^3 + bx^2 + cx + d, \quad \text{where } a, b, c, d \text{ are in } \mathbb{R} \text{ and } a > 0. \tag{5.23}$$

(The case $a < 0$ is treated similarly.) Note that $f''(x) = 6ax + 2b$ has exactly one zero, namely

$$x_0 := \frac{-b}{3a}. \tag{5.24}$$

The change of variables

$$x = z + x_0 \tag{5.25}$$

leads to the well known depressed cubic

$$g(z) := z^3 + pz + q, \quad \text{where } p := \frac{3ac - b^2}{3a^2} \quad \text{and} \quad q := \frac{27a^2d + 2b^3 - 9abc}{27a^3} \tag{5.26}$$

which we reviewed in Section 5.2. Here $ag(z) = f(x) = f(z + x_0)$ so the roots of g are precisely those of f, translated by x_0:

$$x \text{ is a root of } f \Leftrightarrow x - x_0 \text{ is a root of } g. \tag{5.27}$$

So all we need to do is find the roots of g, and then add x_0 to them, to obtain the roots of f. Because the change of variables (5.25) is linear, the multiplicity of the roots is preserved. Translating some of the results from Theorem 5.5 for g to f gives the following:

Theorem 5.7. *f is strictly concave on $]-\infty, x_0]$ and is strictly convex on $[x_0, +\infty[$, where x_0 is the unique inflection point of f defined in (5.24). Recall the definitions of p, q from (5.26) and also set*

$$\Delta := (p/3)^3 + (q/2)^2 = \frac{(3ac-b^2)^3}{(9a^2)^3} + \frac{(27a^2d + 2b^3 - 9abc)^2}{(54a^3)^2}. \tag{5.28}$$

Then exactly one of the following cases occurs:

(i) $\boxed{b^2 > 3ac \Leftrightarrow p < 0}$*: Set $x_\pm := (-b \pm \sqrt{b^2 - 3ac})/(3a)$. Then $x_\pm$ are two distinct simple roots of f', f is strictly increasing on $]-\infty, x_-]$, f is strictly decreasing on $[x_-, x_+]$, f is strictly increasing on $[x_+, +\infty[$. This case trifurcates further as follows:*

(a) $\boxed{\Delta > 0}$*: Then f has exactly one real root; moreover, it is simple and given by*

$$x_0 + u_- + u_+, \quad \textit{where } u_\pm := \sqrt[3]{\frac{-q}{2} \pm \sqrt{\Delta}}. \tag{5.29}$$

The two remaining simple nonreal roots are

$$x_0 - \tfrac{1}{2}(u_- + u_+) \pm \mathrm{i}\tfrac{1}{2}\sqrt{3}(u_- - u_+).$$

(b) $\boxed{\Delta = 0}$*: Then f has two distinct real roots: The simple root is*

$$x_0 + \frac{3q}{p} = x_0 + 2\sqrt[3]{\frac{-q}{2}} = \frac{4abc - b^3 - 9a^2d}{a(b^2 - 3ac)} \tag{5.30}$$

and the double root is

$$x_0 - \frac{3q}{2p} = x_0 - \sqrt[3]{\frac{-q}{2}} = \frac{9ad - bc}{2(b^2 - 3ac)}. \tag{5.31}$$

(c) $\boxed{\Delta < 0}$*: Then f has three simple real roots r_-, r_0, r_+ where $r_- < x_- < r_0 < x_+ < r_+$. Indeed, set*

$$\theta := \arccos \frac{-q/2}{(-p/3)^{3/2}}, \tag{5.32}$$

which lies in $]0, \pi[$, and then define y_0, y_1, y_2 by

$$y_k := x_0 + 2(-p/3)^{1/2} \cos\Big(\frac{\theta + 2k\pi}{3}\Big). \tag{5.33}$$

Then $r_- = y_1$, $r_0 = y_2$, and $r_+ = y_0$.

(ii) $\boxed{b^2 = 3ac \Leftrightarrow p = 0}$*: Then f is strictly increasing on $\mathbb{R}$ and its only real root is*

$$r := x_0 + (-q)^{1/3}. \tag{5.34}$$

If $q = 0$, then r is a triple root. If $q \neq 0$, then r is a simple root and the remaining nonreal simple roots are $x_0 - \frac{1}{2}(-q)^{1/3} \pm \mathrm{i}\frac{1}{2}\sqrt{3}(-q)^{1/3}$.

(iii) $\boxed{b^2 < 3ac \Leftrightarrow p > 0}$ *: Then f is strictly increasing on $\mathbb{R}$, and f has exactly one real root; moreover, it is simple and given by*

$$x_0 + u_- + u_+, \quad \text{where } u_\pm := \sqrt[3]{\frac{-q}{2} \pm \sqrt{\Delta}}. \tag{5.35}$$

The two remaining simple nonreal roots are $x_0 - \frac{1}{2}(u_- + u_+) \pm \mathrm{i}\frac{1}{2}\sqrt{3}(u_- - u_+)$.

In turn, Corollary 5.6 turns into

Corollary 5.8. *Recall* (5.24) *and* (5.26)*, and set*

$$\Delta := (p/3)^3 + (q/2)^2 = \frac{(3ac - b^2)^3}{(9a^2)^3} + \frac{(27a^2d + 2b^3 - 9abc)^2}{(54a^3)^2} \tag{5.36}$$

Then exactly one of the following holds:

(i) $\boxed{b^2 = 3ac \text{ or } \Delta > 0}$ *: Then f has exactly one real root and it is given by*

$$x_0 + \sqrt[3]{\frac{-q}{2} + \sqrt{\Delta}} + \sqrt[3]{\frac{-q}{2} - \sqrt{\Delta}}. \tag{5.37}$$

(ii) $\boxed{b^2 > 3ac \text{ and } \Delta = 0}$ *: Then f has exactly two real roots which are given by*

$$x_0 + \frac{3q}{p} = x_0 + 2\sqrt[3]{\frac{-q}{2}} \quad \text{and} \quad x_0 + \frac{-3q}{2p} = x_0 - \sqrt[3]{\frac{-q}{2}}. \tag{5.38}$$

(iii) $\boxed{\Delta < 0}$ *: Then f has exactly three real (simple) roots r_0, r_1, r_2, where*

$$r_k := x_0 + 2(-p/3)^{1/2} \cos\left(\frac{\theta + 2k\pi}{3}\right), \qquad \theta := \arccos \frac{-q/2}{(-p/3)^{3/2}}, \tag{5.39}$$

and $r_1 < r_2 < r_0$.

5.4 Convex Analysis of the general quartic

In this section, we study the function

$$h(x) := \alpha x^4 + \beta x^3 + \gamma x^2 + \delta x + \varepsilon, \text{ where } \alpha, \beta, \gamma, \delta, \varepsilon \text{ are in } \mathbb{R} \text{ with } \alpha \neq 0. \tag{5.40}$$

We start by characterizing convexity.

Proposition 5.9. (convexity) *The general quartic* (5.40) *is convex if and only if*

$$\alpha > 0 \text{ and } 8\alpha\gamma \geqslant 3\beta^2. \tag{5.41}$$

Proof. Note that $h'(x) = 4\alpha x^3 + 3\beta x^2 + 2\gamma x + \delta$ and, also completing the square,

$$h''(x) = 12\alpha x^2 + 6\beta x + 2\gamma = \frac{3}{4}\alpha\Big(4x + \frac{\beta}{\alpha}\Big)^2 + 2\gamma - \frac{3\beta^2}{4\alpha}. \tag{5.42}$$

Hence $h'' \geqslant 0 \Leftrightarrow [\alpha > 0 \text{ and } 2\gamma \geqslant 3\beta^2/(4\alpha)] \Leftrightarrow$ (5.41). (For further information on deciding the nonnegativity of polynomials[5], see [55, Section 3.1.3].) □

Having characterization convexity, we shall assume this condition for the remainder of this section:

$$h \text{ is convex, i.e., } \alpha > 0 \text{ and } 8\alpha\gamma \geqslant 3\beta^2. \tag{5.43}$$

Before we head on, we recall again some standard notions from [8]. For a function $h : X \to [-\infty, +\infty]$, the *domain* of h is given by

$$\operatorname{dom} h = \big\{x \in X \bigm| h(x) < +\infty\big\},$$

and the *range* of h is $\operatorname{ran} h = \big\{h(x) \bigm| x \in \operatorname{dom} h\big\} \subseteq [-\infty, +\infty]$. A function $f : X \to [-\infty, +\infty]$ is *proper* if

$$(\forall x \in X)\ f(x) > -\infty \text{ and } (\exists x_0 \in X) \text{ such that } f(x_0) < +\infty. \tag{5.44}$$

A function f is *lower semicontinuous* if its *epigraph*,

$$\text{epi } f = \big\{(x, r) \in X \times \mathbb{R} \mid f(x) \leqslant r\big\}, \tag{5.45}$$

is a closed set; equivalently for every sequence $(x_n)_{n\in\mathbb{N}}$ in X,

$$x_n \to x \Rightarrow f(x) \leqslant \liminf_{n\to\infty} f(x_n). \tag{5.46}$$

The set of proper lower semicontinuous convex functions from $X \to\,]-\infty, +\infty]$ is denoted by $\Gamma_0(X)$. Let $h : X \to\,]-\infty, +\infty]$ be proper, where X is Hilbert space. Then ∂h is the *subdifferential* of h is the set-valued operator

$$\partial h : X \rightrightarrows X : \bar{x} \mapsto \big\{y \in X \bigm| \langle x - \bar{x}, y\rangle + h(\bar{x}) \leqslant h(x),\ (\forall x \in X)\big\}, \tag{5.47}$$

The domain of ∂h is $\operatorname{dom}(\partial h) = \big\{x \in X \bigm| \partial h(x) \neq \varnothing\big\}$ and the range of ∂h is $\operatorname{ran}(\partial h) = \big\{y \in X \bigm| \exists\ x \in X \text{ such that } y \in \partial h(x)\big\}$. *The interior* of C is the largest open set contained in C and is given by

$$\operatorname{int} C = \big\{x \bigm| \exists\ \epsilon > 0,\ \mathbb{B}(x, \epsilon) \subseteq C\big\}.$$

Definition 5.10. A function $h \in \Gamma_0(X)$ is *supercoercive* if

$$\lim_{\|x\|\to+\infty} f(x)/\|x\| = +\infty. \tag{5.48}$$

[5] We thank Dr. Amy Wiebe for referring us to [55].

Definition 5.11. Let $f : X \to [-\infty, +\infty]$. The *Fenchel conjugate* of f is given by

$$f^* : X \mapsto [-\infty, +\infty] : u \to \sup_{x \in X} \big(\langle x, u \rangle - f(x) \big). \tag{5.49}$$

Proposition 5.12. (Fenchel conjugate) *Recall our assumptions* (5.40) *and* (5.43)*. Let* $y \in \mathbb{R}$*. Then*

$$h^*(y) = yx_y - h(x_y), \tag{5.50}$$

where $p := (8\alpha\gamma - 3\beta^2)/(16\alpha^2) \geqslant 0$*,* $q := (8\alpha^2(\delta - y) + \beta^3 - 4\alpha\beta\gamma)/(32\alpha^3)$*,* $\Delta := (p/3)^2 + (q/2)^2 \geqslant 0$*, and*

$$x_y := -\frac{\beta}{4\alpha} + \sqrt[3]{\frac{-q}{2} + \sqrt{\Delta}} + \sqrt[3]{\frac{-q}{2} - \sqrt{\Delta}}. \tag{5.51}$$

Proof. Because h is supercoercive, it follows from [8, Proposition 14.15] that $\operatorname{dom} h^* = \mathbb{R}$. Combining with the differentiability of h, it follows that $y \in \operatorname{int}\operatorname{dom}\, h^* \subseteq \operatorname{dom} \partial h^* = \operatorname{ran} \partial h = \operatorname{ran} h'$. However, if $h'(x) = y$, then $h^*(y) = xy - h(x)$ and we have found the conjugate. It remains to solve $h'(x) = y$, i.e.,

$$4\alpha x^3 + 3\beta x^2 + 2\gamma x + \delta - y = 0. \tag{5.52}$$

So we set

$$f(x) := ax^3 + bx^2 + cx + d, \quad \text{where } \ a := 4\alpha, \ \ b := 3\beta, \ \ c := 2\gamma, \ \ d := \delta - y.$$

To solve (5.52), i.e., $f(x) = 0$, we first note that

$$p = \frac{3ac - b^2}{3a^2} = \frac{3(4\alpha)(2\gamma) - (3\beta)^2}{3(4\alpha)^2} = \frac{8\alpha\gamma - 3\beta^2}{16\alpha^2} \geqslant 0,$$

where the inequality follows from (5.43). Next,

$$\begin{aligned} q = \frac{3^3 a^2 d + 2b^3 - 3^2 abc}{(3a)^3} &= \frac{3^3 4^2 \alpha^2(\delta - y) + 2(3^3\beta^3) - 3^2(4\alpha)(3\beta)(2\gamma)}{3^3 4^3 \alpha^3} \\ = \frac{8\alpha^2(\delta - y) + \beta^3 - 4\alpha\beta\gamma}{32\alpha^3} & \end{aligned}$$

and

$$\Delta = (p/3)^3 + (q/2)^2 \geqslant 0,$$

where the inequality follows because $p \geqslant 0$. Then $-b/(3a) = -\beta/(4\alpha)$ and now Corollary 5.8(i) yields the unique solution of $f(x) = 0$ as (5.51). □

Definition 5.13. [8, Definition 12.23] Let $f \in \Gamma_0\big(X\big)$ and let $x \in X$. Then $\operatorname{Prox}_f(x)$ is the unique point in X that satisfies

$$\min_{y \in X} \Big(f\big(y\big) + \frac{1}{2}\,\|x - y\|^2 \Big) = f\big(\operatorname{Prox}_f x\big) + \frac{1}{2}\,\big\|x - \operatorname{Prox}_f x\big\|^2. \tag{5.53}$$

The operator $\operatorname{Prox}_f : X \to X$ is the *proximity operator* or the *proximal mapping* of f.

Proposition 5.14. (proximal mapping) *Recall our assumptions* (5.40) *and* (5.43)*. Let $y \in \mathbb{R}$. Then*

$$\operatorname{Prox}_h(y) = -\frac{\beta}{4\alpha} + \sqrt[3]{\frac{-q}{2} + \sqrt{\Delta}} + \sqrt[3]{\frac{-q}{2} - \sqrt{\Delta}}, \tag{5.54}$$

where

$$p := \frac{4\alpha(1+2\gamma) - 3\beta^2}{16\alpha^2}, \quad q := \frac{8\alpha^2(\delta - y) + \beta^3 - 2\alpha\beta(1+2\gamma)}{32\alpha^3}, \tag{5.55}$$

and $\Delta := (p/3)^3 + (q/2)^2 \geqslant 0$*.*

Proof. Because h is differentiable and full domain, it follows that $\operatorname{Prox}_h(y)$ is the *unique* solution x of the equation $h'(x) + x - y = 0$. The proof thus proceeds analogously to that of Proposition 5.12 — the only difference is we must solve

$$f(x) := ax^3 + bx^2 + cx + d, \quad \text{where } a := 4\alpha,\ b := 3\beta,\ c := 2\gamma + 1,\ d := \delta - y.$$

(The only difference is that $c = 2\gamma + 1$ rather than 2γ due to the additional term "$+x$".) Thus we know *a priori* that the resulting cubic must have a unique real solution. We now have

$$\begin{aligned} 0 < \frac{1}{4\alpha} \leqslant \frac{1}{4\alpha} + \frac{8\alpha\gamma - 3\beta^2}{16\alpha^2} = \frac{4\alpha(1+2\gamma) - 3\beta^2}{16\alpha^2} = p = \frac{12\alpha(1+2\gamma) - 9\beta^2}{48\alpha^2} \\ = \frac{3ac - b^2}{3a^2}, \end{aligned}$$

which is our usual p from discussing roots of the cubic f. Similarly, the q defined here is the same as the usual q for $f(x)$ (see (5.26)). Finally, the formula for $x = \operatorname{Prox}_h(y)$ now follows from Corollary 5.8(i). □

Example 5.15. Suppose that

$$h(x) = x^4 + x^3 + x^2 + x + 1, \tag{5.56}$$

and let $y \in \mathbb{R}$. Then h is convex and

$$h^*(y) = yx_y - h(x_y), \quad \text{where} \tag{5.57a}$$

$$x_y = -\frac{1}{4} + \frac{1}{2}\sqrt[3]{y - \tfrac{5}{8} + \sqrt{(y - \tfrac{5}{8})^2 + (\tfrac{5}{4})^3}} + \frac{1}{2}\sqrt[3]{y - \tfrac{5}{8} - \sqrt{(y - \tfrac{5}{8})^2 + (\tfrac{5}{4})^3}}. \tag{5.57b}$$

Moreover,

$$\operatorname{Prox}_h(y) = -\frac{1}{4} + \frac{1}{2}\sqrt[3]{y - \tfrac{3}{8} + \sqrt{(y - \tfrac{3}{8})^2 + (\tfrac{3}{4})^3}} + \frac{1}{2}\sqrt[3]{y - \tfrac{3}{8} - \sqrt{(y - \tfrac{3}{8})^2 + (\tfrac{3}{4})^3}}. \tag{5.58}$$

See Figure 5.1 for a visualization.

Proof. Note that h fits the pattern of (5.40) with $\alpha = \beta = \gamma = \delta = \varepsilon = 1$. The characterization of convexity presented in Proposition 5.9 turns into $1 > 0$ and $8 \geqslant 3$ which are both obviously true. Hence h is convex.

To compute the Fenchel conjugate, we apply Proposition 5.12 and get $p = 5/16$, $q = (5 - 8y)/32 = -(y - 5/8)/4$, and $\Delta = 5^3/16^3 + (y - 5/8)^2/8^2$. Then $-q/2 = (y - 5/8)/8$. Hence (5.51) turns into

$$\begin{aligned} x_y &= -\frac{1}{4} + \sqrt[3]{\frac{y-5/8}{8} + \sqrt{\frac{(y-5/8)^2}{8^2} + \frac{5^3}{16^3}}} + \\ &\qquad \sqrt[3]{\frac{y-5/8}{8} - \sqrt{\frac{(y-5/8)^2}{8^2} + \frac{5^3}{16^3}}} \\ &= -\frac{1}{4} + \sqrt[3]{\frac{y-5/8}{8} + \sqrt{\frac{(y-5/8)^2}{8^2} + \frac{5^3}{8^2 \cdot 4^3}}} + \\ &\qquad \sqrt[3]{\frac{y-5/8}{8} - \sqrt{\frac{(y-5/8)^2}{8^2} + \frac{5^3}{8^2 \cdot 4^3}}}, \end{aligned}$$

which simplifies to (5.57a).

To compute $\operatorname{Prox}_h(y)$, we utilize Proposition 5.14. Obtaining fresh values for p, q, Δ, we have this time $p = 9/16$, $q = (3 - 8y)/32$, $\Delta = ((8y-3)^2 + 27)/4096 = ((8y-3)^2 + 3^3)/64^2$. Hence $-q/2 = (8y-3)/64$ and $\sqrt{\Delta} = \sqrt{(8y-3)^2 + 3^3}/64$. It follows that

$$\begin{aligned} \sqrt[3]{\frac{-q}{2} \pm \sqrt{\Delta}} &= \sqrt[3]{\frac{8y-3}{64} \pm \frac{\sqrt{(8y-3)^2 + 3^3}}{64}} = \frac{1}{4}\sqrt[3]{8y - 3 \pm \sqrt{(8y-3)^2 + 3^3}} \\ &= \frac{1}{2}\sqrt[3]{y - \tfrac{3}{8} \pm \sqrt{(y - \tfrac{3}{8})^2 + (\tfrac{3}{4})^3}}. \end{aligned}$$

This, $-\beta/(4\alpha) = -1/4$, and (5.54) now yields (5.58). □

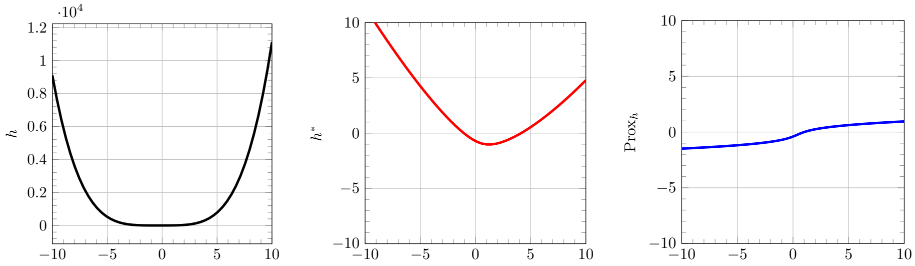

Figure 5.1: A visualization of Example 5.15. Depicted are h (left), its conjugate h^* (middle), and the proximal mapping Prox_h (right).

Example 5.16. Suppose that

$$h(x) = \alpha x^4, \quad \text{where } \alpha > 0, \tag{5.59}$$

and let $y \in \mathbb{R}$. Then

$$h^*(y) = \frac{3}{4(4\alpha)^{1/3}} y^{4/3} \tag{5.60}$$

and

$$\operatorname{Prox}_h(y) = \frac{1}{2}\sqrt[3]{\frac{y}{\alpha} + \sqrt{\frac{1+27\alpha y^2}{27\alpha^3}}} + \frac{1}{2}\sqrt[3]{\frac{y}{\alpha} - \sqrt{\frac{1+27\alpha y^2}{27\alpha^3}}}. \tag{5.61}$$

Proof. Note that h fits the pattern of (5.40) with $\beta = \gamma = \delta = \varepsilon = 0$. The characterization of convexity presented in Proposition 5.9 turns into $\alpha > 0$ and $0 \geqslant 0$ which are both obviously true. Hence h is convex.

We start by computing the Fenchel conjugate of h using Proposition 5.12. We have $p = 0$, $q = -y/(4\alpha)$, $\Delta = (y/(8\alpha))^2$, and $-\beta/(4\alpha) = 0$. Hence $-q/2 = y/(8\alpha)$ and $\sqrt{\Delta} = |y|/(8\alpha)$ which imply $\sqrt[3]{(-q/2) \pm \sqrt{\Delta}} = \sqrt[3]{y/(8\alpha) \pm |y|/(8\alpha)} = \sqrt[3]{\max\{0, y/(4\alpha)\}}$ or $\sqrt[3]{\min\{0, y/(4\alpha)\}}$. Using (5.51), we get

$$x_y = \sqrt[3]{\max\{0, y/(4\alpha)\}} + \sqrt[3]{\min\{0, y/(4\alpha)\}} = \sqrt[3]{y/(4\alpha)}.$$

Using Proposition 5.12, we obtain

$$\begin{aligned} h^*(y) &= yx_y - h(x_y) = yy^{1/3}/(4\alpha)^{1/3} - \alpha y^{4/3}/(4\alpha)^{4/3} \\ &= \frac{|y|^{4/3}}{4^{1/3}\alpha^{1/3}} - \frac{|y|^{4/3}}{4^{4/3}\alpha^{1/3}} = \frac{|y|^{4/3}}{4^{1/3}\alpha^{1/3}}\left(1 - \tfrac{1}{4}\right) = \frac{3|y|^{4/3}}{4(4\alpha)^{1/3}} \end{aligned}$$

as claimed.

To compute $\operatorname{Prox}_h(y)$, we utilize Proposition 5.14. Obtain fresh values of p, q, Δ, we have this time $p = 1/(4\alpha) > 0$ and $q = -y/(4\alpha)$ (see (5.55)). Hence $\Delta = (p/3)^3 + (q/2)^2 = (1 + 27\alpha y^2)/(1728\alpha^3)$ and so $\sqrt{\Delta} = \sqrt{1 + 27\alpha y^2}/(8(3\alpha)^{3/2})$. Now $-\beta/(4\alpha) = 0$ and $-q/2 = y/(8\alpha)$, so (5.54) yields

$$\begin{aligned} \operatorname{Prox}_h(y) &= \sqrt[3]{\frac{y}{8\alpha} + \frac{\sqrt{1+27\alpha y^2}}{8(3\alpha)^{3/2}}} + \sqrt[3]{\frac{y}{8\alpha} - \frac{\sqrt{1+27\alpha y^2}}{8(3\alpha)^{3/2}}} \\ &= \frac{1}{2}\sqrt[3]{\frac{y}{\alpha} + \sqrt{\frac{1+27\alpha y^2}{27\alpha^3}}} + \frac{1}{2}\sqrt[3]{\frac{y}{\alpha} - \sqrt{\frac{1+27\alpha y^2}{27\alpha^3}}} \end{aligned}$$

as claimed. □

Remark 5.17. The Fenchel conjugate formula (5.60) is known and can also be computed by combining, e.g., [8, Example 13.2(i) with Proposition 13.23(i)]. The prox formula (5.61) appears — with a typo though — in [8, Example 24.38(v)].

5.5 The proximal mapping of α/x

In this section, for $\alpha > 0$, we study the convex reciprocal function

$$h(x) := \begin{cases} \alpha/x, & \text{if } x > 0; \\ +\infty, & \text{if } x \leqslant 0. \end{cases} \tag{5.62}$$

The Fenchel conjugate h^*, which requires only solving a *quadratic* equation, is essentially known (e.g., combine [8, Example 13.2(ii) and Proposition 13.23(i)]), and given by

$$h^*(y) = \begin{cases} -2\sqrt{-\alpha y}, & \text{if } y \leqslant 0; \\ +\infty, & \text{if } y > 0. \end{cases} \tag{5.63}$$

The purpose of this section is to explicitly compute Prox_h. We have the following result:

Proposition 5.18. *Suppose that h is given by* (5.62)*, and let $y \in \mathbb{R}$. Set $y_0 := -3\sqrt[3]{\alpha/4} \approx -1.88988\sqrt[3]{\alpha}$. Then we have the following three possibilities:*

(i) If $y_0 < y$, then

$$\operatorname{Prox}_h(y) = \frac{y}{3} + \sqrt[3]{\frac{\alpha}{2} + \Big(\frac{y}{3}\Big)^3 + \sqrt{\alpha\Big(\frac{\alpha}{4} + \Big(\frac{y}{3}\Big)^3\Big)}} + \sqrt[3]{\frac{\alpha}{2} + \Big(\frac{y}{3}\Big)^3 - \sqrt{\alpha\Big(\frac{\alpha}{4} + \Big(\frac{y}{3}\Big)^3\Big)}}.$$

(ii) If $y = y_0$, then

$$\operatorname{Prox}_h(y_0) = \sqrt[3]{\alpha}/\sqrt[3]{4} \approx 0.62996\sqrt[3]{\alpha}.$$

(iii) If $y < y_0$, then

$$\operatorname{Prox}_h(y) = \frac{y}{3}\left(1 - 2\cos\Big(\frac{1}{3}\arccos\frac{(y/3)^3 + \alpha/2}{-(y/3)^3}\Big)\right).$$

Proof. Because $\operatorname{dom} h = \mathbb{R}_{++}$, we must find the *positive* solution of the equation $h'(x) + x - y = 0$. Since $h'(x) = -\alpha x^{-2}$, we are looking for the (necessarily unique) positive solution of $x^2(h'(x) + x - y) = 0$, i.e., of

$$x^3 - yx^2 - \alpha = 0.$$

This fits the pattern of (5.23) in Section 5.3, with parameters $a = 1$, $b = -y$, $c = 0$, and $d = -\alpha$. As in (5.26), we set

$$p := \frac{3ac - b^2}{3a^2} = -\frac{y^2}{3}\begin{cases} < 0, & \text{if } y \neq 0; \\ = 0, & \text{if } y = 0 \end{cases} \tag{5.64}$$

and

$$q := \frac{27a^2d + 2b^3 - 9abc}{27a^3} = -\alpha - 2(y/3)^3.$$

Next,

$$\begin{aligned}\Delta &= (p/3)^3 + (q/2)^2 = -y^6/9^3 + (\alpha + 2(y/3)^3)^2/4 \\ &= -(y/3)^6 + \alpha^2/4 + \alpha(y/3)^3 + (y/3)^6 \\ &= \alpha\big(\alpha/4 + (y/3)^3\big).\end{aligned}$$

Hence

$$\Delta \begin{cases} < 0 \Leftrightarrow y < y_0; \\ = 0 \Leftrightarrow y = y_0; \\ > 0 \Leftrightarrow y > y_0, \end{cases} \qquad \text{where } \; y_0 := -\frac{3}{\sqrt[3]{4}}\sqrt[3]{\alpha} \approx -1.88988\sqrt[3]{\alpha}. \tag{5.65}$$

Now set

$$x_0 := -\frac{b}{3a} = \frac{y}{3}. \tag{5.66}$$

Note that

$$-q/2 = (y/3)^3 + \alpha/2. \tag{5.67}$$

We now discuss the three possibilities from Corollary 5.8 — these will correspond to the three items of the result!

Case 1: $b^2 = 3ac$ or $\Delta > 0$; equivalently, $y = 0$ or $y > y_0$; equivalently, $y_0 < y$.
Then Corollary 5.8(i) yields

$$\begin{aligned}\operatorname{Prox}_h(y) &= x_0 + \sqrt[3]{\frac{-q}{2} + \sqrt{\Delta}} + \sqrt[3]{\frac{-q}{2} - \sqrt{\Delta}} \\ &= \frac{y}{3} + \sqrt[3]{\frac{\alpha}{2} + \Big(\frac{y}{3}\Big)^3 + \sqrt{\alpha\Big(\frac{\alpha}{4} + \Big(\frac{y}{3}\Big)^3\Big)}} + \\ &\qquad\qquad \sqrt[3]{\frac{\alpha}{2} + \Big(\frac{y}{3}\Big)^3 - \sqrt{\alpha\Big(\frac{\alpha}{4} + \Big(\frac{y}{3}\Big)^3\Big)}}\end{aligned}$$

as claimed.

Case 2: $\Delta = 0$; equivalently, $y = y_0$.
Then Corollary 5.8(ii) yields two distinct real roots. We can take a short cut here, though: By exploiting the continuity of Prox_h at y_0 via $\operatorname{Prox}_h(y_0) = \lim_{y\to y_0^+} \operatorname{Prox}_h(y)$, we get

$$\operatorname{Prox}_h(y_0) = \frac{y_0}{3} + 2\sqrt[3]{\frac{\alpha}{2} + \Big(\frac{y_0}{3}\Big)^3} = \sqrt[3]{\alpha}/\sqrt[3]{4} \approx 0.62996\sqrt[3]{\alpha}.$$

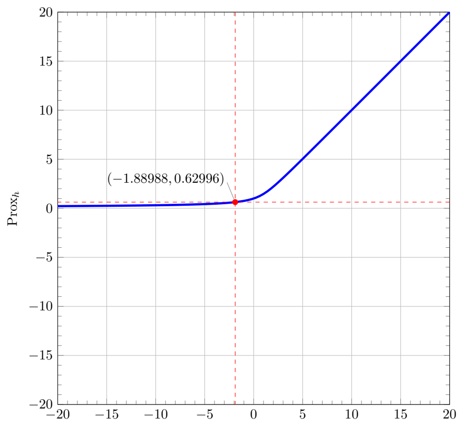

Figure 5.2: A visualization of Proposition 5.18 when $\alpha = 1$.

Case 3: $\Delta < 0$; equivalently, $y < y_0$.
By uniqueness of $\mathrm{Prox}_h(y)$, the desired root must be the *largest* (and the only positive) real root offered in this case (see Corollary 5.8(iii)):

$$\mathrm{Prox}_h(y) = x_0 + 2(-p/3)^{1/2} \cos\left(\frac{\theta}{3}\right), \quad \text{where } \theta := \arccos \frac{-q/2}{(-p/3)^{3/2}}. \tag{5.68}$$

By (5.64), $-p/3 = y^2/9$; thus, using $y < y_0 < 0$, we have $(-p/3)^{1/2} = -y/3$, $(-p/3)^{3/2} = -(y/3)^3$, and (5.67) yields $-(q/2)/(-p/3)^{3/2} = -((y/3)^3 + \alpha/2)/(y/3)^3$. This and (5.66) results in

$$\mathrm{Prox}_h(y) = \frac{y}{3} - 2\frac{y}{3} \cos\left(\frac{\theta}{3}\right), \quad \text{where } \theta := \arccos\left(-\frac{(y/3)^3 + \alpha/2}{(y/3)^3}\right). \tag{5.69}$$

□

Remark 5.19. Suppose that $\alpha = 1$. Then Prox_h was discussed in [27]; however, no explicit formulae were presented. For a visualization of Prox_h in this case, see Figure 5.2.

5.6 Projection onto epigraph of a parabola

In this section, for $\alpha > 0$, we study projection onto the epigraph of the function

$$h\colon \mathbb{R}^n \to \mathbb{R}\colon \mathbf{x} \mapsto \alpha\|\mathbf{x}\|^2. \tag{5.70}$$

Theorem 5.20. *Set* $E := \operatorname{epi} h := \big\{(x,r) \bigm| h(x) \leqslant r\big\} \subseteq \mathbb{R}^{n+1}$. *Let* $(\mathbf{y},\eta) \in (\mathbb{R}^n \times \mathbb{R})$. *If* $(\mathbf{y},\eta) \in E$, *then* $P_E(\mathbf{y},\eta) = (\mathbf{y},\eta)$. *So we assume that* $(\mathbf{y},\eta) \in$

$(\mathbb{R}^n \times \mathbb{R}) \smallsetminus E$, i.e., $\alpha\|\mathbf{y}\|^2 > \eta$. Set $\nu := \|\mathbf{y}\| \geqslant 0$,

$$p := -\frac{(2\alpha\eta - 1)^2}{12\alpha^2}, \quad q := \frac{(2\alpha\eta-1)^3 - 27\alpha^2\nu^2}{108\alpha^3}, \tag{5.71}$$

$\Delta := (p/3)^3 + (q/2)^2 = (27\alpha^2\nu^2 - 2(2\alpha\eta-1)^3)\nu^2/(1728\alpha^4)$, and

$$x := \begin{cases} -\dfrac{\alpha\eta+1}{3\alpha} + \sqrt[3]{-q/2+\sqrt{\Delta}} + \sqrt[3]{-q/2-\sqrt{\Delta}}, & \text{if } \Delta \geqslant 0; \\[2ex] -\dfrac{\alpha\eta+1}{3\alpha} + \dfrac{|2\alpha\eta-1|}{3\alpha}\cos\Big(\dfrac{1}{3}\arccos\dfrac{-q/2}{(-p/3)^{3/2}}\Big), & \text{if } \Delta < 0. \end{cases} \tag{5.72}$$

Then

$$P_E(\mathbf{y},\eta) = \Big(\frac{\mathbf{y}}{1+2\alpha x}, \eta + x\Big). \tag{5.73}$$

See Figure 5.3 for an illustration for the case $\alpha = 1/2$.

Proof. For $x \geqslant 0$, we have $xh = x\alpha\|\cdot\|^2$, $x\nabla h = 2\alpha x \operatorname{Id}$, $\operatorname{Id} + x\nabla h = (1+2\alpha x)\operatorname{Id}$ and therefore $\operatorname{Prox}_{xh} = (1+2\alpha x)^{-1}\operatorname{Id}$. In view of [16, Theorem 6.36], we must first find a positive root x of $\varphi(x) := h(\operatorname{Prox}_{xh}(\mathbf{y})) - x - \eta = \alpha\|\mathbf{y}\|^2/(1+2\alpha x)^2 - x - \eta = 0$. Note that $\varphi(0) > 0$, that φ is strictly decreasing on $\mathbb{R}_+$, and that $\varphi(x) \to -\infty$ as $x \to +\infty$. Hence φ has *exactly one* positive root. Multiplying by $(1+2\alpha x)^2 > 0$, where $x > 0$, results in the cubic $\alpha\nu^2 - (x+\eta)(1+2\alpha x)^2 = 0$, which must have *exactly one* positive root. Re-arranging, we are led to

$$f(x) := 4\alpha^2 x^3 + 4\alpha(\alpha\eta+1)x^2 + (4\alpha\eta+1)x + \eta - \alpha\nu^2 = 0, \tag{5.74}$$

a cubic which we know has exactly one positive root. As in Section 5.3, we set

$$a := 4\alpha^2, \;\; b := 4\alpha(\alpha\eta+1), \;\; c := 4\alpha\eta + 1, \;\; d := \eta - \alpha\nu^2 < 0, \tag{5.75}$$

$$p := \frac{3ac - b^2}{3a^2} = -\frac{(2\alpha\eta-1)^2}{12\alpha^2} \leqslant 0, \tag{5.76}$$

and

$$q := \frac{27a^2d + 2b^3 - 9abc}{27a^3} = \frac{8\alpha^3\eta^3 - 12\alpha^2\eta^2 + 6\alpha\eta - 27\alpha^2\nu^2 - 1}{108\alpha^3}. \tag{5.77}$$

We then have

$$\Delta := (p/3)^3 + (q/2)^2 = \frac{\big(8\alpha^3\eta^3 - 12\alpha^2\eta^2 + 6\alpha\eta - 27\alpha^2\nu^2 - 1\big)^2 - (2\alpha\eta-1)^6}{(6\alpha)^6} \tag{5.78a}$$

$$= -\frac{\nu^2}{1728\alpha^4}\big(16\alpha^3\eta^3 - 24\alpha^2\eta^2 + 12\alpha\eta - 27\alpha^2\nu^2 - 2\big) \tag{5.78b}$$

$$= \frac{\nu^2}{1728\alpha^4}\big(27\alpha^2\nu^2 - 2(2\alpha\eta-1)^3\big), \tag{5.78c}$$

as claimed. Utilizing Corollary 5.8, we have

$$x = -\frac{\alpha\eta+1}{3\alpha} + \sqrt[3]{-q/2+\sqrt{\Delta}} + \sqrt[3]{-q/2-\sqrt{\Delta}}, \quad \text{if } \Delta \geqslant 0 \tag{5.79a}$$

and

$$x = -\frac{\alpha\eta+1}{3\alpha} + \frac{|2\alpha\eta-1|}{3\alpha}\cos\Big(\frac{1}{3}\arccos\frac{-q/2}{(-p/3)^{3/2}}\Big) \quad \text{if } \Delta < 0. \tag{5.79b}$$

(Because we know there is *exactly one* positive root, it is clear that we must pick r_0 in Corollary 5.8(iii) when $\Delta < 0$.) Finally, [16, Theorem 6.36] yields

$$P_E(\mathbf{y},\eta) = \big(\mathrm{Prox}_{xh}(\mathbf{y}), \eta + x\big) = \Big(\frac{\mathbf{y}}{1+2\alpha x}, \eta + x\Big) \tag{5.80}$$

as claimed. □

5.7 A proximal mapping of a closure of a perspective functions

Recall for a convex function $\varphi : X \to \left]-\infty,+\infty\right]$, the perspective function is given by

$$f : \mathbb{R}\times X \to \left]-\infty,+\infty\right] \;:\; (\xi,x) \mapsto \begin{cases} \xi\varphi(x/\xi), & \text{if } \xi > 0;\\ +\infty, & \text{otherwise.}\end{cases}$$

The following complete [8, Example 24.57] which stopped short of providing solutions for a cubic encountered.

Example 5.21. Define the function h on $\mathbb{R}^n \times \mathbb{R}$ by

$$h(y,\eta) := \begin{cases} \|y\|^2/(2\eta), & \text{if } \eta > 0;\\ 0, & \text{if } y = 0 \text{ and } \eta = 0;\\ +\infty, & \text{otherwise.}\end{cases} \tag{5.81}$$

Let $\gamma > 0$ and $(y,\eta) \in \mathbb{R}^n \times \mathbb{R}$. Then

$$\mathrm{Prox}_{\gamma h}(y,\eta) = \begin{cases} (0,0), & \text{if } \|y\|^2 + 2\gamma\eta \leqslant 0;\\ \Big(\big(1 - \frac{\gamma\lambda}{\|y\|}\big)y, \eta + \frac{\gamma\lambda^2}{2}\Big), & \text{if } \|y\|^2 + 2\gamma\eta > 0,\end{cases} \tag{5.82}$$

where $p = 2(\eta+\gamma)/\gamma$, $\Delta = (p/3)^3 + (\|y\|/\gamma)^2$, and

$$\lambda = \begin{cases} \sqrt[3]{\frac{\|y\|}{\gamma} + \sqrt{\Delta}} + \sqrt[3]{\frac{\|y\|}{\gamma} - \sqrt{\Delta}}, & \text{if } \Delta \geqslant 0;\\[2ex] 2(-p/3)^{1/2}\cos\Big(\frac{1}{3}\arccos\frac{\|y\|/\gamma}{(-p/3)^{3/2}}\Big), & \text{if } \Delta < 0.\end{cases} \tag{5.83}$$

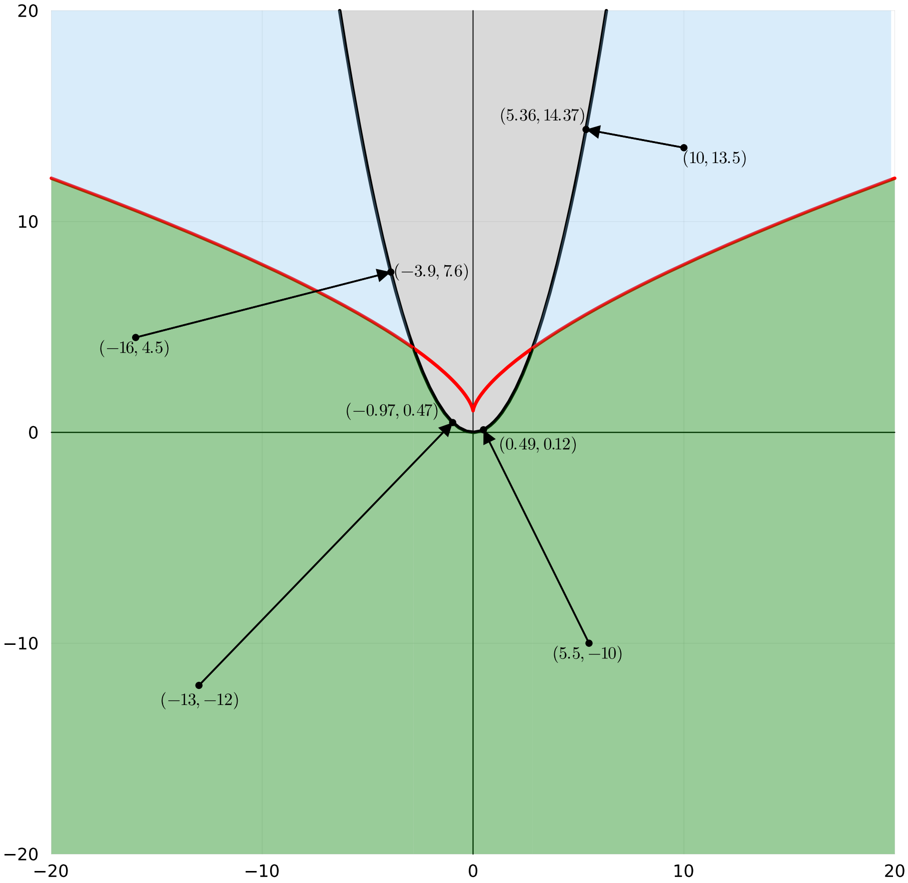

Figure 5.3: A visualization of Theorem 5.20 when $n = 1$ and $\alpha = 1/2$. The epigraph is shown in gray. The red curve corresponds to $\Delta = 0$, the green region to $\Delta < 0$ (where trig functions are used), and the blue region to $\Delta > 0$.

Proof. The first cases were already provided in [8, Example 24.57]. Now assume $\|y\|^2 + 2\gamma\eta > 0$. It was also observed in [8, Example 24.57] that if $y = 0$, then $\mathrm{Prox}_{\gamma h}(y, \eta) = (0, \eta)$.

So assume also that $y \neq 0$. It follows from the discussion that in [8, Example 24.57] that λ is the unique positive solution of the already depressed cubic

$$\lambda^3 + \frac{2(\eta+\gamma)}{\gamma}\lambda - \frac{2\|y\|}{\gamma} = 0, \tag{5.84}$$

which is where the discussion in [8] halted. Continuing here, we set

$$p := \frac{2(\eta+\gamma)}{\gamma}, \quad q := -\frac{2\|y\|}{\gamma} < 0, \tag{5.85}$$

and $\Delta := (p/3)^3 + (q/2)^2$. Using Corollary 5.6, we see that if $\Delta < 0$, then

$$\lambda = 2(-p/3)^{1/2} \cos\Big(\frac{1}{3} \arccos \frac{-q/2}{(-p/3)^{3/2}}\Big) \tag{5.86}$$

while if $\Delta \geqslant 0$, then

$$\lambda = \sqrt[3]{\frac{-q}{2} + \sqrt{\Delta}} + \sqrt[3]{\frac{-q}{2} - \sqrt{\Delta}} \tag{5.87}$$

which slightly simplifies to the expression provided in (5.83).

Finally, notice that if $y = 0$, then the assumption that $\|y\|^2 + 2\gamma\eta > 0$ yields $\eta > 0$; thus, $p > 0$, $q = 0$, and hence $\Delta > 0$. Formally, our λ then simplifies to 0 which conveniently allows us to combine this case with the case $y \neq 0$. □

5.8 On the projection of a hyperbolic paraboloid

Recall that we work with rectangular hyperbolic paraboloids. In this section, X is a real Hilbert space and we set

$$S := \big\{(\mathbf{x}, \mathbf{y}, \gamma) \in X \times X \times \mathbb{R} \;\big|\; \langle \mathbf{x}, \mathbf{y}\rangle = \alpha\gamma\big\}, \quad \text{where } \alpha \in \mathbb{R} \smallsetminus \{0\}. \tag{5.88}$$

Using the Hilbert product space norm $\|(\mathbf{x}, \mathbf{y}, \gamma)\| := \sqrt{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + \beta^2\gamma^2}$, where $\beta > 0$, we are interested in finding the projection onto S. Various cases were discussed in [13], but 3 were treated only implicitly. Armed with the cubic, we are now able to treat two of these cases explicitly (the remaining case features a quintic and remains hard). The first case concerns

$$P_S(\mathbf{z}, -\mathbf{z}, \gamma), \quad \text{when } \mathbf{z} \in X \smallsetminus \{0\} \text{ and } \alpha(\gamma - \alpha/\beta^2) < -\|\mathbf{z}\|^2/4, \tag{5.89}$$

while the second case is

$$P_S(\mathbf{z}, \mathbf{z}, \gamma), \quad \text{when } \mathbf{z} \in X \smallsetminus \{0\} \text{ and } \alpha(\gamma + \alpha/\beta^2) > \|\mathbf{z}\|^2/4. \tag{5.90}$$

5.8.1 The case when (5.89) holds

Theorem 5.22. *Suppose* $\mathbf{z} \in X \smallsetminus \{0\}$*, set*

$$\zeta := \|\mathbf{z}\| > 0, \tag{5.91}$$

and assume that

$$\alpha(\gamma - \alpha/\beta^2) < -\zeta^2/4. \tag{5.92}$$

Set

$$p := -\frac{(\alpha+\beta^2\gamma)^2}{3\alpha^2}, \quad q := \frac{2(\alpha+\beta^2\gamma)^3}{27\alpha^3} + \frac{\beta^2\zeta^2}{\alpha^2}, \tag{5.93}$$

and

$$\Delta := (p/3)^3 + (q/2)^2 = \frac{\beta^2\zeta^2}{\alpha^2}\Big(\frac{\beta^2\zeta^2}{4\alpha^2} + \frac{(\alpha+\beta^2\gamma)^3}{27\alpha^3}\Big). \tag{5.94}$$

If $\Delta \geqslant 0$*, then set*

$$x := \frac{2\alpha - \beta^2\gamma}{3\alpha} + \sqrt[3]{\frac{-q}{2} + \sqrt{\Delta}} + \sqrt[3]{\frac{-q}{2} - \sqrt{\Delta}}; \tag{5.95}$$

and if $\Delta < 0$*, then set*

$$x := \frac{2\alpha - \beta^2\gamma}{3\alpha} + \delta\frac{2(\alpha+\beta^2\gamma)}{3\alpha}\cos\left(\frac{1}{3}\Big((3+\delta)\pi + \arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right) \tag{5.96a}$$

where

$$\delta := \operatorname{sign}(\alpha^2 + \alpha\beta^2\gamma) \in \{-1, 0, 1\}. \tag{5.96b}$$

Then $-1 < x < 1$ *and*

$$P_S(\mathbf{z}, -\mathbf{z}, \gamma) = \Big(\frac{\mathbf{z}}{1-x}, \frac{-\mathbf{z}}{1-x}, \gamma + \frac{\alpha x}{\beta^2}\Big). \tag{5.97}$$

Proof. By [13, Theorem 4.1(ii)(a)], there exists a *unique* $x \in\,]-1, 1[$ such that

$$\frac{2\zeta^2}{(1-x)^2} + \frac{2\alpha^2 x}{\beta^2} + 2\alpha\gamma = 0; \tag{5.98}$$

multiplying by $\beta^2(1-x)^2/2 > 0$ yields the cubic

$$f(x) := ax^3 + bx^2 + cx + d = 0 \tag{5.99}$$

where

$$a := \alpha^2 > 0,\ \ b := \alpha\beta^2\gamma - 2\alpha^2,\ \ c := \alpha^2 - 2\alpha\beta^2\gamma,\ \ d := \alpha\beta^2\gamma + \beta^2\zeta^2. \tag{5.100}$$

Our strategy is to systematically discuss all cases of Theorem 5.7 and then combine cases as much as possible. As usual, we set

$$x_0 := -\frac{b}{3a} = \frac{2\alpha - \beta^2\gamma}{3\alpha} = \frac{2\alpha^2 - \alpha\beta^2\gamma}{3\alpha^2} \text{ and } p := \frac{3ac - b^2}{3a^2} = -\frac{(\alpha+\beta^2\gamma)^2}{3\alpha^2} \leqslant 0, \tag{5.101}$$

and we note that the definition of p is consistent with the one given in (5.93). We have the characterization

$$p = 0 \;\Leftrightarrow\; \alpha + \beta^2\gamma = 0 \;\Leftrightarrow\; \gamma = -\alpha/\beta^2. \tag{5.102}$$

Again as usual, we set

$$q := \frac{27a^2d + 2b^2 - 9abc}{27a^3} = \frac{2(\alpha+\beta^2\gamma)^3}{27\alpha^3} + \frac{\beta^2\zeta^2}{\alpha^2}, \tag{5.103}$$

which matches (5.93), and of course

$$\Delta := (p/3)^3 + (q/2)^2 = \frac{\beta^2\zeta^2}{\alpha^2}\Big(\frac{\beta^2\zeta^2}{4\alpha^2} + \frac{(\alpha+\beta^2\gamma)^3}{27\alpha^3}\Big), \tag{5.104}$$

which matches (5.94). We now systematically discuss the case of Theorem 5.7. *Case 1:* $p < 0$, i.e., $\alpha + \beta^2\gamma \neq 0$ by (5.93).

Case 1(a): $p < 0$ and $\Delta > 0$.
Then Theorem 5.7(a) and the definition x_0 in (5.101) yield (5.95).

Case 1(b): $p < 0$ and $\Delta = 0$.
By Theorem 5.7(b), there are two roots, $x_0 + 3q/p$ and $x_0 - 3q/(2p)$, one of which lies in $]-1,1[$. Now

$$\begin{aligned} x_0 - \frac{3q}{2p} - 1 &= \frac{2\alpha - \beta^2\gamma}{3\alpha} - \frac{3}{2}\frac{2(\alpha+\beta^2\gamma)^3 + 27\alpha\beta^2\zeta^2}{27\alpha^3}\frac{-3\alpha^2}{(\alpha+\beta^2\gamma)^2} - 1 \\ &= \frac{2\alpha - \beta^2\gamma}{3\alpha} + \frac{(\alpha+\beta^2\gamma)^3 + 27\alpha\beta^2\zeta^2/2}{3\alpha(\alpha+\beta^2\gamma)^2} - 1 \\ &= \frac{\big((2\alpha - \beta^2\gamma) + (\alpha+\beta^2\gamma) - (3\alpha)\big)}{3\alpha(\alpha+\beta^2\gamma)^2}(\alpha+\beta^2\gamma)^2 + \frac{27\alpha\beta^2\zeta^2/2}{3\alpha(\alpha+\beta^2\gamma)^2} \\ &= \frac{9\beta^2\zeta^2}{2(\alpha+\beta^2\gamma)^2} \\ &\geqslant 0; \end{aligned}$$

hence the root $x_0 - 3q/(2p)$ lies in $[1,+\infty[$ and therefore our desired root is the remaining one, namely $x_0 + 3q/p$, which also allows us to use the representation (5.95).

Case 1(c): $p < 0$ and $\Delta < 0$.
According to Theorem 5.7(c), we have three distinct real roots, but there is information about their location. We must locate the root in $]-1,1[$. First, $b^2 - 3ac = (\alpha^2 + \alpha\beta^2\gamma)^2$ which yields $\sqrt{b^2 - 3ac} = |\alpha^2 + \alpha\beta^2\gamma|$. This and the definition of b yields

$$\begin{aligned} x_\pm &:= \frac{-b \pm \sqrt{b^2 - 3ac}}{3a} = \frac{2\alpha^2 - \alpha\beta^2\gamma \pm |\alpha^2 + \alpha\beta^2\gamma|}{3\alpha^2} \\ &= \frac{1}{3\alpha^2}\frac{(3\alpha^2) + (\alpha^2 - 2\alpha\beta^2\gamma) \pm \big|(3\alpha^2) - (\alpha^2 - 2\alpha\beta^2\gamma)\big|}{2}. \end{aligned}$$

Hence

$$x_- = \frac{\min\{3\alpha^2, \alpha^2 - 2\alpha\beta^2\gamma\}}{3\alpha^2} < \frac{\max\{3\alpha^2, \alpha^2 - 2\alpha\beta^2\gamma\}}{3\alpha^2} = x_+. \tag{5.105}$$

We now bifurcate one last time.

Case 1(c)(+): $p < 0$, $\Delta < 0$, and $\alpha^2 + \alpha\beta^2\gamma > 0$.
Then $3\alpha^2 > \alpha^2 - 2\alpha\beta^2\gamma$ and therefore $x_+ = 1$. It follows that our desired root x is the "middle root" corresponding to $k = 2$ in Theorem 5.7(c):

$$\begin{aligned} x &= x_0 + 2(-p/3)^{1/2} \cos\left(\frac{1}{3}\Big(4\pi + \arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right) \\ &= \frac{2\alpha - \beta^2\gamma}{3\alpha} + \frac{2|\alpha + \beta^2\gamma|}{3|\alpha|} \cos\left(\frac{1}{3}\Big(4\pi + \arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right) \\ &= \frac{2\alpha - \beta^2\gamma}{3\alpha} + \frac{2(\alpha + \beta^2\gamma)}{3\alpha} \cos\left(\frac{1}{3}\Big(4\pi + \arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right), \end{aligned}$$

where in the last line we used the assumption to deduce that $|\alpha + \beta^2\gamma|/|\alpha| = |\alpha^2 + \alpha\beta^2\gamma|/\alpha^2 = (\alpha^2 + \alpha\beta^2\gamma)/\alpha^2 = (\alpha + \beta^2\gamma)/\alpha$.

Case 1(c)(−): $p < 0$, $\Delta < 0$, and $\alpha^2 + \alpha\beta^2\gamma \leqslant 0$.
Then $3\alpha^2 \leqslant \alpha^2 - 2\alpha\beta^2\gamma$ and therefore $x_- = 1$. It follows that our desired root is the "smallest root" corresponding to $k = 1$ in Theorem 5.7(c):

$$\begin{aligned} x &= x_0 + 2(-p/3)^{1/2} \cos\left(\frac{1}{3}\Big(2\pi + \arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right) \\ &= \frac{2\alpha - \beta^2\gamma}{3\alpha} + \frac{2|\alpha + \beta^2\gamma|}{3|\alpha|} \cos\left(\frac{1}{3}\Big(2\pi + \arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right) \\ &= \frac{2\alpha - \beta^2\gamma}{3\alpha} - \frac{2(\alpha + \beta^2\gamma)}{3\alpha} \cos\left(\frac{1}{3}\Big(2\pi + \arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right), \end{aligned}$$

where in the last line we used the assumption to deduce that $|\alpha + \beta^2\gamma|/|\alpha| = |\alpha^2 + \alpha\beta^2\gamma|/\alpha^2 = -(\alpha^2 + \alpha\beta^2\gamma)/\alpha^2 = -(\alpha + \beta^2\gamma)/\alpha$.

Note that the last two cases can be combined to obtain (5.96).

Case 2: $p = 0$, i.e., $\alpha + \beta^2\gamma = 0$ by (5.93). Then $\Delta = (q/2)^2 \geqslant 0$; hence, $\sqrt{\Delta} = |q|/2$ and thus $\{-q/2 \pm \sqrt{\Delta}\} = \{-q, 0\}$. By Theorem 5.7(ii), the only real root is $x_0 + (-q)^{1/3} = x_0 + (-q/2 + \sqrt{\Delta})^{1/3} + (-q/2 - \sqrt{\Delta})^{1/3}$ which is the same as (5.95) using (5.101).

Case 3: $p > 0$. In vie of (5.93), this case never occurs. □

5.8.2 The case when (5.90) holds

Theorem 5.23. *Suppose* $\mathbf{z} \in X \smallsetminus \{0\}$*, set*

$$\zeta := \|\mathbf{z}\| > 0, \tag{5.106}$$

and assume that

$$\alpha(\gamma + \alpha/\beta^2) > \zeta^2/4. \tag{5.107}$$

Set

$$p := -\frac{(\beta^2\gamma-\alpha)^2}{3\alpha^2}, \quad q := \frac{2(\beta^2\gamma-\alpha)^3}{27\alpha^3} - \frac{\beta^2\zeta^2}{\alpha^2}, \tag{5.108}$$

and

$$\Delta := (p/3)^3 + (q/2)^2 = \frac{\beta^2\zeta^2}{\alpha^2}\Big(\frac{\beta^2\zeta^2}{4\alpha^2} - \frac{(\beta^2\gamma-\alpha)^3}{27\alpha^3}\Big). \tag{5.109}$$

If $\Delta \geqslant 0$, *then set*

$$x := -\frac{2\alpha+\beta^2\gamma}{3\alpha} + \sqrt[3]{\frac{-q}{2}+\sqrt{\Delta}} + \sqrt[3]{\frac{-q}{2}-\sqrt{\Delta}}; \tag{5.110}$$

and if $\Delta < 0$, *then set*

$$x := -\frac{2\alpha+\beta^2\gamma}{3\alpha} + \delta\frac{2(\alpha-\beta^2\gamma)}{3\alpha}\cos\left(\frac{1}{3}\Big((2+2\delta)\pi + \arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right) \tag{5.111a}$$

where

$$\delta := \operatorname{sign}(\alpha^2 - \alpha\beta^2\gamma) \in \{-1, 0, 1\}. \tag{5.111b}$$

Then $-1 < x < 1$ *and*

$$P_S(\mathbf{z}, \mathbf{z}, \gamma) = \Big(\frac{\mathbf{z}}{1+x}, \frac{\mathbf{z}}{1+x}, \gamma + \frac{\alpha x}{\beta^2}\Big). \tag{5.112}$$

Proof. By [13, Theorem 4.1(iii)(a)], there exists a *unique* $x \in \,]-1, 1[$ such that

$$\frac{2\zeta^2}{(1+x)^2} - \frac{2\alpha^2 x}{\beta^2} - 2\alpha\gamma = 0; \tag{5.113}$$

multiplying by $-\beta^2(1+x)^2/2 < 0$ yields the cubic

$$f(x) := ax^3 + bx^2 + cx + d = 0 \tag{5.114}$$

where

$$a := \alpha^2 > 0, \quad b := \alpha\beta^2\gamma + 2\alpha^2, \quad c := \alpha^2 + 2\alpha\beta^2\gamma, \quad d := \alpha\beta^2\gamma - \beta^2\zeta^2. \tag{5.115}$$

Our strategy is to systematically discuss all cases of Theorem 5.7 and then combine cases as much as possible. As usual, we set

$$x_0 := -\frac{b}{3a} = -\frac{2\alpha+\beta^2\gamma}{3\alpha} = -\frac{2\alpha^2+\alpha\beta^2\gamma}{3\alpha^2} \text{ and } p := \frac{3ac-b^2}{3a^2} = -\frac{(\beta^2\gamma-\alpha)^2}{3\alpha^2} \leqslant 0, \tag{5.116}$$

and we note that the definition of p is consistent with the one given in (5.108). We have the characterization

$$p = 0 \;\Leftrightarrow\; \beta^2\gamma - \alpha = 0 \;\Leftrightarrow\; \gamma = \alpha/\beta^2. \tag{5.117}$$

Again as usual, we set

$$q := \frac{27a^2d + 2b^2 - 9abc}{27a^3} = \frac{2(\beta^2\gamma-\alpha)^3}{27\alpha^3} - \frac{\beta^2\zeta^2}{\alpha^2}, \tag{5.118}$$

which matches (5.108), and of course

$$\Delta := (p/3)^3 + (q/2)^2 = \frac{\beta^2\zeta^2}{\alpha^2}\Big(\frac{\beta^2\zeta^2}{4\alpha^2} - \frac{(\beta^2\gamma-\alpha)^3}{27\alpha^3}\Big), \tag{5.119}$$

which matches (5.109). We now systematically discuss the case of Theorem 5.7. *Case 1:* $p < 0$, i.e., $\beta^2\gamma - \alpha \neq 0$ by (5.108).

Case 1(a): $p < 0$ and $\Delta > 0$.
Then Theorem 5.7(a) and the definition of x_0 in (5.116) yield (5.110).

Case 1(b): $p < 0$ and $\Delta = 0$.
By Theorem 5.7(b), there are two roots, $x_0 + 3q/p$ and $x_0 - 3q/(2p)$, one of which lies in $]-1, 1[$. Now

$$\begin{aligned}
x_0 - \frac{3q}{2p} + 1 &= -\frac{2\alpha+\beta^2\gamma}{3\alpha} - \frac{3}{2}\frac{2(\beta^2\gamma-\alpha)^3 - 27\alpha\beta^2\zeta^2}{27\alpha^3}\frac{-3\alpha^2}{(\beta^2\gamma-\alpha)^2} + 1\\
&= -\frac{2\alpha+\beta^2\gamma}{3\alpha} + \frac{(\beta^2\gamma-\alpha)^3 - 27\alpha\beta^2\zeta^2/2}{3\alpha(\beta^2\gamma-\alpha)^2} + 1\\
&= \frac{\big((-2\alpha-\beta^2\gamma) + (\beta^2\gamma-\alpha) + (3\alpha)\big)}{3\alpha(\beta^2\gamma-\alpha)^2}(\beta^2\gamma-\alpha)^2 - \frac{27\alpha\beta^2\zeta^2/2}{3\alpha(\beta^2\gamma-\alpha)^2}\\
&= -\frac{9\beta^2\zeta^2}{2(\beta^2\gamma-\alpha)^2}\\
&\leqslant 0;
\end{aligned}$$

hence the root $x_0 - 3q/(2p)$ lies in $]-\infty, -1]$ and therefore our desired root is the remaining one, namely $x_0 + 3q/p$, which also allows us to use the representation (5.110).

Case 1(c): $p < 0$ and $\Delta < 0$.
According to Theorem 5.7(c), we have three distinct real roots, but there is information about their location. We must locate the root in $]-1, 1[$. First, $b^2 - 3ac = (\alpha^2 - \alpha\beta^2\gamma)^2$ which yields $\sqrt{b^2-3ac} = |\alpha^2 - \alpha\beta^2\gamma|$. This and the definition of b yields

$$\begin{aligned}
x_\pm := \frac{-b \pm \sqrt{b^2-3ac}}{3a} &= \frac{-2\alpha^2 - \alpha\beta^2\gamma \pm |\alpha^2 - \alpha\beta^2\gamma|}{3\alpha^2}\\
&= \frac{1}{3\alpha^2}\frac{(-3\alpha^2) + (-\alpha^2 - 2\alpha\beta^2\gamma) \pm \big|(-3\alpha^2) - (-\alpha^2 - 2\alpha\beta^2\gamma)\big|}{2}.
\end{aligned}$$

Hence

$$x_- = \frac{\min\{-3\alpha^2, -\alpha^2 - 2\alpha\beta^2\gamma\}}{3\alpha^2} < \frac{\max\{-3\alpha^2, -\alpha^2 - 2\alpha\beta^2\gamma\}}{3\alpha^2} = x_+. \tag{5.120}$$

We now bifurcate one last time.

Case 1(c)(+): $p < 0$, $\Delta < 0$, and $\alpha^2 - \alpha\beta^2\gamma > 0$.
Then $-3\alpha^2 < -\alpha^2 - 2\alpha\beta^2\gamma$ and therefore $x_- = -1$. It follows that our desired

root x is the "middle root" corresponding to $k = 2$ in Theorem 5.7(c):

$$\begin{aligned} x &= x_0 + 2(-p/3)^{1/2} \cos\left(\frac{1}{3}\Big(4\pi + \arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right) \\ &= -\frac{2\alpha+\beta^2\gamma}{3\alpha} + \frac{2|\beta^2\gamma-\alpha|}{3|\alpha|}\cos\left(\frac{1}{3}\Big(4\pi + \arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right) \\ &= -\frac{2\alpha+\beta^2\gamma}{3\alpha} + \frac{2(\alpha-\beta^2\gamma)}{3\alpha}\cos\left(\frac{1}{3}\Big(4\pi + \arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right), \end{aligned}$$

where in the last line we used the assumption to deduce that $|\beta^2\gamma - \alpha|/|\alpha| = |\alpha\beta^2\gamma - \alpha^2|/\alpha^2 = (\alpha^2 - \alpha\beta^2\gamma)/\alpha^2 = (\alpha - \beta^2\gamma)/\alpha$.

Case 1(c)(−): $p < 0$, $\Delta < 0$, and $\alpha^2 - \alpha\beta^2\gamma \leqslant 0$.
Then $-3\alpha^2 \geqslant -\alpha^2 - 2\alpha\beta^2\gamma$ and therefore $x_+ = -1$. It follows that our desired root is the "largest root" corresponding to $k = 0$ in Theorem 5.7(c):

$$\begin{aligned} x &= x_0 + 2(-p/3)^{1/2} \cos\left(\frac{1}{3}\Big(\arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right) \\ &= -\frac{2\alpha+\beta^2\gamma}{3\alpha} + \frac{2|\beta^2\gamma-\alpha|}{3|\alpha|}\cos\left(\frac{1}{3}\Big(\arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right) \\ &= -\frac{2\alpha+\beta^2\gamma}{3\alpha} + \frac{2(\beta^2\gamma-\alpha)}{3\alpha}\cos\left(\frac{1}{3}\Big(\arccos\frac{-q/2}{(-p/3)^{3/2}}\Big)\right), \end{aligned}$$

where in the last line we used the assumption to deduce that $|\beta^2\gamma - \alpha|/|\alpha| = |\alpha\beta^2\gamma - \alpha^2|/\alpha^2 = (\alpha\beta^2\gamma - \alpha^2)/\alpha^2 = (\beta^2\gamma - \alpha)/\alpha$.

Note that the last two cases can be combined to obtain (5.111).

Case 2: $p = 0$, i.e., $\alpha - \beta^2\gamma = 0$ by (5.108). Then $\Delta = (q/2)^2 \geqslant 0$; hence, $\sqrt{\Delta} = |q|/2$ and thus $\{-q/2 \pm \sqrt{\Delta}\} = \{-q, 0\}$. By Theorem 5.7(ii), the only real root is $x_0 + (-q)^{1/3} = x_0 + (-q/2 + \sqrt{\Delta})^{1/3} + (-q/2 - \sqrt{\Delta})^{1/3}$ which is the same as (5.110) using (5.116).

Case 3: $p > 0$. In view of (5.108), this case never occurs. □

Chapter 6

Directional asymptotics of Fejér-monotone sequences

In this chapter, we present the results of the paper [11] titled "Directional asymptotics of Fejér monotone sequences" which appeared in Optimization Letters, 17 (2023), 531–544, and which has already been cited by researchers from Israel and Ukraine [60, 63]. The notion of Fejér monotonicity is instrumental in unifying the convergence proofs of many iterative methods, such as the Krasnoselskii-Mann iteration, the proximal point method, the Douglas-Rachford splitting algorithm, and many others; see, e.g., [8, Chapters 5, 26, 28] and [9, 24, 28–30]. Recently, among many important advances, Rockafellar showed in [61] that in a finite-dimensional Hilbert space[6], the sequences generated by the proximal point algorithm enjoy directionally asymptotic properties. In this chapter, we study directionally asymptotical results of strongly convergent subsequences of Fejér monotone sequences in general Hilbert spaces and also provide examples to show that the sets of directionally asymptotic cluster points can be large and that weak convergence is needed in infnite-dimensional spaces.

The notation that we employ is for the most part standard and follows [8, Chapter 5]; however, a partial list is provided for the reader's convenience.

Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X. We denote the set of *(weak) cluster points of* $(x_n)_{n\in\mathbb{N}}$ by

$$\mathcal{C}\big((x_n)_{n\in\mathbb{N}}\big) := \big\{x \in X \;\big|\; x \text{ is the weak limit of some subsequence of } (x_n)_{n\in\mathbb{N}}\big\}.$$

Of course, if X is finite-dimensional, then $\mathcal{C}\big((x_n)_{n\in\mathbb{N}}\big)$ is the same as the set of strong cluster points of $(x_n)_{n\in\mathbb{N}}$. We write $x_n \to x$ if $(x_n)_{n\in\mathbb{N}}$ converges strongly to x, and $x_n \rightharpoonup x$ if $(x_n)_{n\in\mathbb{N}}$ converges weakly to x. Let C be a subset of X and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X. Then $(x_n)_{n\in\mathbb{N}}$ is *Fejér monotone with respect to* C if

$$(\forall c \in C)(\forall n \in \mathbb{N})\ \|x_{n+1} - c\| \leqslant \|x_n - c\|,$$

and we also call C a *Fejér monotone set of* $(x_n)_{n\in\mathbb{N}}$. The corresponding *support function* of C is defined by $\sigma_C(x) := \sup\langle C, x\rangle$.

Definition 6.1. Let $\mathrm{C} \subset X$. The *support function* of C is given by

$$\sigma_{\mathrm{C}}(y) : X \to\ ]-\infty, +\infty] : y \mapsto \sup\langle \mathrm{C}, y\rangle\,. \tag{6.1}$$

[6] We thank Terry Rockafellar for his inspirational talk at the virtual West Coast Optimization Meeting in May 2021 which stimulated this research.

while the corresponding *distance function* is $d_C(x) := \inf \|C - x\|$, for every $x \in X$. The *polar cone of* C is $C^{\ominus} := \{u \in X \mid \sigma_C(u) \leqslant 0\}$.

Definition 6.2. Let C be a nonempty convex subset of X and $x \in X$. Then the *normal cone operator* to C at x is

$$N_{\mathrm{C}} : X \rightrightarrows X : x \mapsto \begin{cases} \{y \in X \mid \sup\langle \mathrm{C} - x, y\rangle \leqslant 0\}, & \text{if } x \in \mathrm{C}; \\ \varnothing, & \text{otherwise.} \end{cases} \tag{6.2}$$

Note that if $z \in C$, then $N_C(z) = (C - z)^{\ominus}$ is the *normal cone* of C at x. For a set-valued monotone operator $A : X \rightrightarrows X$, the corresponding *resolvent* is $J_A := (\operatorname{Id} + A)^{-1}$.

6.1 Auxiliary results

We start by reviewing some preparatory results on Fejér monotone sequences.

Lemma 6.3. *Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X. Then the following hold:*

(i) The largest Fejér monotone set of $(x_n)_{n\in\mathbb{N}}$ is the (possibly empty) closed convex set

$$\bigcap_{n\in\mathbb{N}} \{z \in X \mid 2\langle x_n - x_{n+1}, z\rangle \leqslant \|x_n\|^2 - \|x_{n+1}\|^2\},$$

and is closed convex.

(ii) If C is a Fejér monotone set of $(x_n)_{n\in\mathbb{N}}$, then $\overline{\operatorname{conv}}\, C$, the closed convex hull of C, is a Fejér monotone set of $(x_n)_{n\in\mathbb{N}}$.

(iii) If C_1, C_2 are Fejér monotone sets of $(x_n)_{n\in\mathbb{N}}$, then $C_1 \cup C_2$ is a Fejér monotone set of $(x_n)_{n\in\mathbb{N}}$.

(iv) If C is a Fejér monotone set of $(x_n)_{n\in\mathbb{N}}$, then C is a Fejér monotone set of every subsequence of $(x_n)_{n\in\mathbb{N}}$.

Proof. (i): Let $z \in X$ and $n \in \mathbb{N}$. Then $\|x_{n+1} - z\| \leqslant \|x_n - z\| \Leftrightarrow \|x_{n+1} - z\|^2 \leqslant \|x_n - z\|^2 \Leftrightarrow \|x_{n+1}\|^2 + \|z\|^2 - 2\langle x_{n+1}, z\rangle \leqslant \|x_n\|^2 + \|z\|^2 - 2\langle x_n, z\rangle \Leftrightarrow 2\langle x_n - x_{n+1}, z\rangle \leqslant \|x_n\|^2 - \|x_{n+1}\|^2$. (ii)&(iii): These follow from (i). (iv): Obvious from the definition of Fejér monotonicity. □

The proof of the following result is an extension of a part of Rockafellar's proof of [61, Theorem 2.3].

Lemma 6.4. *Let C be a nonempty closed convex subset of X, let $\bar{z} \in X$, and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X. Suppose that $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to C. Then the following hold for all n, m in $\mathbb{N}$ such that $m \geqslant n + 1$:*

$$(\forall z \in C) \quad \langle x_n - x_{n+1}, z - \bar{z}\rangle \leqslant \tfrac{1}{2}\left(\|x_n - \bar{z}\|^2 - \|x_{n+1} - \bar{z}\|^2\right) \tag{6.3a}$$

$$= \langle x_{n+1} - \bar{z}, x_n - x_{n+1}\rangle + \tfrac{1}{2}\|x_n - x_{n+1}\|^2 \tag{6.3b}$$

$$\leqslant \|x_{n+1} - \bar{z}\|\|x_n - x_{n+1}\| + \tfrac{1}{2}\|x_n - x_{n+1}\|^2 \tag{6.3c}$$

and

$$(\forall z \in C) \quad \langle x_n - x_m, z - \overline{z}\rangle \leqslant \tfrac{1}{2}(\|x_n - \overline{z}\|^2 - \|x_m - \overline{z}\|^2). \tag{6.4}$$

Proof. Let $z \in C$ and let $k \in \{n, n+1, \ldots, m\}$. Because $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to C, we have

$$\begin{aligned} 0 &\leqslant \|x_k - z\|^2 - \|x_{k+1} - z\|^2 \\ &= \|x_k - x_{k+1}\|^2 + 2\langle x_k - x_{k+1}, x_{k+1} - z\rangle \\ &= \|x_k - x_{k+1}\|^2 + 2\left\langle x_k - x_{k+1}, (x_{k+1} - \overline{z}) - (z - \overline{z})\right\rangle. \end{aligned}$$

Therefore,

$$\begin{aligned} \langle x_k - x_{k+1}, z - \overline{z}\rangle &\leqslant \tfrac{1}{2}\|x_k - x_{k+1}\|^2 + \langle x_k - x_{k+1}, x_{k+1} - \overline{z}\rangle && \text{(6.5a)} \\ &= \tfrac{1}{2}\big(\|x_k - \overline{z}\|^2 - \|x_{k+1} - \overline{z}\|^2\big) && \text{(6.5b)} \end{aligned}$$

which yields (6.3a) and (6.3b). Next, (6.3c) is just Cauchy-Schwarz. Finally, (6.4) follows by summing (6.5) from $k = n$ to $k = m$ and telescoping. □

We now obtain information on the location of the set of weak cluster points of a sequence. We start with a fact on lower semicontinuity from [8, Theorem 9.1].

Fact 6.5. *Let $f : X \to \left]-\infty, +\infty\right]$ be a proper lower semicontinuous convex function. Then f is weakly lower semicontinuous.*

Lemma 6.6. *Let C be a nonempty subset of X and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X. Suppose that*

$$\varlimsup_{n\to\infty} \sigma_C(x_n) \leqslant 0.$$

Then

$$\mathcal{C}\big((x_n)_{n\in\mathbb{N}}\big) \subseteq C^{\ominus}.$$

Proof. Suppose that $x \in \mathcal{C}\big((x_n)_{n\in\mathbb{N}}\big)$ and to the contrary that $x \notin C^{\ominus}$. Then $\sigma_C(x) > 0$ and there exists a weakly convergent subsequence $(x_{k_n})_{n\in\mathbb{N}}$ of $(x_n)_{n\in\mathbb{N}}$ such that $x_{k_n} \rightharpoonup x$. The weak lower semicontinuity of σ_C now implies

$$0 < \sigma_C(x) \leqslant \varliminf_{n\to\infty} \sigma_C(x_{k_n}) \leqslant \varlimsup_{n\to\infty} \sigma_C(x_{k_n}) \leqslant \varlimsup_{n\to\infty} \sigma_C(x_n) \leqslant 0,$$

which is absurd! □

Lemma 6.7. *Suppose that X is finite-dimensional, let C be a nonempty closed subset of X, and let $(x_n)_{n\in\mathbb{N}}$ be a bounded sequence in X. Then*

$$d_C(x_n) \to 0 \;\;\Leftrightarrow\;\; \mathcal{C}\big((x_n)_{n\in\mathbb{N}}\big) \subseteq C.$$

Proof. "⇒": Let x be a cluster point in $\mathcal{C}\big((x_n)_{n\in\mathbb{N}}\big)$, say $x_{k_n} \to x$. The continuity of d_C and the assumption yield

$$d_C(x) = \lim_{n\to\infty} d_C(x_{k_n}) = \lim_{n\to\infty} d_C(x_n) = 0.$$

Hence $x \in C$ because C is closed.

"⇐": Suppose to the contrary that $\overline{\lim}_{n\to\infty} d_C(x_n) > 0$. Then there exists a subsequence $(x_{k_n})_{n\in\mathbb{N}}$ of $(x_n)_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} d_C(x_{k_n}) = \varlimsup_{n\to\infty} d_C(x_n) > 0. \tag{6.6}$$

Recall that $(x_n)_{n\in\mathbb{N}}$ is bounded and X is finite-dimensional. Using Bolzano-Weierstrass and after passing to another subsequence and relabeling, we may and do assume that $x_{k_n} \to x$. By assumption, $x \in C$. But then $d_C(x_{k_n}) \to d_C(x) = 0$ which contradicts (6.6). □

We end this section with results on linear isometries.

Definition 6.8. A mapping $T : X \to X$ is an *isometry* if

$$\big(\forall x \in X\big)\big(\forall y \in X\big) \quad \|Tx - Ty\| = \|x - y\|.$$

Lemma 6.9. *Let $A : X \to X$ be a linear isometry. Then the following hold:*

(i) If $T := \frac{1}{2}\operatorname{Id} + \frac{1}{2}A$, then $(\forall x \in X)$ $Tx \perp (x - Tx)$.

(ii) If $A^* = A^{-1} = -A$, then $J_A = \frac{1}{2}\operatorname{Id} - \frac{1}{2}A$ and $(\forall x \in X)$ $J_Ax \perp (x - J_Ax)$.

Proof. Let $x \in X$.

(i): Note that $\operatorname{Id} - T = \frac{1}{2}\operatorname{Id} - \frac{1}{2}$. Hence

$$4\langle Tx, x - Tx\rangle = \langle x + Ax, x - Ax\rangle = \|x\|^2 - \|Ax\|^2 = 0.$$

(ii): Clearly, $\pm A$ is monotone and $A^2 = -\operatorname{Id}$. Hence, [10, Proposition 2.10] yields $J_A = \frac{1}{2}\operatorname{Id} - \frac{1}{2}A = \operatorname{Id} - T$, where $T = \frac{1}{2}\operatorname{Id} + \frac{1}{2}A$. Now apply (i). □

Corollary 6.10. *Let $A : X \to X$ be a linear operator such that $A^* = A^{-1} = -A$, let $x_0 \in X \smallsetminus \{0\}$, and set*

$$(\forall n \in \mathbb{N}) \quad x_{n+1} := J_Ax_n = \tfrac{1}{2}x_n - \tfrac{1}{2}Ax_n.$$

Then $x_n \to 0$ and $(\forall n \in \mathbb{N})$ $x_{n+1} \neq x_n$, and

$$(\forall n \in \mathbb{N}) \quad \left\langle \frac{x_{n+1}}{\|x_{n+1}\|}, \frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|} \right\rangle = 0. \tag{6.7}$$

Proof. Clearly, A is a maximally monotone isometry. The formula for J_Ax_n is a consequence of Lemma 6.9(ii) which also yields (6.7) after we prove that the quotients are well defined which we do next. Let $x \in X$. Then $(\operatorname{Id} + A)^{-1}x = J_Ax = 0 \Leftrightarrow x = (\operatorname{Id} + A)(0) = 0$ and $(\operatorname{Id} + A)^{-1}x = J_Ax = x \Leftrightarrow x = (\operatorname{Id} + A)x \Leftrightarrow Ax = 0 \Leftrightarrow x = 0$. We have shown that if $x \neq 0$, then $J_Ax \neq 0$ and $J_Ax \neq x$. A straightforward induction yields $(\forall n \in \mathbb{N})$ $x_n \neq 0$ and $x_{n+1} \neq x_n$, as claimed. Finally, [21, Corollary 1.2] implies that $x_n \to 0$. □

Remark 6.11. When $X = \mathbb{R}^2$ and

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

then Corollary 6.10 recovers [61, the example on page 11]. Note that (see [47, page 206]) a linear isometry need not be surjective.

6.2 Directional asymptotics and Fejér monotonicity

We are now ready for our main results on the directionally asymptotic behaviour of Fejér monotone sequences. Inspired by Rockafellar's proof of [61, Theorem 2.3], we show that his result reaches to the setting of general Fejér monotone sequences.

Theorem 6.12. *Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X that is Fejér monotone with respect to some nonempty closed convex subset Z of X. Suppose that $x_n \to \bar{z} \in X$ and that $(\forall n \in \mathbb{N})$ $x_{n+1} \neq x_n \neq \bar{z}$. Then*

$$\mathcal{C}\bigg(\Big(\frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|}\Big)_{n\in\mathbb{N}}\bigg) \cup \mathcal{C}\bigg(\Big(\frac{x_n - \bar{z}}{\|x_n - \bar{z}\|}\Big)_{n\in\mathbb{N}}\bigg) \subseteq (Z - \bar{z})^{\ominus}; \tag{6.8}$$

in particular, if $\bar{z} \in Z$, then we may replace $(Z - \bar{z})^{\ominus}$ by $N_Z(\bar{z})$ in (6.8)*.*

Proof. Let $n \in \mathbb{N}$. Taking the supremum over $z \in Z$ in (6.3) yields

$$\sigma_{Z-\bar{z}}(x_n - x_{n+1}) \leqslant \|x_{n+1} - \bar{z}\|\|x_n - x_{n+1}\| + \tfrac{1}{2}\|x_n - x_{n+1}\|^2. \tag{6.9}$$

Dividing (6.9) by $\|x_n - x_{n+1}\|$ and using the positive homogeneity of $\sigma_{Z-\bar{z}}$, we have

$$\sigma_{Z-\bar{z}}\Big(\frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|}\Big) \leqslant \|x_{n+1} - \bar{z}\| + \tfrac{1}{2}\|x_n - x_{n+1}\|. \tag{6.10}$$

Because $x_n \to \bar{z}$ and so $x_n - x_{n+1} \to 0$, we let $n \to \infty$ in (6.10) to learn that

$$\varlimsup_{n\to\infty} \sigma_{Z-\bar{z}}\Big(\frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|}\Big) \leqslant 0. \tag{6.11}$$

Combining (6.11) with Lemma 6.6, we obtain

$$\mathcal{C}\bigg(\Big(\frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|}\Big)_{n\in\mathbb{N}}\bigg) \subseteq (Z - \bar{z})^{\ominus}. \tag{6.12}$$

Next, let $m \geqslant n + 1$. Taking the supermum over $z \in Z$ in (6.4) yields

$$\sigma_{Z-\bar{z}}(x_n - x_m) \leqslant \tfrac{1}{2}\big(\|x_n - \bar{z}\|^2 - \|x_m - \bar{z}\|^2\big). \tag{6.13}$$

Passing to the limit as $m \to \infty$ and using the lower semicontinuity of $\sigma_{Z-\bar{z}}$ in (6.13), we obtain

$$\sigma_{Z-\bar{z}}(x_n - \bar{z}) \leqslant \varliminf_{m\to\infty} \sigma_{Z-\bar{z}}(x_n - x_m) \leqslant \tfrac{1}{2}\|x_n - \bar{z}\|^2. \tag{6.14}$$

Dividing (6.14) by $\|x_n - \bar{z}\|$ and using the positive homogeneity of $\sigma_{Z-\bar{z}}$, we have

$$\sigma_{Z-\bar{z}}\Big(\frac{x_n - \bar{z}}{\|x_n - \bar{z}\|}\Big) \leqslant \tfrac{1}{2}\|x_n - \bar{z}\|. \tag{6.15}$$

Because $x_n \to \bar{z}$, we let $n \to \infty$ in (6.15) and get

$$\varlimsup_{n\to\infty} \sigma_{Z-\bar{z}}\Big(\frac{x_n - \bar{z}}{\|x_n - \bar{z}\|}\Big) \leqslant 0. \tag{6.16}$$

Combining (6.16) with our trusted Lemma 6.6, we obtain

$$\mathcal{C}\bigg(\Big(\frac{x_n - \bar{z}}{\|x_n - \bar{z}\|}\Big)_{n\in\mathbb{N}}\bigg) \subseteq (Z - \bar{z})^{\ominus}. \tag{6.17}$$

Altogether, (6.12) and (6.17) imply (6.8). □

Although ostensibly more general, the following result is actually an easy consequence of Theorem 6.12:

Corollary 6.13. *Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X that is Fejér monotone with respect to some nonempty closed convex subset Z of X. Suppose that $(x_{k_n})_{n\in\mathbb{N}}$ is a subsequence of $(x_n)_{n\in\mathbb{N}}$ such that $x_{k_n} \to \bar{z} \in X$ and that $(\forall n \in \mathbb{N})\ x_{k_{n+1}} \neq x_{k_n} \neq \bar{z}$. Then*

$$\mathcal{C}\bigg(\Big(\frac{x_{k_n} - x_{k_{n+1}}}{\|x_{k_n} - x_{k_{n+1}}\|}\Big)_{n\in\mathbb{N}}\bigg) \cup \mathcal{C}\bigg(\Big(\frac{x_{k_n} - \bar{z}}{\|x_{k_n} - \bar{z}\|}\Big)_{n\in\mathbb{N}}\bigg) \subseteq (Z - \bar{z})^{\ominus}; \tag{6.18}$$

in particular, if $\bar{z} \in Z$, then we may replace $(Z - \bar{z})^{\ominus}$ by $N_Z(\bar{z})$ in (6.18).

Proof. Recalling Lemma 6.3(iv), we simply apply Theorem 6.12 to $(x_{k_n})_{n\in\mathbb{N}}$. □

When X is finite-dimensional, we have the following two nice results:

Corollary 6.14. *Suppose that X is finite-dimensional. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X that is Fejér monotone with respect to some nonempty closed convex subset Z of X. Suppose that $(x_{k_n})_{n\in\mathbb{N}}$ is a subsequence of $(x_n)_{n\in\mathbb{N}}$ such that $x_{k_n} \to \bar{z} \in Z$ and that $(\forall n \in \mathbb{N})\ x_{k_{n+1}} \neq x_{k_n} \neq \bar{z}$. Then*

$$\mathcal{C}\bigg(\Big(\frac{x_{k_n} - x_{k_{n+1}}}{\|x_{k_n} - x_{k_{n+1}}\|}\Big)_{n\in\mathbb{N}}\bigg) \cup \mathcal{C}\bigg(\Big(\frac{x_{k_n} - \bar{z}}{\|x_{k_n} - \bar{z}\|}\Big)_{n\in\mathbb{N}}\bigg) \subseteq \mathbb{S} \cap N_Z(\bar{z});$$

equivalently,

$$\lim_{n\to\infty} d_{\mathbb{S}\cap N_Z(\bar{z})}\Big(\frac{x_{k_n} - x_{k_{n+1}}}{\|x_{k_n} - x_{k_{n+1}}\|}\Big) = 0 \quad \textit{and} \quad \lim_{n\to\infty} d_{\mathbb{S}\cap N_Z(\bar{z})}\Big(\frac{x_{k_n} - \bar{z}}{\|x_{k_n} - \bar{z}\|}\Big) = 0.$$

Proof. Combine Corollary 6.13 with Lemma 6.7. □

Corollary 6.15 (**no zigzagging**). *Suppose that X is finite-dimensional and let $(x_n)_{n\in\mathbb{N}}$ be a sequence that is Fejér monotone with respect to some closed convex subset Z of X. Suppose that $x_n \to \bar{z} \in Z$, that $(\forall n \in \mathbb{N})\ x_{n+1} \neq x_n \neq \bar{z}$, and that $N_Z(\bar{z})$ is a ray. Then*

$$\lim_{n\to\infty} \frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|} = \lim_{n\to\infty} \frac{x_n - \bar{z}}{\|x_n - \bar{z}\|} \in \mathbb{S} \cap N_Z(\bar{z}). \tag{6.19}$$

Proof. Clear from Corollary 6.14 because $\mathbb{S} \cap N_Z(\bar{z})$ is a singleton when $N_Z(\bar{z})$ is a ray. □

Remark 6.16. In Corollary 6.15 the assumption $N_Z(\bar{z})$ being a ray forces int $Z \neq \varnothing$.

Remark 6.17. The results of Theorem 6.12 have been utilized in the discussion for convergence rate bounds of alternating projection methods, see [60, Remark 4.2].

6.3 Large sets of directionally asymptotic cluster points

In this section, we give an example illustrating that the sets of directionally asymptotic cluster points can be large. It also shows that without the ray assumption in Corollary 6.15, (6.19) can go quite wrong.

Fact 6.18 (**Dirichlet**). (See, e.g., [64, page 88]) *Let $\alpha \in \mathbb{R} \smallsetminus \mathbb{Q}$. Then the set $\{n\alpha - \lfloor n\alpha \rfloor \mid n \in \mathbb{N}\}$ is dense in* $[0, 1]$.

Definition 6.19. [8, Definition 4.1] Let C be a nonempty subset X. A mapping $T : \mathrm{C} \to X$ is

(i) *nonexpansive*, or *Lipschitz continuous* with constant 1, if

$$\big(\forall x \in \mathrm{C}\big)\big(\forall y \in \mathrm{C}\big) \quad \|Tx - Ty\| \leqslant \|x - y\|\,; \tag{6.20}$$

(ii) *firmly nonexpansive* if

$$\big(\forall x \in \mathrm{C}\big)\big(\forall y \in \mathrm{C}\big) \quad \|Tx - Ty\|^2 + \big\|\big(\operatorname{Id} - T\big)x - \big(\operatorname{Id} - T\big)y\big\|^2 \leqslant \|x - y\|^2. \tag{6.21}$$

For the remainder of this section, R_α denotes the counterclockwise rotator in the Euclidean plane by α.

Example 6.20. Suppose that $X = \mathbb{R}^2$, let $0 < \theta \notin \frac{1}{2}\pi\mathbb{N}$, Then $T := \frac{1}{2}\operatorname{Id} + \frac{1}{2}R_{2\theta} = \cos(\theta)R_\theta$ is firmly nonexpansive, with $Z := \operatorname{Fix} T = \{0\}$. Let $x_0 \in X \smallsetminus \{0\}$, and set

$$(n \in \mathbb{N}) \quad x_{n+1} := Tx_n. \tag{6.22}$$

Then $x_n \to \bar{z} := 0$, and $(\forall n \in \mathbb{N})\ x_{n+1} \neq x_n \neq \bar{z}$ and $\langle x_n - x_{n+1}, x_{n+1}\rangle = 0$. Moreover, we have the following dichotomy:

(i) $\theta \in 2\pi\mathbb{Q}$ and

$$\mathcal{C}\Big(\Big(\frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|}\Big)_{n\in\mathbb{N}}\Big) \cup \mathcal{C}\Big(\Big(\frac{x_n}{\|x_n\|}\Big)_{n\in\mathbb{N}}\Big) \quad \text{is a } \textit{finite} \text{ subset of } \mathbb{S}.$$

(ii) $\theta \notin 2\pi\mathbb{Q}$ and

$$\mathcal{C}\Big(\Big(\frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|}\Big)_{n\in\mathbb{N}}\Big) = \mathcal{C}\Big(\Big(\frac{x_n}{\|x_n\|}\Big)_{n\in\mathbb{N}}\Big) = \mathbb{S}.$$

Proof. Because $\theta \notin 2\pi\mathbb{Z}$, we have $\operatorname{Fix} T = \{0\}$. Note that

$$T^n = (\cos\theta)^n R_{n\theta} = (\cos\theta)^n \begin{pmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{pmatrix} \quad \text{and} \quad x_n = T^n x_0.$$

Since $0 < |\cos\theta| < 1$ and $R_{n\theta}$ is an isometry, we have $\|T^n\| \to 0$. Thus $x_n \to 0$. By Lemma 6.9(i), we have $\langle x_n - x_{n+1}, x_{n+1}\rangle = 0$. Moreover, $\|x_{n+1}\| = \|\cos\theta R_\theta x_n\| = |\cos\theta|\|x_n\| < \|x_n\|$ because $x_0 \neq 0$, so $(\forall n \in \mathbb{N})$ $x_{n+1} \neq x_n$ and $x_n \neq 0$.

To study the set of cluster points of $(T^n x_0/\|T^n x_0\|)_{n\in\mathbb{N}}$, we consider two cases.

Case 1: $\cos\theta > 0$. We have

$$\frac{T^n x_0}{\|T^n x_0\|} = R_{n\theta}\frac{x_0}{\|x_0\|} = \begin{pmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{pmatrix}\frac{x_0}{\|x_0\|}. \tag{6.23}$$

We proceed with two subcases.

Subcase 1: $\frac{\theta}{2\pi} \in \mathbb{Q}$. Then $\theta = 2\pi\frac{k}{l}$ with $k, l \in \mathbb{N}$ and $l \neq 0$, and $\cos n\theta = \cos\frac{n}{l}(2k\pi)$ and $\sin n\theta = \sin\frac{n}{l}(2k\pi)$. By using $n = ml, ml+1, \ldots, ml+l-1$ with $m \in \mathbb{N}$, the set

$$\big\{R_{n\theta} \bigm| n \in \mathbb{N}\big\} = \big\{R_{t2k\pi/l} \bigm| t = 0, \ldots, l-1\big\}.$$

The sequence $(R_{n\theta})_{n\in\mathbb{N}}$ has at most l cluster points. From (6.23) we see that $(T^n x_0/\|T^n x_0\|)_{n\in\mathbb{N}}$ has at most l cluster points. In fact, if $k = 2$, then there are precisely l cluster points.

Subcase 2: $\frac{\theta}{2\pi} \notin \mathbb{Q}$. Then $\theta = 2\pi\alpha$ with $\alpha \in \mathbb{R}_{++} \smallsetminus \mathbb{Q}$, and

$$\cos n\theta = \cos n(2\pi\alpha) = \cos(n\alpha)(2\pi) = \cos(n\alpha - \lfloor n\alpha\rfloor)(2\pi), \text{ and}$$

$$\sin n\theta = \sin n(2\pi\alpha) = \sin(n\alpha)(2\pi) = \sin(n\alpha - \lfloor n\alpha\rfloor)(2\pi).$$

By Fact 6.18, $\{n\alpha - \lfloor n\alpha\rfloor \mid n \in \mathbb{N}\}$ is dense in $[0,1]$. Hence the set of cluster points of $\{R_{n\theta} \mid n \in \mathbb{N}\}$ is $\{R_\beta \mid \beta \in [0, 2\pi]\}$. By (6.23), the set of cluster points of $(T^n x_0/\|T^n x_0\|)_{n\in\mathbb{N}}$ is $\mathbb{S}$.

Case 2: $\cos\theta < 0$. We have

$$\frac{T^n x_0}{\|T^n x_0\|} = (-1)^n R_{n\theta}\frac{x_0}{\|x_0\|} = (-1)^n \begin{pmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{pmatrix}\frac{x_0}{\|x_0\|}. \tag{6.24}$$

We proceed with two subcases.

Subcase 1: $\frac{\theta}{2\pi} \in \mathbb{Q}$. Due to $(-1)^n$, we have to consider n being even and odd. When n is even, write $n = 2k$ with $k \in \mathbb{N}$,

$$R_{n\theta} = \begin{pmatrix} \cos k(2\theta) & -\sin k(2\theta) \\ \sin k(2\theta) & \cos k(2\theta) \end{pmatrix}.$$

Since $\frac{2\theta}{2\pi} \in \mathbb{Q}$, similar arguments as in Case 1 subcase 1 show that $(R_{k(2\theta)})_{k\in\mathbb{N}}$ has a finite number of cluster points. When n is odd, let $n = 2k+1$ and $\theta = m2\pi/l$ with $k, l, m \in \mathbb{N}$ and $l \neq 0$. If we set $k = tl + s$ for $t, s \in \mathbb{N}$ and $0 \leqslant s \leqslant l-1$, then

$$(2k+1)\frac{m2\pi}{l} = \frac{2(tl+s)+1}{l}m2\pi = \left(2t + \frac{2s+1}{l}\right)m2\pi$$

so that

$$\cos(2k+1)\theta = \cos\left(\frac{2s+1}{l}m2\pi\right), \quad \sin(2k+1)\theta = \sin\left(\frac{2s+1}{l}m2\pi\right)$$

where $0 \leqslant s \leqslant l-1$. This shows that when n is odd, we have most l cluster points. Combining the even and odd cases, the set $\{R_{n\theta} \mid n \in \mathbb{N}\}$ has at most a finite number of cluster points, so is $(T^n x_0/\|T^n x_0\|)_{n\in\mathbb{N}}$ by (6.24).

Subcase 2: $\frac{\theta}{2\pi} \notin \mathbb{Q}$. Put $\theta = \alpha(2\pi)$ with $\alpha \notin \mathbb{Q}$. When n is even, write $n = 2k$ with $k \in \mathbb{N}$. Then $R_{n\theta} = R_{k(2\theta)}$, similar arguments as in Case 1 subcase 2 show that the set of cluster points of $\{R_{k(2\theta)} \mid k \in \mathbb{N}\}$ is $\{R_\beta \mid 0 \leqslant \beta \leqslant 2\pi\}$, because $\frac{2\theta}{2\pi} \in \mathbb{R}_{++} \smallsetminus \mathbb{Q}$.

When n is odd, write $n = 2k+1$ with $k \in \mathbb{N}$. Since that

$$\cos[(2k+1)\alpha 2\pi] = \cos[k(2\alpha)2\pi + 2\alpha\pi] = \cos[(k(2\alpha) - \lfloor k(2\alpha)\rfloor)2\pi + 2\alpha\pi], \text{ and}$$

$$\sin[(2k+1)\alpha 2\pi] = \sin[k(2\alpha)2\pi + 2\alpha\pi] = \sin[(k(2\alpha) - \lfloor k(2\alpha)\rfloor)2\pi + 2\alpha\pi],$$

and that

$$\{k(2\alpha) - \lfloor k(2\alpha)\rfloor \mid k \in \mathbb{N}\} \text{ is dense in } [0,1]$$

we see that the set of cluster points of $\{R_{(2k+1)\theta} \mid k \in \mathbb{N}\}$ is

$$\{R_\beta \mid 2\alpha\pi \leqslant \beta \leqslant 2\alpha\pi + 2\pi\}.$$

Hence, in both cases the set of cluster points of $(T^n x_0/\|T^n x_0\|)_{n\in\mathbb{N}}$ is $\mathbb{S}$.

Finally, we consider the set of cluster points of

$$\left(\frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|}\right)_{n\in\mathbb{N}}.$$

Now

$$\begin{aligned}\operatorname{Id} - T = \frac{\operatorname{Id} - R_{2\theta}}{2} &= \sin\theta \begin{pmatrix} \sin\theta & \cos\theta \\ -\cos\theta & \sin\theta \end{pmatrix} \\ &= \sin\theta \begin{pmatrix} \cos(\theta + 3\pi/2) & -\sin(\theta + 3\pi/2) \\ \sin(\theta + 3\pi/2) & \cos(\theta + 3\pi/2) \end{pmatrix} = \sin\theta R_{(\theta+3\pi/2)},\end{aligned}$$

so that

$$\begin{aligned}\frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|} &= \frac{(\operatorname{Id} - T)x_n}{\|(\operatorname{Id} - T)x_n\|} \\ &= \begin{cases} R_{(\theta+3\pi/2)} \frac{x_n}{\|x_n\|}, & \text{if } \sin\theta > 0; \\ -R_{(\theta+3\pi/2)} \frac{x_n}{\|x_n\|}, & \text{if } \sin\theta < 0. \end{cases}\end{aligned}$$

Then the set of cluster points of

$$\left(\frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|}\right)_{n\in\mathbb{N}}$$

is just $\pm R_{(\theta+3\pi/2)}$ rotations of the set of cluster points of $(x_n/\|x_n\|)_{n\in\mathbb{N}}$. Consequently, the set of cluster points of $((x_n - x_{n+1})/\|x_n - x_{n+1}\|)_{n\in\mathbb{N}}$ is a finite set if $\frac{\theta}{2\pi} \in \mathbb{Q}$; and is $\mathbb{S}$ if $\frac{\theta}{2\pi} \notin \mathbb{Q}$. □

By choosing $\theta \in \pi\mathbb{Q}$ small, then the quotient sequences in Theorem 6.12 will have the same finite number of cluster points. On the other hand, if $\theta \in \pi(\mathbb{R} \smallsetminus \mathbb{Q})$, then *every* unit vector is a cluster point of the two quotient sequences.

Remark 6.21. We do not consider the case when $\theta \in \frac{1}{2}\pi\mathbb{N}$ because then $T = \operatorname{Id}$ or 0 in which case one has finite convergence of $(x_n)_{n\in\mathbb{N}}$.

6.4 Missing the sphere: an infinite-dimensional example

It is interesting to ask whether in infinite-dimensional Hilbert spaces the nonempty sets of weak cluster points in Theorem 6.12 lie in the sphere $\mathbb{S}$. It turns out that the answer is negative, and the sequence provided is obtained by iterating a resolvent. To this end, we assume in this section that

$$X = \ell^2\big(\{1, 2, \ldots\}\big),$$

with the standard Schauder basis $e_1 := (1, 0, 0, \ldots), e_2 := (0, 1, 0, 0, \ldots)$, and so on. We define the *right-shift operator* by

$$R\colon X \to X\colon (\xi_1, \xi_2, \ldots) \mapsto (0, \xi_1, \xi_2, \ldots),$$

Then R is a linear isometry with $\operatorname{Fix} R = \{0\}$. We shall also require the following classical identity.

Fact 6.22 (**Vandermonde's identity**). *(See [42, Section 5.1].) Let m, n, r be in $\mathbb{N}$. Then*

$$\binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k}.$$

Example 6.23. Define the firmly nonexpansive operator $T\colon X \to X$ [7] by

$$T := \tfrac{1}{2}\operatorname{Id} + \tfrac{1}{2}R,$$

set $x_0 := e_1$, and $(x_n)_{n\in\mathbb{N}} := (T^n x_0)_{n\in\mathbb{N}}$ with $x_0 = e_1$. Then the following hold:

(i) $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\operatorname{Fix} T = \{0\}$, and $x_n \to 0$. Moreover, $(\forall n \in \mathbb{N})\ x_{n+1} \neq x_n \neq 0$.

(ii) $(\forall n \in \mathbb{N})\ \langle x_{n+1}, x_n - x_{n+1}\rangle = 0$.

(iii) Both

$$\left(\frac{x_n}{\|x_n\|}\right)_{n\in\mathbb{N}} \quad\text{and}\quad \left(\frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|}\right)_{n\in\mathbb{N}}$$

converge weakly — but not strongly — to $0 \notin \mathbb{S}$.

Proof. (i): It is well known that $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\operatorname{Fix} T = \{0\}$, because T is nonexpansive. Since $R^k e_1 = e_{k+1}$, we have

$$x_n = T^n x_0 = \frac{1}{2^n}(\operatorname{Id} + R)^n x_0 = \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}R^k x_0 = \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}e_{k+1}. \tag{6.25}$$

Hence, by Fact 6.22,

$$\|x_n\|^2 = \frac{1}{4^n}\sum_{k=0}^{n}\binom{n}{k}^2 = \frac{1}{4^n}\binom{2n}{n} = \frac{1}{4^n}\frac{(2n)!}{(n!)^2};$$

in particular, $x_n \neq 0$. $x_{n+1} \neq x_n$ because x_{n+1} contains a nonzero term of e_{n+2} and $e_{n+2} \perp e_{k+1}$ for $1 \leqslant k \leqslant n$.

Now recall *Stirling's formula* (see, e.g., [67, Theorem 5.44]) which states that

$$n! \approx \sqrt{2\pi n}\frac{n^n}{e^n} \tag{6.26}$$

for large n, and which implies

$$\|x_n\|^2 = \frac{1}{4^n}\binom{2n}{n} = \frac{(2n)!}{4^n(n!)^2} \approx \frac{1}{4^n}\frac{\sqrt{2\pi(2n)}(2n/e)^{2n}}{(\sqrt{2\pi n})^2(n/e)^{2n}} = \frac{1}{\sqrt{\pi}}\frac{1}{\sqrt{n}} \to 0.$$

(The qualitative fact that $x_n \to 0$ also follows from [8, Example 5.29].)

(ii): Since R is an isometry, this follows from Lemma 6.9(i).

(iii). For *fixed* k, we have from Stirling's formula (6.26) that

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \approx \frac{1}{k!}\frac{\sqrt{2\pi n}(n/e)^n}{\sqrt{2\pi(n-k)}((n-k)/e)^{n-k}} \approx \frac{n^k}{k!} \tag{6.27}$$

[7] We thank Walaa Moursi for suggesting to investigate the operator T in this section.

for large n. Hence

$$\frac{x_n}{\|x_n\|} \approx \sqrt[4]{\pi}\sqrt[4]{n}\frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}e_{k+1} \approx \sum_{k=0}^{n}\frac{\sqrt[4]{\pi}\sqrt[4]{n}}{2^n}\frac{n^k}{k!}e_{k+1} \rightharpoonup 0$$

because that $(e_k)_{k\in\mathbb{N}}$ is a total set in $\ell^2(\mathbb{N})$ and that for each fixed $k \in \mathbb{N}$ the coefficient of e_{k+1} in $x_n/\|x_n\|$ clearly converges to 0 as $n \to \infty$; see, e.g., [47, Example 4.8-6].

Next,

$$x_n - x_{n+1} = \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}e_{k+1} - \frac{1}{2^{n+1}}\sum_{k=0}^{n+1}\binom{n+1}{k}e_{k+1}. \tag{6.28}$$

Since $e_{n+2} \perp e_{k+1}$ for $0 \leqslant k \leqslant n$, by Fact 6.22 and (6.25) we have

$$\begin{aligned}
\langle x_n, x_{n+1}\rangle &= \left\langle \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}e_{k+1}, \frac{1}{2^{n+1}}\sum_{k=0}^{n+1}\binom{n+1}{k}e_{k+1}\right\rangle \qquad (6.29a)\\
&= \frac{1}{2}\frac{1}{4^n}\left\langle \sum_{k=0}^{n}\binom{n}{k}e_{k+1}, \sum_{k=0}^{n}\binom{n+1}{k}e_{k+1}\right\rangle \qquad (6.29b)\\
&= \frac{1}{2}\frac{1}{4^n}\sum_{k=0}^{n}\binom{n}{k}\binom{n+1}{k} = \frac{1}{2}\frac{1}{4^n}\sum_{k=0}^{n}\binom{n}{n-k}\binom{n+1}{k} \qquad (6.29c)\\
&= \frac{1}{2}\frac{1}{4^n}\binom{2n+1}{n}. \qquad (6.29d)
\end{aligned}$$

It follows from (6.25) and (6.29) that

$$\begin{aligned}
\|x_n - x_{n+1}\|^2 &= \|x_n\|^2 + \|x_{n+1}\|^2 - 2\langle x_n, x_{n+1}\rangle \qquad (6.30a)\\
&= \frac{1}{4^n}\binom{2n}{n} + \frac{1}{4^{n+1}}\binom{2(n+1)}{n+1} - 2\frac{1}{2}\frac{1}{4^n}\binom{2n+1}{n} \qquad (6.30b)\\
&= \frac{1}{4^n}\left[\binom{2n}{n} + \frac{1}{4}\binom{2n+2}{n+1} - \binom{2n+1}{n}\right] \qquad (6.30c)\\
&= \frac{1}{2(n+1)4^n}\binom{2n}{n} = \frac{1}{2(n+1)}\|x_n\|^2 \qquad (6.30d)\\
&\approx \frac{1}{2n}\frac{1}{\sqrt{\pi n}} = \frac{1}{2\sqrt{\pi}}\frac{1}{n^{3/2}} \qquad (6.30e)
\end{aligned}$$

for large n. Combining (6.28), (6.30), and (6.27), we obtain

$$\frac{x_n - x_{n+1}}{\|x_n - x_{n+1}\|} \approx \frac{\sqrt{2}\sqrt[4]{\pi}n^{3/4}}{2^n}\sum_{k=0}^{n}\binom{n}{k}e_{k+1} - \frac{\sqrt{2}\sqrt[4]{\pi}n^{3/4}}{2^{n+1}}\sum_{k=0}^{n+1}\binom{n+1}{k}e_{k+1} \tag{6.31a}$$

$$\approx \sum_{k=0}^{n} \frac{\sqrt{2}\sqrt[4]{\pi}n^{3/4}}{2^n}\binom{n}{k}e_{k+1} - \sum_{k=0}^{n+1} \frac{\sqrt{2}\sqrt[4]{\pi}n^{3/4}}{2^{n+1}}\binom{n+1}{k}e_{k+1} \tag{6.31b}$$

$$\approx \sum_{k=0}^{n} \frac{\sqrt{2}\sqrt[4]{\pi}n^{3/4}}{2^n}\frac{n^k}{k!}e_{k+1} - \sum_{k=0}^{n+1} \frac{\sqrt{2}\sqrt[4]{\pi}n^{3/4}}{2^{n+1}}\frac{(n+1)^k}{k!}e_{k+1} \tag{6.31c}$$

$$\rightharpoonup 0, \tag{6.31d}$$

because for every fixed $k \in \mathbb{N}$ the coefficients of e_{k+1} in $(x_n - x_{n+1})/\|x_n - x_{n+1}\|$ converge to 0 as $n \to \infty$.

In summary, both quotient limits converge weakly but not strongly to 0. □

Chapter 7

Conclusion and open problems

7.1 Main contributions of this thesis

In this thesis, we developed closed-form formulas of orthogonal projections onto the following sets:

- *Crosses* (See Chapter 2).
- *Hyperbolas* (See Chapter 3).
- *Hyperbolic paraboloids* (See Chapter 4 and Section 5.8 of Chapter 5).

We also presented a mathematical model of Elser's framework for the matrix factorization problem in Chapter 1. In Chapter 5, we systematically studied cubics, and utilized a novel and convenient presentation of the roots of cubics to develop formulas for projections operators and proximal mappings. In Chapter 6, we studied the directional asymptotics of Fejér monotone sequences.

7.2 Future work and open problems

We computed the projections $P_{C_\gamma}(x_0, y_0)$ onto hyperbolas of the form

$$C_\gamma = \big\{(x, y) \in X \times X \;\big|\; \langle x, y\rangle = \gamma\big\}, \tag{7.1}$$

in Chapter 3 as solutions to the following equality-constrained problem

$$\text{minimize} \quad f(x, y) := \|x - x_0\|^2 + \|y - y_0\|^2 \text{ sub. to } h(x, y) := \langle x, \operatorname{Id} y\rangle - \gamma = 0.$$

All cases were discussed explicitly, except the case Theorem 3.13(i) in Section 3.3 of Chapter 3 which seeks the unique $\lambda \in \,]-1, 1[$, a solution of a quartic equation $H(\lambda) = 0$, where the quartic polynomial $H(\lambda)$ is given by

$$H(\lambda) := \frac{(\lambda^2 + 1)p - 2\lambda q}{2(1 - \lambda^2)^2} - \gamma, \; p := 2\langle x_0, y_0\rangle, \text{ and } q := \|x_0\|^2 + \|y_0\|^2. \tag{7.2}$$

Open Problem 7.1. *Is it possible to find a closed-form expression of the root λ that solves* (7.2)*?*

Remark 7.2. We suspect a novel analysis to be built upon the results of Chapter 5. Elser solves this (essentially) quartic with Newton's method [40].

In Chapter 4, we study projections on hyperbolic paraboloids, where we encounter cubic and quintic polynomials. We were able to provide closed-form projection formulas in two cases (in Section 5.8 of Chapter 5) by using the systematic study of cubics in Section 5.2 and Section 5.3 of Chapter 5. The case of our general quintic still remains open and is stated below. (See Theorem 4.9(i) in Section 4.4.)

Open Problem 7.3. *Given constants $\alpha \in \mathbb{R}\backslash\{0\}, \beta > 0, \gamma \in \mathbb{R}$, the goal is to find the expression of the unique $\lambda \in \left]-1,1\right[$ that solves the (essentially) quintic equation*

$$g(\lambda) := \frac{(\lambda^2+1)p - 2\lambda q}{(1-\lambda^2)^2} - \frac{2\lambda\alpha^2}{\beta^2} - 2\alpha\gamma_0 = 0, \tag{7.3}$$

where $p := 2\langle x_0, y_0\rangle$ and $q := \|x_0\|^2 + \|y_0\|^2$.

Remark 7.4. We suspect a possible approach would be to rely on the study of cubics and Chapter 5 and a very particular case study of Open Problem 7.1. Recall that there does not exist any quintic formula built out of a finite combination of field operations, continuous functions, and radicals. The quintic (7.3) seems to have a special structure that relates to the geometry of hyperbolic paraboloids. This structure may shed light on the idea of introducing the decompositions of quintics using cubics and quartics.

Let us replace the identity matrix Id in (7.1) with a square matrix $Q_1 \in \mathbb{R}^{n\times n}$ or even a rectangular matrix $Q_2 \in \mathbb{R}^{m\times n}$, which leads to the following sets:

Definition 7.5 (Generalized sets of bilinear forms)**.** Given $\gamma \in \mathbb{R}$ and $\alpha \in \mathbb{R}$, the sets

$$C_\gamma^{Q_1} := \big\{(x,y) \in \mathbb{R}^n \times \mathbb{R}^n \;\big|\; \langle x, Q_1 y\rangle = \gamma\big\}, \tag{7.4}$$

$$C_\alpha^{Q_1} := \big\{(x,y,\gamma) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \;\big|\; \langle x, Q_1 y\rangle = \alpha\gamma\big\}, \tag{7.5}$$

$$C_\gamma^{Q_2} := \big\{(x,y) \in \mathbb{R}^m \times \mathbb{R}^n \;\big|\; x^T Q_2 y = \gamma\big\}, \tag{7.6}$$

$$C_\alpha^{Q_2} := \big\{(x,y,\gamma) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R} \;\big|\; x^T Q_2 y = \alpha\gamma\big\}, \tag{7.7}$$

are some examples of sets that have bilinear structure.

Open Problem 7.6. *A systematic study of orthogonal projections onto the sets given by* (7.4)–(7.7) *still remains open.*

Remark 7.7. The study of these sets would assist in developing different architectures in Elser's framework of deep learning. For example, see [69, Section 4] for different examples of Q matrices, which led to convolutional networks and forward multi-tier networks.

Now, we turn our attention to possible future work in the context of deep learning.

Open Problem 7.8. *To train the neural networks arising from Elser's feasibility frameworks, one employs projection methods [4–7]. However, due to the nonconvexity of bilinear sets, a convergence theory should be further investigated.*

It is a common fact the computational success of the stochastic gradient descent method relies on *automatic differentiation*, also known as *algorithmic differentiation*. It is therefore tempting to develop similar code-based tools for feasibility algorithms including a software package of "automatic projections", which heavily relies on the solution of Open Problem 7.6.

A follow-up open question of the proposed results in Chapter 6 is:

Open Problem 7.9. *Can we extend the results on Fejér monotone sequences to quasi-Fejér monotone sequences [29]?*

Bibliography

[1] Francisco J Aragón Artacho, Jonathan M Borwein, and Matthew K Tam. Recent results on Douglas–Rachford methods for combinatorial optimization problems. *Journal of Optimization Theory and Applications*, 163:1–30, 2014. → pages 1

[2] Francisco J Aragón Artacho, Rubén Campoy, and Veit Elser. An enhanced formulation for solving graph coloring problems with the Douglas–Rachford algorithm. *Journal of Global Optimization*, 77:383–403, 2020. → pages 1

[3] Jean-Pierre Aubin and Hélène Frankowska. *Set-Valued Analysis*. Springer Science & Business Media, 2009. → pages 2, 56

[4] Heinz H Bauschke. The approximation of fixed points of compositions of nonexpansive mappings in hilbert space. *Journal of Mathematical Analysis and Applications*, 202(1):150–159, 1996. → pages 115

[5] Heinz H Bauschke. Projection algorithms: results and open problems. In *Studies in Computational Mathematics*, volume 8, pages 11–22. Elsevier, 2001. → pages

[6] Heinz H Bauschke and Jonathan M Borwein. On the convergence of von Neumann's alternating projection algorithm for two sets. *Set-Valued Analysis*, 1(2):185–212, 1993. → pages

[7] Heinz H Bauschke and Jonathan M Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996. → pages 115

[8] Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017. → pages 2, 3, 4, 18, 19, 20, 36, 82, 83, 86, 87, 91, 93, 100, 102, 106, 110

[9] Heinz H Bauschke, Minh N Dao, and Walaa M Moursi. On Fejér monotone sequences and nonexpansive mappings. *Linear and Nonlinear Analysis*, 1:287–295, 2015. → pages 100

[10] Heinz H Bauschke, Warren L Hare, and Walaa M Moursi. Generalized solutions for the sum of two maximally monotone operators. *SIAM Journal on Control and Optimization*, 52(2):1034–1047, 2014. → pages 103

[11] Heinz H Bauschke, Manish Krishan Lal, and Xianfu Wang. Directional asymptotics of Fejér monotone sequences. *Optimization Letters*, 17(3):531–544, 2023. → pages 100

[12] Heinz H Bauschke, Manish Krishan Lal, and Xianfu Wang. The projection onto the cross. *Set-Valued and Variational Analysis*, 30(3):997–1009, 2022. → pages 17, 36, 56, 72

[13] Heinz H Bauschke, Manish Krishan Lal, and Xianfu Wang. Projecting onto rectangular hyperbolic paraboloids in Hilbert space. *Applied Set-Valued Analysis and Optimization*, 5(2):163–180, 2023. → pages 53, 93, 94, 97

[14] Heinz H Bauschke, Manish Krishan Lal, and Xianfu Wang. Projections onto hyperbolas or bilinear constraint sets in Hilbert spaces. *Journal of Global Optimization*, 86(1):25–36, 2023. → pages 36, 54

[15] Heinz H Bauschke, Manish Krishan Lal, and Xianfu Wang. Real roots of real cubics and optimization. *Journal of Convex Analysis*, 32(1):119–144, 2025. → pages 73

[16] Amir Beck. *First-Order Methods In Optimization*. SIAM, 2017. → pages 90, 91

[17] Frédéric Bernard and Lionel Thibault. Prox-regular functions in Hilbert spaces. *Journal of Mathematical Analysis and Applications*, 303(1):1–14, 2005. → pages 40, 54

[18] Dimitri Bertsekas. *Nonlinear Programming*, volume 4. Athena Scientific, 3rd edition, 2016. → pages 20, 43, 59

[19] LE Blumenson. A derivation of n-dimensional spherical coordinates. *The American Mathematical Monthly*, 67(1):63–66, 1960. → pages 21

[20] Jonathan M Borwein. Proximality and Chebyshev sets. *Optimization Letters*, 1(1):21–32, 2007. → pages 18

[21] Ronald E Bruck and Simeon Reich. Nonexpansive projections and resolvents of accretive operators in banach spaces. *Houston Journal of Mathematics*, 3(4):459–470, 1977. → pages 103

[22] Richard L Burden, J Douglas Faires, and Annette M Burden. *Numerical Analysis*. Cengage Learning, 2015. → pages 45

[23] Enzo Busseti, Walaa M Moursi, and Stephen Boyd. Solution refinement at regular points of conic problems. *Computational Optimization and Applications*, 74:627–643, 2019. → pages 18

[24] Andrzej Cegielski. *Iterative Methods For Fixed Point Problems In Hilbert Spaces*, volume 2057. Springer, 2012. → pages 100

[25] N Chernov and S Wijewickrema. Algorithms for projecting points onto conics. *Journal of Computational and Applied Mathematics*, 251:8–21, 2013. → pages 36

[26] Nikolai Chernov and Hui Ma. Least squares fitting of quadratic curves and surfaces. *Computer Vision*, 285:302, 2011. → pages 36

[27] Giovanni Chierchia, Emilie Chouzenoux, Patrick L Combettes, and Jean-Christophe Pesquet. The proximity operator repository. `http://proximity-operator.net/download/guide.pdf`. → pages 89

[28] Patrick L Combettes. Fejér-monotonicity in convex optimization. *Encyclopedia of optimization*, 2:106–114, 2001. → pages 100

[29] Patrick L Combettes. Quasi-Fejérian analysis of some optimization algorithms. In *Studies in Computational Mathematics*, volume 8, pages 115–152. Elsevier, 2001. → pages 115

[30] Patrick L Combettes. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53(5-6):475–504, 2004. → pages 100

[31] Sara Confalonieri. *The Unattainable Attempt to Avoid the Casus Irreducibilis for Cubic Equations: Gerolamo Cardano's De Regula Aliza*. Springer, 2015. → pages 73, 74

[32] Sean Deyo. *Solving Network Problems by Embedding Discrete Structure in Continuous Space*. PhD thesis, Cornell University, 2023. → pages 36

[33] Sean Deyo and Veit Elser. Avoiding traps in nonconvex problems. *Journal of Applied & Numerical Optimization*, 4(2), 2022. → pages 1, 2

[34] Sean Deyo and Veit Elser. Learning grammar with a divide-and-concur neural network. *Physical Review E*, 105(6):064303, 2022. → pages 2

[35] Sean Deyo and Veit Elser. A logical word embedding for learning grammar. *arXiv preprint arXiv:2304.14590*, 2023. → pages 2

[36] Sean Deyo and Veit Elser. A transparent approach to data representation. *arXiv preprint arXiv:2304.14209*, 2023. → pages 2, 36

[37] Dave Eberly. *Robust and Error-Free Geometric Computing*. CRC Press, 2021. → pages 45

[38] Veit Elser. Phase retrieval by iterated projections. *Journal of the Optical Society of America A*, 20(1):40–55, 2003. → pages 1

[39] Veit Elser. Matrix product constraints by projection methods. *Journal of Global Optimization*, 68(2):329–355, 2017. → pages 2, 36

[40] Veit Elser. Learning without loss. *Fixed Point Theory and Algorithms for Sciences and Engineering*, 2021(1):1–51, 2021. → pages vii, 2, 9, 12, 15, 18, 29, 36, 45, 53, 70, 71, 114

[41] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, 2003. → pages 18

[42] Ronald L Graham, Donald E Knuth, Oren Patashnik, and Stanley Liu. Concrete Mathematics: A Foundation for Computer Science. *Computers in Physics*, 3(5):106–107, 1989. → pages 110

[43] Simon Gravel and Veit Elser. Divide and concur: A general approach to constraint satisfaction. *Physical Review E*, 78(3):036706, 2008. → pages 1

[44] Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*, volume 2. Springer Science & Business Media, 2012. → pages 1

[45] Shuvomoy Das Gupta, Bartolomeo Stellato, and Bart PG Van Parys. Exterior-point optimization for nonconvex learning. *arXiv preprint arXiv:2011.04552*, 2020. → pages 36

[46] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009. → pages 5

[47] Erwin Kreyszig. *Introductory Functional Analysis with Applications*, volume 17. John Wiley & Sons, 1991. → pages 2, 39, 42, 56, 104, 111

[48] Alexander Y Kruger, D Russell Luke, and Nguyen H Thao. Set regularities and feasibility problems. *Mathematical Programming*, 168:279–311, 2018. → pages 17

[49] Hiroki Kuroda, Daichi Kitahara, Eiichi Yoshikawa, Hiroshi Kikuchi, and Tomoo Ushio. Convex estimation of sparse-smooth power spectral densities from mixtures of realizations with application to weather radar. *arXiv preprint arXiv:2309.16215*, 2023. → pages 73

[50] Jena-Lonis Lauriere. A language and a program for stating and solving combinatorial problems. *Artificial Intelligence*, 10(1):29–127, 1978. → pages 1

[51] Yan-Chao Liang and Jane J Ye. Optimality conditions and exact penalty for mathematical programs with switching constraints. *Journal of Optimization Theory and Applications*, 190(1):1–31, 2021. → pages 17

[52] David G Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, 1997. → pages 20

[53] D Russell Luke, Shoham Sabach, and Marc Teboulle. Optimization on spheres: models and proximal algorithms with computational performance comparisons. *SIAM Journal on Mathematics of Data Science*, 1(3):408–445, 2019. → pages 14

[54] Boris Odehnal, Hellmuth Stachel, and Georg Glaeser. *The Universe of Quadrics*. Springer Nature, 2020. → pages 53

[55] Pablo A Parrilo, Grigoriy Blekherman, and Rekha R Thomas. *Semidefinite Optimization and Convex Algebraic Geometry*. SIAM Society for Industrial and Applied Mathematics, 2013. → pages 82

[56] Guy Pierra. Decomposition through formalization in a product space. *Mathematical Programming*, 28:96–115, 1984. → pages 3, 5

[57] Ting Kei Pong and Henry Wolkowicz. The generalized trust region subproblem. *Computational Optimization and Applications*, 58(2):273–322, 2014. → pages 30, 31

[58] Qazi Ibadur Rahman and Gerhard Schmeisser. *Analytic Theory of Polynomials*. Oxford University Press, 2002. → pages 73, 74

[59] A Ramachandra Rao and Pochiraju Bhimasankaram. *Linear Algebra*, volume 19. Springer, 2000. → pages 7

[60] Simeon Reich and Rafał Zalas. Comparing the methods of alternating and simultaneous projections for two subspaces. *arXiv preprint arXiv:2306.12219*, 2023. → pages 100, 106

[61] R Tyrrell Rockafellar. Advances in convergence and scope of the proximal point algorithm. *Journal Nonlinear and Convex Analysis*, 22:2347–2375, 2021. → pages 100, 101, 104

[62] R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science, 2009. → pages 36, 40, 54, 56, 57

[63] Olga Rovenska. Approximation of classes of poisson integrals by incomplete Fejér means. *Archiv der Mathematik*, pages 1–10, 2023. → pages 100

[64] Houshang H Sohrab. *Basic Real Analysis*, volume 231. Springer, 2003. → pages 106

[65] Ronald J Stern and Henry Wolkowicz. Trust region problems and nonsymmetric eigenvalue perturbations. *SIAM Journal on Matrix Analysis and Applications*, 15(3):755–778, 1994. → pages 30

[66] Ronald J Stern and Henry Wolkowicz. Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations. *SIAM Journal on Optimization*, 5(2):286–313, 1995. → pages 30

[67] Karl R Stromberg. *An Introduction to Classical Real Analysis*, volume 376. American Mathematical Soc., 2015. → pages 110

[68] Edward Tsang. *Foundations of Constraint Satisfaction*. BoD–Books on Demand, 2014. → pages 1

[69] Ezra Winston and J Zico Kolter. Monotone operator equilibrium networks. *Advances in Neural Information Processing Systems*, 33:10718–10728, 2020. → pages 114

[70] Lin Zhang, Yi Shen, Hua Xiang, Quan Qian, and Bo Li. Visualization of all two-qubit states via partial-transpose moments. *Physical Review A*, 108(1):012414, 2023. → pages 73